Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
A Framework for Understanding and Evaluating the Quality of Data Sets in Empirical Software Engineering

Marshima Mohd Rosli

Software engineering data sets provide valuable information related to software development and the evolution of software projects. Researchers often obtain software engineering data sets from public data repositories, which contain data and metadata about software development artefacts, such as the bug reports and source code. Typically, the term data refers to measurements associated with metrics and entities, whereas the term metadata describes the meaning and context of the data. However, there is growing concern regarding the quality of such data sets because they can lead to questionable empirical results if the data is of poor quality.

Researchers, who are the creators of data sets, describe the data and metadata in many different ways. This creates challenges that make it difficult to interpret data sets or to report clearly the quality of data. Moreover, researchers intending to use data sets from public data repositories rely on their quality. Although many techniques have been proposed in the literature to resolve the challenges associated with data quality, there is a lack of research that evaluates the quality of data sets in public data repositories.

The main research question in this study is, ‘How can we help researchers to understand the quality of data sets in software engineering?’ Hence, the aim of the study was to develop a data quality assessment framework to interpret the data sets better and evaluate their quality. The framework will allow researchers to understand the quality of data and identify any potential problems that may be present in a data set.

The research was organised into six steps. The first step was a review of data quality in software engineering. The results of a systematic mapping study carried out as part of the review showed that many definitions of data quality issues are unclear because of the inconsistent terminology used in these definitions. The second step was an observation of existing and artificial data sets to explore their different formats and structures. The results from the observation of data sets indicated that few data sets contain metadata to describe what the data mean, which might lead to misinterpretation. The third step was designing a metamodel to describe the structure and concepts associated with data sets, and the relationships between each concept. Every concept in the metamodel is defined using standard terminology to allow common interpretation of data.

The fourth step involved developing a framework for data quality assessment to determine whether a data set contains sufficient information to facilitate the correct interpretation of data. The fifth step was constructing formal guidelines for the creation of good-quality data sets. The guidelines were constructed based on the essential terminology of the dataset metamodel and procedures from the framework. The final step
consisted of the evaluation of part of the framework by means of a user study. The part of the framework consists of definitions of elements in the dataset metamodel and formal definitions for data quality issues.

The research makes four significant contributions which are: (i) a systematic mapping study on data quality in software engineering. (ii) a dataset metamodel for describing the structure of data sets. (iii) a data quality assessment framework to better understand the quality of data sets. (iv) formal guidelines for creating good-quality data sets.
First and foremost, I would like to thank the Ministry of Higher Education Malaysia for providing a scholarship to enable me to pursue this doctoral study, and the Universiti Teknologi Mara for granting me a study leave.

My sincere gratitude to my supervisors: Associate Professor Ewan Tempero and Associate Professor Andrew Luxton-Reilly for their help, guidance, encouragement, support, friendship, patience and dedication during the last four years.

For their friendship and support, I would like to thank all my friends at the Department of Computer Science, in particular Safurah Abdul Jalil and Mozhgan Memari. Moreover, thank you to all my Malaysian friends in Auckland for their friendship, love and help for my family throughout our stay in the New Zealand.

Finally, I would like thank my family for providing me support and for coping with me throughout the time of this research.
# Table of Contents

## 1 Introduction

1.1 Background ......................................................... 1
1.2 Research motivation .............................................. 2
1.3 Importance of data quality in data sets ......................... 4
1.4 Research questions .............................................. 4
1.5 Research objective .............................................. 5
1.6 Research scope .................................................. 5
1.7 Research approach .............................................. 6
1.8 Research contributions .......................................... 7
1.9 Thesis organisation ............................................. 8

## 2 Understanding data sets

2.1 Overview of data sets ............................................ 11
   2.1.1 Data set contents ......................................... 12
   2.1.2 Data sets formats and structures ......................... 12
2.2 Examples of data sets ........................................... 13
   2.2.1 Real data sets ........................................... 13
   2.2.2 Artificial data sets ....................................... 20
   2.2.3 Discussion of examples of data sets ..................... 23
2.3 Data sets in software engineering .............................. 24
2.4 Software metrics tools .......................................... 25
2.5 Conclusion ...................................................... 26

## 3 A review of data quality research in software engineering

3.1 Introduction ...................................................... 28
3.2 Data quality research ........................................... 29
3.3 Systematic mapping study ....................................... 31
   3.3.1 Definition of Systematic Mapping Research Questions ... 32
   3.3.2 Conducting a search for relevant literature .............. 33
   3.3.3 Selection of studies ..................................... 36
   3.3.4 Developing the classification scheme ..................... 37
   3.3.5 Data Extraction and Mapping of Studies ................ 37
   3.3.6 Results of the systematic mapping ....................... 38
   3.3.7 Discussion of the Systematic Mapping Findings ........ 45
## TABLE OF CONTENTS

3.3.8 Implications for research ........................................ 47
3.4 Recently published data quality studies not included in the systematic mapping study .................................................. 47
3.5 Review of data quality issues ........................................ 49
  3.5.1 Definitions of data quality issues ............................. 50
  3.5.2 Examples of data quality issues in software engineering data sets ............................................................... 52
  3.5.3 Dimension of data quality issues ............................. 54
  3.5.4 Data quality studies ........................................... 57
  3.5.5 Dealing with poor quality of data sets ....................... 59
  3.5.6 Review of current data quality model in software engineering ................................................................. 61
3.6 Discussion and conclusions ........................................ 63

4 A metamodel for modelling data sets ................................ 66
  4.1 Introduction ..................................................... 66
  4.2 Related work on conceptual and ontological modelling of data ................................................................. 67
    4.2.1 Framework for software measurement validation .......... 68
    4.2.2 Model of software measurement data .......................... 68
    4.2.3 Model for software measurement (MOSME) .................. 69
    4.2.4 Ontology for software metrics and indicators ............. 69
    4.2.5 Software measurement ontology (SMO) ....................... 70
    4.2.6 KEEL (Knowledge Extraction based on Evolutionary Learning) tool .......................................................... 70
  4.3 Designing a dataset metamodel .................................... 71
    4.3.1 Definitions of measurement data set and measurement metadata ............................................................... 71
    4.3.2 Dataset metamodel ........................................... 72
    4.3.3 Extended data sets ........................................... 76
  4.4 Modelling a data set: The process ................................ 79
    4.4.1 Identifying the physical structure elements ............... 79
    4.4.2 Interpreting the physical structure elements ............. 81
  4.5 Modelling the existing data sets .................................. 83
    4.5.1 Example 1: PROMISE Boetticher ............................ 84
    4.5.2 Example 2: PROMISE Datatrieve ............................. 84
    4.5.3 Example 3: PROMISE KC2 .................................... 86
    4.5.4 Example 4: Qualitas Weka 3.7.5 ............................. 88
    4.5.5 Example 5: PROMISE Ant 1.3 ............................... 90
    4.5.6 Summary of results of modelling the data sets ............ 92
    4.5.7 Discussion of the results of modelling the data sets .... 94
  4.6 Conclusions .................................................... 94

5 A data quality assessment framework ................................ 96
  5.1 Introduction ..................................................... 96
  5.2 Formal definition of data quality issues ....................... 97
    5.2.1 Common data quality issues ............................... 98
    5.2.2 Metadata-related quality issues ............................ 100
5.3 Quality assessment .............................................. 102
  5.3.1 Assessing the data set .................................... 103
  5.3.2 Identifying data quality issues ......................... 105
  5.3.3 Evaluating the metadata of data sets ................. 107
  5.3.4 Preparing the assessment report ....................... 109
5.4 Results of Quality Assessment ............................... 109
  5.4.1 Summary report ....................................... 110
  5.4.2 Full report ........................................... 111
5.5 Applying quality assessment to the data sets ............. 112
  5.5.1 Discussion of the results for applying the quality assessment process 114
5.6 Conclusion .................................................. 115

6 Guidelines for creating good-quality data sets .................. 117
  6.1 Introduction ............................................. 117
  6.2 Constructing the guidelines ............................... 118
  6.3 The guidelines for creating data sets ..................... 119
  6.4 How to use the guidelines ................................ 120
    6.4.1 Using the guidelines while creating data sets ...... 121
    6.4.2 Using the guidelines after evaluating the quality of data set ... 122
  6.5 Conclusion ............................................... 128

7 A survey to evaluate the data quality framework ............... 130
  7.1 Introduction ............................................. 130
  7.2 Survey Design ............................................ 131
    7.2.1 Goal and survey research questions .................. 131
    7.2.2 Participants ........................................ 132
    7.2.3 Sampling methods .................................. 132
    7.2.4 Survey material .................................... 133
    7.2.5 Tasks ............................................ 133
  7.3 Survey Execution ........................................ 133
    7.3.1 Preparation ........................................ 133
    7.3.2 Procedure ......................................... 134
  7.4 Results and Analysis ..................................... 136
    7.4.1 Demographics ....................................... 136
    7.4.2 Dataset framework information ...................... 139
    7.4.3 Open-ended question ................................ 159
  7.5 Discussion of the findings of our survey ................... 159
    7.5.1 Analysis of the application of the definitions of dataset category elements to the data sets ........... 160
    7.5.2 Analysis of the time spent identifying the dataset category elements 162
    7.5.3 Analysis of the application of formal definitions of data quality issues to the data sets .................. 163
    7.5.4 Analysis of time spent identifying data quality issues ... 164
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5.5</td>
<td>Analysis of the open-ended questions about the framework</td>
<td>164</td>
</tr>
<tr>
<td>7.6</td>
<td>Improvement in the data quality framework</td>
<td>165</td>
</tr>
<tr>
<td>7.7</td>
<td>Threats to validity</td>
<td>166</td>
</tr>
<tr>
<td>7.7.1</td>
<td>Conclusion validity</td>
<td>166</td>
</tr>
<tr>
<td>7.7.2</td>
<td>Internal validity</td>
<td>167</td>
</tr>
<tr>
<td>7.7.3</td>
<td>Construct validity</td>
<td>167</td>
</tr>
<tr>
<td>7.7.4</td>
<td>External validity</td>
<td>168</td>
</tr>
<tr>
<td>7.8</td>
<td>Conclusion</td>
<td>169</td>
</tr>
<tr>
<td>8</td>
<td>An observational study to evaluate the data quality framework</td>
<td>171</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>171</td>
</tr>
<tr>
<td>8.2</td>
<td>Design of the study</td>
<td>172</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Goal and study research questions</td>
<td>172</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Participants</td>
<td>173</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Sampling</td>
<td>173</td>
</tr>
<tr>
<td>8.2.4</td>
<td>Materials</td>
<td>173</td>
</tr>
<tr>
<td>8.2.5</td>
<td>Tasks</td>
<td>175</td>
</tr>
<tr>
<td>8.2.6</td>
<td>Observational study hypotheses</td>
<td>175</td>
</tr>
<tr>
<td>8.3</td>
<td>Execution</td>
<td>176</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Preparation</td>
<td>176</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Procedure</td>
<td>176</td>
</tr>
<tr>
<td>8.4</td>
<td>Analysis</td>
<td>177</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Task list and Observation</td>
<td>177</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Demographics</td>
<td>182</td>
</tr>
<tr>
<td>8.4.3</td>
<td>Analysis of participants' application of the new definitions of dataset category elements according to participants' background</td>
<td>184</td>
</tr>
<tr>
<td>8.4.4</td>
<td>Analysis of participants' application of the new definitions of dataset category elements according to observation data</td>
<td>186</td>
</tr>
<tr>
<td>8.4.5</td>
<td>Analysis of participants' application of the new definitions of dataset category elements according to experience in analysing datasets</td>
<td>187</td>
</tr>
<tr>
<td>8.4.6</td>
<td>Analysis of participants' application of the formal definitions of quality issues according to participants' background</td>
<td>189</td>
</tr>
<tr>
<td>8.4.7</td>
<td>Analysis of participants' application of the formal definitions of data quality issues according to observation data</td>
<td>190</td>
</tr>
<tr>
<td>8.4.8</td>
<td>Analysis of participants' application of the formal definitions of data quality issues according to experience in analysing datasets</td>
<td>192</td>
</tr>
<tr>
<td>8.4.9</td>
<td>Analysis of participants' evaluation of the metadata in datasets according to participants' background</td>
<td>193</td>
</tr>
<tr>
<td>8.4.10</td>
<td>Analysis of participants' evaluation of the metadata in datasets according to observation data</td>
<td>194</td>
</tr>
<tr>
<td>8.4.11</td>
<td>Analysis of participants' evaluation of the metadata in datasets according to experience in analysing datasets</td>
<td>196</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8.4.12</td>
<td>Correctness of application of the new definitions of dataset category elements</td>
<td>197</td>
</tr>
<tr>
<td>8.4.13</td>
<td>Correctness of application of the formal definitions of data quality issues</td>
<td>198</td>
</tr>
<tr>
<td>8.4.14</td>
<td>Correctness of identifying and evaluating the metadata in datasets</td>
<td>200</td>
</tr>
<tr>
<td>8.4.15</td>
<td>Discussion of the findings of our observational study</td>
<td>202</td>
</tr>
<tr>
<td>8.5</td>
<td>Threats to validity</td>
<td>204</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Conclusion validity</td>
<td>204</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Internal validity</td>
<td>205</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Construct validity</td>
<td>205</td>
</tr>
<tr>
<td>8.5.4</td>
<td>External validity</td>
<td>205</td>
</tr>
<tr>
<td>8.6</td>
<td>Conclusions</td>
<td>206</td>
</tr>
<tr>
<td>9</td>
<td>Discussion</td>
<td>207</td>
</tr>
<tr>
<td>9.1</td>
<td>Revisiting the research questions</td>
<td>207</td>
</tr>
<tr>
<td>9.2</td>
<td>Evaluating the use of the dataset metamodel</td>
<td>209</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Application of dataset metamodel</td>
<td>210</td>
</tr>
<tr>
<td>9.3</td>
<td>Evaluating the use of the quality assessment process</td>
<td>211</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Application of quality assessment process</td>
<td>212</td>
</tr>
<tr>
<td>9.4</td>
<td>Evaluating part of the data quality framework</td>
<td>213</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Observations from the analysis of survey</td>
<td>213</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Observations from the analysis of observational study</td>
<td>215</td>
</tr>
<tr>
<td>9.5</td>
<td>Implications</td>
<td>215</td>
</tr>
<tr>
<td>9.5.1</td>
<td>The state of quality for data sets in software engineering</td>
<td>215</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Why should researchers care about data quality in software engineering?</td>
<td>216</td>
</tr>
<tr>
<td>9.5.3</td>
<td>How researchers use data sets in software engineering?</td>
<td>217</td>
</tr>
<tr>
<td>9.5.4</td>
<td>Interpreting threats to validity and data quality</td>
<td>218</td>
</tr>
<tr>
<td>9.6</td>
<td>Limitations</td>
<td>219</td>
</tr>
<tr>
<td>9.6.1</td>
<td>Limited to software engineering data sets</td>
<td>219</td>
</tr>
<tr>
<td>9.6.2</td>
<td>Manual modelling of data sets</td>
<td>219</td>
</tr>
<tr>
<td>9.6.3</td>
<td>Sample size of survey</td>
<td>219</td>
</tr>
<tr>
<td>9.6.4</td>
<td>Minimal guideline in the observational study</td>
<td>219</td>
</tr>
<tr>
<td>9.6.5</td>
<td>Incomplete evaluation of framework</td>
<td>220</td>
</tr>
<tr>
<td>9.7</td>
<td>Improvements on survey and observational study</td>
<td>220</td>
</tr>
<tr>
<td>9.8</td>
<td>Conclusion</td>
<td>221</td>
</tr>
<tr>
<td>10</td>
<td>Conclusions &amp; Future work</td>
<td>222</td>
</tr>
<tr>
<td>10.1</td>
<td>Research Summary</td>
<td>222</td>
</tr>
<tr>
<td>10.1.1</td>
<td>A review of data quality research in software engineering</td>
<td>223</td>
</tr>
<tr>
<td>10.1.2</td>
<td>Examining the content of data sets</td>
<td>224</td>
</tr>
<tr>
<td>10.1.3</td>
<td>Designing a dataset metamodel</td>
<td>224</td>
</tr>
<tr>
<td>10.1.4</td>
<td>Developing a data quality assessment framework</td>
<td>225</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>10.1.5 Constructing guidelines for creating good-quality data sets</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>10.1.6 Performing a user evaluation of a data quality framework</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>10.2 Research Contributions</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>10.3 Future Work</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>10.3.1 Evaluating other types of data sets</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>10.3.2 Automated modelling of the data sets</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>10.3.3 To unify dataset terminology into an ontology</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>10.3.4 To improve and refine the quality assessment process in the data quality framework</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>10.4 Final Remarks</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>A Appendix A</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>A.1 Summary of search results from online databases</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>A.2 List of included studies in systematic mapping</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>A.3 Data extracted from S7</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>B Appendix B</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>B.1 Dataset metamodel</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>C Appendix C</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>C.1 Formal procedures for modelling data set</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>D Appendix D</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td>D.1 Matrix for applying the modelling data sets process</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td>E Appendix E</td>
<td>251</td>
<td></td>
</tr>
<tr>
<td>E.1 Formal procedures for quality assessment</td>
<td>251</td>
<td></td>
</tr>
<tr>
<td>F Appendix F</td>
<td>254</td>
<td></td>
</tr>
<tr>
<td>F.1 Quality assessment for Ant 1.3 data set</td>
<td>254</td>
<td></td>
</tr>
<tr>
<td>G Appendix G</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>G.1 Summary report for Data Quality Assessment Framework</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>G.2 Full report for Data Quality Assessment Framework</td>
<td>258</td>
<td></td>
</tr>
<tr>
<td>G.3 Matrix for applying the Quality Assessment Process</td>
<td>260</td>
<td></td>
</tr>
<tr>
<td>H Appendix H</td>
<td>262</td>
<td></td>
</tr>
<tr>
<td>H.1 Full report for Example 1: PROMISE Boetticher dataset</td>
<td>262</td>
<td></td>
</tr>
<tr>
<td>H.2 Full report for Example 2: PROMISE Datatrieve dataset</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>I Appendix I</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>I.1 Participant Information Sheet</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>I.2 Advertisement</td>
<td>270</td>
<td></td>
</tr>
<tr>
<td>J Appendix J</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>J.1 Dataset Quality Assessment Survey</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>J.2</td>
<td>Ethics application for an online survey</td>
<td>288</td>
</tr>
<tr>
<td>K</td>
<td><strong>Appendix K</strong></td>
<td>307</td>
</tr>
<tr>
<td>K.1</td>
<td>Participants responses of correct answers for the definitions of elements in three data sets in the survey</td>
<td>307</td>
</tr>
<tr>
<td>L</td>
<td><strong>Appendix L</strong></td>
<td>308</td>
</tr>
<tr>
<td>L.1</td>
<td>Survey: Evaluation of data quality framework for software engineering data sets</td>
<td>308</td>
</tr>
<tr>
<td>L.2</td>
<td>The observation schedule</td>
<td>312</td>
</tr>
<tr>
<td>L.3</td>
<td>Ethics application for an observational study</td>
<td>314</td>
</tr>
<tr>
<td>M</td>
<td><strong>Appendix M</strong></td>
<td>357</td>
</tr>
<tr>
<td>M.1</td>
<td>Participant Information Sheet and Consent Form for participant</td>
<td>357</td>
</tr>
<tr>
<td>M.2</td>
<td>Participant Information Sheet and Consent Form for Organisation</td>
<td>361</td>
</tr>
<tr>
<td>M.3</td>
<td>Advertisement</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>367</td>
</tr>
<tr>
<td>Abbreviation (SLR)</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>SLR</td>
<td>Systematic literature review</td>
<td></td>
</tr>
<tr>
<td>ISBSG</td>
<td>International Software Benchmarking Standards Group</td>
<td></td>
</tr>
<tr>
<td>AQD</td>
<td>Automated quality detection</td>
<td></td>
</tr>
<tr>
<td>ARFF</td>
<td>Attribute-relation file format</td>
<td></td>
</tr>
<tr>
<td>CK</td>
<td>Chidamber and Kemerer</td>
<td></td>
</tr>
<tr>
<td>CSV</td>
<td>Comma separated values</td>
<td></td>
</tr>
<tr>
<td>CVS</td>
<td>Concurrent Version System</td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>DCTF</td>
<td>Data cleaning techniques —filtering</td>
<td></td>
</tr>
<tr>
<td>DCTP</td>
<td>Data cleaning techniques —polishing</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td>Function points</td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>Lines of code</td>
<td></td>
</tr>
<tr>
<td>MDP</td>
<td>Metrics Data Program</td>
<td></td>
</tr>
<tr>
<td>MQD</td>
<td>Manual quality detection</td>
<td></td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
<td></td>
</tr>
<tr>
<td>PICOC</td>
<td>Population, intervention, comparison, outcomes and context</td>
<td></td>
</tr>
<tr>
<td>PROMISE</td>
<td>Predictor models in software engineering</td>
<td></td>
</tr>
<tr>
<td>RQ</td>
<td>Research question</td>
<td></td>
</tr>
<tr>
<td>RTDQ</td>
<td>Robust technique with respect to data quality</td>
<td></td>
</tr>
<tr>
<td>SAT</td>
<td>Special analysis techniques</td>
<td></td>
</tr>
<tr>
<td>SIR</td>
<td>Software Artifact Infrastructure Repository</td>
<td></td>
</tr>
<tr>
<td>PIS</td>
<td>Participant information sheet</td>
<td></td>
</tr>
<tr>
<td>UAHPEC</td>
<td>University of Auckland Human Participants Ethics Committee</td>
<td></td>
</tr>
<tr>
<td>UCI</td>
<td>University of California, Irvine</td>
<td></td>
</tr>
<tr>
<td>SI</td>
<td>International System of Units</td>
<td></td>
</tr>
<tr>
<td>SER</td>
<td>Software Engineering Researcher</td>
<td></td>
</tr>
<tr>
<td>OQ</td>
<td>Observation Question</td>
<td></td>
</tr>
<tr>
<td>ORQ</td>
<td>Observation Research Question</td>
<td></td>
</tr>
<tr>
<td>UiTM</td>
<td>Universiti Teknologi MARA</td>
<td></td>
</tr>
</tbody>
</table>
Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Title: Can we trust our results? A mapping study on data quality. Published in Asia Pacific Software Engineering Conference, 2013.

Parts of this publication are reported in Chapter 3/Section 3.3 Systematic mapping study/ pages 29-45 and Chapter 7/Section 7.4 How to use guidelines/ pages 124-125.

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>The PhD candidate performed the systematic mapping study, analysed the data, wrote the text and created figures for this publication.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>85%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ewan Tempero</td>
<td>Provided conceptual ideas, suggestions for the interpretation of results and editing.</td>
</tr>
<tr>
<td>Andrew Luxton-Reilly</td>
<td>Provided conceptual ideas, suggestions for the interpretation of results and editing.</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ewan Tempero</td>
<td></td>
<td>17 November 2016</td>
</tr>
<tr>
<td>Andrew Luxton-Reilly</td>
<td></td>
<td>17 November 2016</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

**Title:** What is in a data set? Describing a structure of data sets. Published in Australasian Computer Science Conference, 2016.

Parts of this publication are reported in Chapter 2/ Section 2.2 Examples of data sets/ pages 11-22 and Section Chapter 5/ Section 5.3 Designing a dataset metamodel/ pages 75-83, Section 5.5 Modelling the existing data sets/pages 87-99.

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>The PhD candidate designed the metamodel, analysed the data sets, wrote the text and created figures for this publication.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>80%</td>
</tr>
</tbody>
</table>

### CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ewan Tempero</td>
<td>Provided conceptual ideas to improve the metamodel, provided suggestions for the interpretation of evaluation and editing.</td>
</tr>
<tr>
<td>Andrew Luxton-Reilly</td>
<td>Provided conceptual ideas to improve the metamodel, provided suggestions for the interpretation of evaluation and editing.</td>
</tr>
</tbody>
</table>

### Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ewan Tempero</td>
<td>[Signature]</td>
<td>17 November 2016</td>
</tr>
<tr>
<td>Andrew Luxton-Reilly</td>
<td>[Signature]</td>
<td>17 November 2016</td>
</tr>
</tbody>
</table>

*Last updated: 19 October 2015*
This chapter provides an overview of the research in this thesis. It outlines the background of the research area, followed by the motivation and main objectives of the research. The problem to be investigated is then presented, followed by the expected contributions of this research. The chapter ends with an overview of the thesis structure.

1.1 Background

Research indicates that the quality of data sets is critical to the results of empirical studies [1–3]. Although data sets play a central role in empirical research, few studies consider the quality of the data sets used [4–7]. If the quality of the data is poor, then the results of empirical studies cannot be trusted, and any models or conclusions based on the data sets are questionable. Ensuring data quality is therefore fundamentally important to the field of empirical software engineering.

Recently, a growing body of literature has questioned data quality in software engineering data sets. A number of researchers have reported data quality issues such as noise, missing values, duplicate data and incorrect data [8–11]. However, these issues have not received as much attention as that given to the development of prediction models [11–13].

Data quality is important because poor data quality may lead to invalid conclusions. This issue has been mentioned in previous studies, but there have been small number of serious discussions on the validity of results [2, 8, 9]. For example, Gray et al. [8] demonstrated that empirical studies based on data sets that contain duplicate data may lead to erroneous results. Shepperd et al. [9] confirmed this finding by investigating data integrity and inconsistencies between different versions of defect data sets from the National Aeronautics and Space Administration (NASA) Metrics Data Program (MDP). Gray et al. and Shepperd et al. urged research communities to take action to establish good-quality data sets in order to ensure the validity of results.
To conduct research with good-quality data sets, we need a way to evaluate the quality of the data sets. A number of researchers have proposed techniques for dealing with data quality issues in data sets [4, 5, 14, 15]. For example, Mockus [5] and Zhang [4] proposed imputation techniques to handle missing data problems. Further, Kim et al. [14] and Khoshgoftaar et al. [15] developed noise identification and filtering techniques to improve data quality. Although many techniques have been proposed to resolve the challenges associated with data quality, there is a lack of research that evaluates the quality of data sets—particularly those from public data repositories.

Recently, there has been a move towards creating public data repositories to provide access and make data sets available to the research community—for example, NASA MDP and Predictor Models in Software Engineering (PROMISE) [16]. These data repositories contain collections of data sets from various sources (e.g., research, open source projects and organisations). Data sets from data repositories are designed individually with specific definitions and structures by creators of the data sets, and researchers intending to use data sets from public data repositories rely on their quality. For example, Zhang-[17] conducted an experiment using data sets from Eclipse and NASA public repository and made no attempt to identify quality issues in the data sets.

Ideally, the creators of data sets should know about the quality of the data. The creators of data sets should document a complete information of data sets including the known quality issues when sharing their data sets in the public data repositories. This information is important for researchers who are intend to use the data sets. If researchers are not informed about the quality of data, then they need to conduct a pre-processing analysis to find the quality of data. In some cases, researchers often conduct their own analysis of the quality of data sets within data repositories. For example, Jiang et al. [6] used three NASA MDP data sets (CM1, JM1 and PC1) stored in the PROMISE repository to build defect prediction models. They analysed the three data sets and suspected that some of the attributes had noisy data. Thus, they excluded the noisy data and used subsets of the three data sets for their prediction models.

Before using data sets from data repositories, researchers need to determine the quality of the data that form the basis of the results. No studies have created a standard way of assessing the quality of data sets in data repositories. Therefore, this study aims to find a standard way to help researchers determine good-quality data sets for analysis in empirical research.

1.2 Research motivation

Public data repositories such as NASA MDP [8], PROMISE [16], Software Artifact Infrastructure Repository (SIR) [18], Qualitas Corpus [19] and The International Software Benchmarking Standards Group (ISBSG) [20] contain a large number of data sets for research purposes. For some data repositories, known quality issues are not clearly documented in the existing data sets [8]. As a result, inexperienced researchers may assume that the available data sets are of reasonable quality for analysis in empirical
As a result of the growing importance of describing information about data sets, the ISBSG established standard data collection procedures for software projects [20]. These procedures capture various characteristics of software projects, including the project team, effort per development phase, and methods and techniques. A large amount of data can be collected, but only a few of these data are mandatory [22]. This can cause missing data in some of important variables of software projects use to build software model [23]. When there is a large amount of missing data in data sets, this might cause insufficient information about the software projects which may lead to a problem with the interpretation of the data.

If data sets are poorly documented, the meaning of the data may be lost, and researchers may not be aware of this problem. For example, there are 37,699 lines of code (LOC) in the Ant 1.3 data set, as reported in PROMISE repository [24], and there are 30,445 LOC in Qualitas Corpus [19] repository. Each repository shows a different value for the LOC, even though both values refer to the same entity. In this case, it is unclear which of these LOC values should be used when analysing the Ant 1.3 data set.

There are at least two possible reasons to explain the different values for the LOC of the Ant 1.3 data set. First, each repository may use a different method to calculate the LOC of the entity, despite using the same label for a metric. Qualitas Corpus defines LOC as the number of physical LOC, whereas PROMISE uses the same name but calculates the LOC value as the number of LOC in the Java binary code. This shows that PROMISE has a different interpretation of LOC compared with Qualitas Corpus.

Second, each repository may contain a different number of files, which influences the LOC of the entity. In this case, only Qualitas Corpus identifies the files counted in the Ant 1.3 data set. PROMISE does not state the value for the number of files, so it is possible that the difference in the LOC values is because different files were measured. This misinterpretation problem does not occur if both repositories explain the contents of the data set. This information is called ‘data about data’ or ‘metadata’.

Metadata can be defined as additional information that describes the purpose, meaning and context of the data in the data sets. Metadata enable researchers to better understand the data and reproduce results. More importantly, appropriate metadata allow researchers to avoid misinterpretation problems such as the one identified in the example of Ant 1.3 data set above.

Although researchers have been sharing and making their data sets available in data repositories for quite some time, a standard way of describing data sets is yet to be established. Researchers describe information about data sets in data repositories in many different ways [8, 16] —particularly in regard to the structure of data and metadata. This creates challenges with the interpretation of data sets that make it difficult to clearly report the quality of data.

The overall motivation of this research is to help researchers better understand and interpret data sets, and to identify the potential problems a data set may have. The focus of this research has led to the design and development of a framework that incorporates
1.3 Importance of data quality in data sets

Research has shown that data quality issues exist in data sets from public data repositories [8–10]. For example, duplicate data occurs when a record with the same values for the same entity is entered multiple times by accident. If there are duplicate data, the aggregate count changes for the distribution of data in the data set.

Another data quality issue is inconsistent data, which occurs when a record with different values for the same entity is entered multiple times by accident. This issue may result from mistakes made by people when using a software analysis tool to generate data sets. This quality issue creates questionable records in a data set because it is difficult to determine which record is correct.

The existence of quality issues in data sets may affect empirical research. If researchers choose incorrect records, it may result in biased data that affected the results of prediction models or classifiers. If the results of prediction models or classifiers are affected, it may lead to invalid conclusions in the empirical research.

Further, in some data sets from public data repositories, the data may not be able to be interpreted correctly. This issue arises due to insufficient information in the data sets, which may lead to misinterpreting the data and reaching incorrect conclusions. For example, the measurement for LOC in the Ant 1.3 data set could be misinterpreted because the meaning of LOC is not explicitly described and the number of files being measured is not stated in the PROMISE repository. Thus, data sets that contain insufficient metadata may lead to incorrect conclusions if researchers misinterpret the data.

1.4 Research questions

As identified in the previous sections, the domain of empirical software engineering needs a standardised way of modelling data that can be used to describe the structure of a data set and provide a better understanding of its quality.

Therefore, the main research question in this research can be framed as:

‘How can we help researchers to understand the quality of data sets in software engineering?’

We divide the main research question into subordinate research questions (SRQ) to determine approaches to answer the main research question:

- SRQ1. To what extent does existing research take into consideration the quality of data sets in software engineering?

To answer this question, we reviewed related research on data quality topics and problems that have been addressed, the techniques used to manage quality in software
The main objective of this research is to provide a framework to evaluate the quality of data sets in order to help researchers better interpret data sets and identify potential problems. In particular, this research aims:

• RO1. To review related research on data quality problems that have been addressed, the techniques used to manage quality in data sets, the type of data sets used for analysis in empirical research and to investigate to what extent does poor data quality affect the empirical results.

• RO2. To examine the different formats and structures of existing data sets in data repositories and to identify the aspects of data sets necessary for developing a precise means to describe data sets.

• RO3. To develop a data quality framework for describing the content of data sets and evaluating their quality

• RO4. To evaluate the usability and effectiveness of the proposed framework

• RO5. To provide guidelines for creating good-quality data sets using the framework.

1.6 Research scope

The term ‘data sets’ used throughout this research refers to software engineering data sets and not those related to database data sets. This research focuses on data sets for analysis in empirical software engineering —particularly those obtained from public data
repositories. As mentioned earlier, some data sets from public data repositories cannot be interpreted correctly because insufficient information is provided. This issue may result in misinterpreting the intended meaning of data sets and reaching invalid conclusions.

With regard to the misinterpretation issue, this research will focus on a specific dimension of data quality from the area of information system—namely, interpretability. This dimension concerns the documentation and metadata that are available to correctly interpret the data sets [25].

This research will also consider four data quality dimensions from the area of information system for assessing the quality of data sets. These dimensions are accuracy, completeness, consistency and uniqueness [25, 26]. We selected the four dimensions because real data sets from public data repositories contain quality issues that correspond to these dimensions [8, 10, 11]. Others dimensions of data quality, such as timeliness, relevance and currency, are beyond the scope of this research.

1.7 Research approach

The research process consists of three stages to address the research questions and achieve the research objectives. These are illustrated in Figure 1.1.

The first stage of the research addressed SRQ1 and SRQ2, and to achieve RO1 and RO2. To answer SRQ1 and achieve RO1, we conducted a literature review on data quality in software engineering to obtain an understanding of the current state of the existing research and to investigate the extent to which the research community has addressed data quality. We started the review by conducting a systematic mapping study to provide an overview of recent research (2008-2012). We further reviewed the literature that focused on definitions and examples of data quality issues. We also reviewed the literature on data quality studies that were concerned with research results and techniques or methods used to conduct research with poor-quality data sets.

To answer SRQ2 and achieve RO2, we conducted an observation of both real and artificial data sets to understand what the data sets represent and how they are organised. We started the observation by examining the content, formats and structures of real data sets from public data repositories. We also examined the formats and structures of artificial data sets to show the possible ways that data may be presented and also to consider them when developing a standard way to understand data sets.

The second stage of the research aim to answer SRQ3 and achieve RO3. In this stage, we first constructed precise definitions for data sets and their potential elements of data sets. These potential elements were identified based on the observation of real data sets in the first stage, which enabled us to explore the formats and structures of data sets, and identify common aspects. The challenges we experienced in constructing precise definitions for elements in data sets motivated us to design a dataset metamodel. We designed the metamodel to describe the structure and concepts associated with data sets, and the relationships between each concept.

We also developed a data quality framework to evaluate the quality of data sets. The
aim of the data quality framework is to determine whether a data set contains sufficient information to facilitate the correct interpretation of data for analysis in empirical research. The framework is divided into two main processes: the modelling of data sets and a quality assessment process. The first process of the framework is to model the data set using a formal procedure based on the dataset metamodel. The second process of the framework is to assess the quality of the data sets, using a quality assessment process.

The third stage of the research aims to answer SRQ3, and achieve RO4 and RO5. We constructed a guideline for creating good-quality data sets. We started by compiling a list of the best practices reported throughout the thesis to provide a useful approach to creating a data set. In particular, some guidelines were developed from the dataset metamodel, which focused on the essential elements of data sets and the relationships between elements. We also developed some guidelines from the quality assessment process, which focused on the evaluation scale to assess the metadata in data sets.

We performed a user evaluation of the framework to evaluate the effectiveness and usability of the framework. We conducted a user evaluation to assess the effectiveness and usability of part of our framework, in particular, the definitions of elements in the dataset metamodel and the formal definitions of data quality issues. Finally, we derived conclusions from our literature review, observations of data sets, and evaluation.

Figure 1.1: Research approach

1.8 Research contributions

The research discussed in this thesis contributes to the field of software engineering—particularly in the area of data quality. The main research contributions are as follows:
1. As part of this research, we conducted a systematic mapping study to determine the current state of existing research on data quality and, most importantly, to discover gaps in the empirical research. The findings of the mapping study provide insights for us to conduct further review on data quality issues in empirical research. As a first step towards this insight, we developed a model of what happens to data, which we call the research data life cycle. This model describes the stages a data set may go through, and it can be used to identify when problems with data quality may be introduced. A conference paper that describes this mapping study, entitled ‘Can we trust our results? A systematic mapping of data quality’, was published in the Proceedings of the Asia Pacific Software Engineering Conference, 2013 [27].

2. This research designed a dataset metamodel to describe the structure and concepts in a data set, as well as the relationships between the concepts. It includes precise definitions of data sets and their potential elements. We anticipate that the definitions and the metamodel will enable researchers to understand what data sets intend to represent, and to identify whether data sets have sufficient information to be usable for analysis in empirical research. A conference paper that describes this metamodel, entitled ‘What is in a data set? Describing a structure of data sets’, was published in the Proceedings of the Australasian Computer Science Conference, 2016 [28].

3. This research developed a framework to evaluate the quality of data sets in software engineering. The framework consisted of the dataset metamodel and a quality assessment process. The quality assessment process introduced two important components: a formal definition for data quality issues and an evaluation scale of metadata. The formal definition for data quality issues was constructed using the standard terminology from the data set metamodel to specify the structure of data sets that have quality issues. The evaluation scale of metadata was constructed to help researchers assess the completeness and accessibility of metadata in data sets. We presented the initial work on the framework for data quality assessment as a short conference paper entitled ‘What is in a data set? Describing a structure of data sets for data quality assessment’ at the 13th New Zealand Computer Science Research Student Conference, 2015.

4. This research constructed guidelines for creating good-quality data sets. The guidelines compiled the best practices reported throughout the thesis to provide a useful approach for creating a data set from scratch using the essential elements in the data set metamodel and specifying the relationships between the elements.

1.9 Thesis organisation

The remainder of this thesis is organised as follows:

- Chapter 2 presents information about data sets and focuses on understanding what data sets represent and how they are organised. It presents examples of real data
sets from data repositories and artificial data sets. These examples illustrate the different structures and contents of data sets, as well as potential data quality issues. The chapter also describes the context of data sets in software engineering.

- Chapter 3 presents a review of data quality in empirical software engineering. The chapter starts with systematic literature reviews reported in the literature, followed by systematic mapping study methods carried out as the foundation of this research. The results and findings obtained from the mapping study highlight the gaps in the literature. This research aims to address these gaps by conducting a further review of data quality issues in software engineering data sets. This is discussed with an emphasis on the definitions of data quality issues, examples of the issues and how to deal with them.

- Chapter 4 introduces a dataset metamodel that provides a standard way to describe the structures and concepts in a data set. The chapter presents the definitions for measurement data sets, measurement metadata and elements of data sets. It illustrates the relationships between elements in the data set through the metamodel and introduces the notion of extended data sets to identify all elements of data sets in all locations.

- Chapter 5 presents a data quality framework for evaluating the quality of data sets, including the dataset metamodel and a quality assessment process. It details the relevant tasks in the quality assessment process. The chapter also introduces formal definitions for data quality issues and the evaluation scale of metadata to facilitate the quality assessment process in the framework.

- Chapter 6 presents guidelines that contain best practices for creating good-quality data sets. It describes the components used to construct the guidelines and also describes how to apply the guidelines to data sets.

- Chapter 7 presents a survey to evaluate part of the framework. The chapter explains the design of the survey carried out to evaluate the effectiveness and usability of the part of the framework that includes definitions of elements in a data set and formal definitions for data quality issues. The chapter also presents the results and discusses the findings of the survey.

- Chapter 8 presents an observational study to evaluate part of the framework. The chapter explains the design of the observational study carried out to evaluate the effectiveness of application of new definitions of elements in a data set and usability of the formal definitions of data quality issues. The chapter also presents the results and discusses the findings of the observational study.

- Chapter 9 presents a discussion of our research findings. It discusses the evidence gathered in our research work and interesting observations from the research findings. Further, the chapter discusses implications for future research, limitations of this research and improvements for the survey.
Chapter 10 presents a summary of the thesis and the contributions it makes to existing research in this area. The chapter also suggests areas for future research that could be conducted to extend this research.
This chapter presents information about software engineering data sets. It is essential to understand what data sets represent and how they are organised in order to measure the quality of the data. This chapter describes a variety of formats and structures of real data sets from public data repositories, and it presents observations of potential data quality issues in data sets. It also discusses the formats and structures of artificial data sets, as well as some issues that may not be observed in data repositories. Further, this chapter discusses how to model some aspects of data sets based on their structure. The chapter concludes with a justification for standardising how data sets are modelled.

### 2.1 Overview of data sets

Data sets contain a collection of data and additional information for a specific purpose. The term data typically refers to measurements or observations that are recorded to represent the results of research activities (e.g., measuring or observing). As mentioned in Chapter 1, the term additional information refers to metadata that describe the purpose, meaning and context of data.

Metadata describe the intended meaning of the data and provide a context for interpreting the data. The context makes the data meaningful and useful. Such contexts include the unit and scale types to measure the data. For example, the units to measure temperature are Fahrenheit, Celsius and Kelvin, and the scale type for the Kelvin unit is a ratio scale.

Metadata enable researchers to make decisions regarding whether data sets are appropriate for the required purpose. Even if it does not contain explicit detailed information, it contains foundation information that is needed to understand common assumptions about the data set. This indicates that data sets are comprehensible and can be interpreted correctly.

Data sets are created for various purposes. In science, researchers use data sets to
support scientific activities such as testing hypotheses and explaining new ideas. In general, researchers use data sets as an input to analysis to solve problems—in particular, to build knowledge in a specific research area. For example, researchers use a defect data set as an input to defect prediction models to determine the fault-prone classification for software improvement.

To use data sets for analysis, researchers need to understand what is in the data sets and the conventions supporting the representation of data sets. The contents of data sets, along with their formats and structures, are discussed in the next sections.

### 2.1.1 Data set contents

Researchers use many terms to describe the content of data sets, including ‘values’, ‘measurements’, ‘observations’, ‘facts’ and ‘record of values’ [29]. Some terms are used to indicate the data, such as ‘values’ and ‘measurements’. Other terms are used to indicate the representation of data, such as ‘record of values’.

As mentioned earlier, metadata provide context to understand the data. The meaning of the data depends on the context that describes how the data sets are structured and what they represent. The context also includes conventions of representation that are applied when data sets are created. These conventions help to convey the meaning of the data. For example, measurements (e.g., 180 cm) that consist of integers and the units of measurement represent characteristics such as length. These measurements can be understood in relation to a system of measurement such as the International System of Units [30].

### 2.1.2 Data sets formats and structures

As mentioned in Chapter 1, data sets come from various sources such as research, open source projects and organisations. These sources publish their data sets using many different formats. Different formats use different ways to represent the structure of data.

In general, four types of format are used to represent the structure of data: comma-separated values (CSV), Excel, attribute-relation file format (ARFF) and plain text. CSV and Excel present the structure of the data in a tabular format. In CSV format, a comma or a tab is used to separate the values in the rows, whereas Excel uses columns and rows to separate values. ARFF structures data in a format that consists of two sections: header and data. The header section contains the name of the data set, a list of properties of entities and their data types. The data section contains the data declaration and values. Plain text format is different from the other formats, as the structure of the data can be represented in a variety of ways. For example, a data set contains only values that are organised in different rows, and each row contains values organised in many different ways.

Many data sets are presented in tabular format. This tabular format is a simple structure for representing data in rectangular tables made up of rows and columns. The columns are almost always labelled and used to distinguish different properties of entities.
2.2. Examples of data sets

This section briefly discusses 10 examples of data sets, including five real data sets from data repositories and five artificial data sets. The real data sets are selected based on the common formats of data sets that appear in data repositories. We created the artificial data sets with reasonable structures to demonstrate certain conditions that we anticipated might appear in the real data sets.

2.2.1 Real data sets

Real data sets from data repositories are created for various purposes by researchers, and they are presented in many formats and structures. For example, some contain metadata that correspond to measurements at external locations, such as web pages and external files.

This section presents four examples of data sets from the PROMISE repository and one example from Qualitas Corpus. It describes the most common formats of data sets to observe the variety of structures that exist in data repositories. We identify some aspects of data and discuss assumptions on how to interpret them based on the structures. In addition, we describe some issues related to data quality that exist in data sets.

2.2.1.1 Example 1: PROMISE Boetticher

Figure 2.1 illustrates part of the Boetticher data set [16], which presents survey data in a typical tabular format. In this example, we assume that each row corresponds to values for the properties of the entity. We also assume that the entity is a response from a respondent who participated in the survey. The rows are not labelled and there are no metadata in the data set. However, there are metadata on the web page where the data set originated.

The web page of the data set provides context information that includes the publication using the data set. The publication contains a general discussion of the demographic data of the data set. However, the discussion does not explicitly describe the actual column names used in the data set. This shows that the publication provides some metadata, but it does not clearly describe the necessary metadata. Further, it requires significant effort to find the metadata required in order to understand the data set. For example, we need to read the publication and identify the information therein that describes the necessary metadata.

In this example, we need to determine what the column headers are intended to represent. Generally, we assume that all column headers represent different properties of the entities, and that some properties are metrics while others are not. This is because only some of them look like metrics that measure the attributes of the entities and are
2.2. EXAMPLES OF DATA SETS

associated with values that are integers. This assumption indicates that some of the values associated with the properties are measurements for the metrics.

As mentioned earlier, every row contains values corresponding to different properties of the entities, except the column headers. We assume that each row contains information about a particular entity because each row has a different list of values. This assumption indicates that there is a need to find a way to describe the representation of values for a particular entity.

```
Gender Highest Degree Numeric Degree Comp Sc Undergrad Courses Comp Sc Grad Courses Hardware Undergrad Courses Hardware Grad Courses MIS Undergrad Courses MIS Grad Courses Proj Mgmt Undergrad Courses Proj Mgmt Grad Courses
Male Bachelor 3 0 0 6 0 0 0 0 4 0
Male Master 4 0 1 0 1 0 1 0 0 0
Male Bachelor 3 0 0 0 0 0 0 0 0 0
Male Master 4 3 6 2 0 2 6 10 6
Male Bachelor 3 0 0 0 0 0 0 0 0 20
Male Master 4 0 1 0 3 0 0 0 0 5
Male HS 2 0 0 0 0 0 0 0 6 0
Male Bachelor 3 2 0 3 0 0 1 0 3
Male Bachelor 3 1 2 6 2 1 1 4 2
Male Master 4 4 2 0 0 0 0 0 2 0
Male Master 4 0 1 0 0 0 0 0 1 0
Male Bachelor 3 0 0 0 0 0 0 0 0 0
Male Phd 5 1 0 0 0 0 0 0 0
Female Bachelor 4 0 0 0 0 0 0 0 0
Male Master 4 2 6 3 5 1 2 2 3
```

Figure 2.1: PROMISE Boetticher data set.

2.2.1.2 Example 2: PROMISE Datatrieve

Figure 2.2 illustrates part of the Datatrieve data set [57], which contains defect data presented in ARFF. There are no columns in this example, and the data are organised differently to Example 1. Rows contain the properties of the entities and values, and there are some metadata in the data set and on the web page of the data set. The entities are assumed to be source code files based on the metadata.

The elements after the heading @attribute appear to be similar to the column header in the tabular data. For example, LOC6_0 looks like a column header that represents a property corresponding to values in the rows after the heading @data, and the numeric looks like a data type for LOC6_0. Thus, it is assumed that every element after the heading @attribute represents a property of the entity, and that the element after the property represents the data type for the correspond value (e.g., numeric). We assume the data type as metadata because it indicates the type of values for the correspond value.

In this example, we assume that some of the properties are metrics because they look like attributes of entities and are associated with values that are integers. We also assume that the values associated with metrics are measurement values because they look like measurements.

In Figure 2.2, some elements appear to be headings for the name of the data set, properties and values. To describe this type of data set, we need to capture the following elements: @relation, @attribute and @data. These elements indicate something relating to the structure separation for elements corresponding to the measurements. This
indicates that we need to determine how to describe elements that provide information about the structure.

We observe that after the heading \texttt{@data}, every row contains a list of values separated by commas. We assume that these values correspond to different properties of the entities, which are organised in rows after the heading \texttt{@attribute}. This illustrates that each row after the heading \texttt{@data} contains values for a particular entity that refers to the same representation of values in each row in Example 1. In this case, we need to be able to model this representation of values for a particular entity in a standard way for different data formats.

Figure 2.3 shows the metadata on the web page that contain information about the data set such as \textit{Relevant information}, \textit{Past usage} and \textit{Attribute information}. \textit{Past usage} describes the publication that uses the DATATRIEVE data set, whereas \textit{Relevant information} describes some of the context of the data set. \textit{Attribute information} is metadata for the properties of the entity because it describes the meaning of the properties in the data set. For example, \texttt{LOC6\_0} is a property of the entity, and the description \textit{LOC60: number of lines of code of module m in version 6.0} describes information about \texttt{LOC6\_0} in the data set.

However, the metadata for \texttt{LOC6\_0} do not explicitly describe how to deal with blank lines and comment lines. It seems reasonable to interpret \texttt{LOC6\_0} as measuring the number of LOC excluding blank lines and comment lines; however, we do not have information to confirm it. In this case, it is essential to know the explicit meaning for every property of the entity in the data set because if the property is interpreted incorrectly, it will affect the other properties of the entity that relies upon the interpretation. For
example, *Mod_Rate* is described as the rate of modification of module m, i.e., \((\text{Added-LOC} + \text{DeletedLOC}) / (\text{LOC6.0} + \text{AddedLOC})\) in the metadata. This indicates that the interpretation for *Mod_Rate* depends on the meaning of *LOC6.0* in the data set.

In this example, some of the metadata are located on a web page. If the data set is updated, it might not be reflected on the web page of the data set. Separating the data set and metadata into two different places might make synchronisation difficult. The synchronisation of data in this example may involve more than one place.

### 2.2.1.3 Example 3: PROMISE KC2

Figure 2.4 illustrates part of the KC2 data set [16], which contains defect data presented in ARFF. The structure of the data set is similar to Example 2, and the metadata are in the data set. Based on the metadata, the entities of the data set are assumed to be source code files.

There are many different kinds of information in this data set. Some of the information describes the data set creation information, such as the author and the time the data set file was created. Some provide information about the population data (e.g., number of properties and number of instances), while others provide information about the properties of the entities (e.g., description for the properties of the entities). In this example, we need to be able to model the kinds of information that correspond to measurements in the data set.

Similar to Example 2, this data set also contains some elements that describe the structure of the data set, such as the headings for *@relation* and *@data*. Some of these elements are assumed to be separators for other elements in the data set. This illustrates
that some elements provide structure separation in the data sets.

After the @data line in Figure 2.4, there are two rows in lines four to five that contain identical data (9,2.0,1,...,10,3,no). These rows could be declared duplicate data because there is no way to differentiate between them. If an identifier for the entity was given for every row, it would be obvious that every row was different and that an entity was being measured.

In this case, these rows are assumed to belong to different entities, but there is no information to confirm it. Thus, we cannot distinguish between duplicate data or determine whether the rows were accidentally reproduced. This shows that when there is less information to describe the data set, we cannot completely interpret the data set.
2.2.1.4 Example 4: Qualitas Weka 3.7.5

Figure 2.5 shows part of the Weka 3.7.5 data set [19], which contains clone data. This example contains tabular data and metadata. It is more complex than the previous examples because it contains data for different types of entities. The entities are clone pairs of source code files, clusters of the clone pairs and population data (groups of entities).

This data set contains two groups of data for different entities: clone pairs and clusters. Each group has column headers that represent different properties of the entities. Each entity group has a set of descriptions that represent the meaning for each column header. This can be seen in the lines that begin with a number sign (#) followed by numbers and descriptions. The descriptions are assumed to be the metadata for the properties of the entities because the column headers appear explicitly in the descriptions.

This data set contains population data for the groups of entities. It describes data relating to the measurements applying to the two groups of entities. The population data are assumed to have the same form as the typical tabular data because they contain measurements, properties of entities and metadata for properties of entities. This shows that we need to be able to model the elements of population data in the same way we model the elements of regular data.

In Figure 2.5, there are properties associated with values in the highlighted row, and some of these properties are metrics while others are not. This is because some metadata for the properties of the entities describe information about the properties that identify the entities. For example, the metadata for property Method1 is The name of the...
2.2. EXAMPLES OF DATA SETS

lexically first method in the clone pair, which describes information about the property that identifies the entity. This shows that we need to be able to capture the fact that there are different types of properties of entities associated with different types of values that could potentially be a measurement or identifier that identifies the entity.

2.2.1.5 Example 5: PROMISE Ant 1.3

Figure 2.6 shows part of the Ant 1.3 dataset [16], which contains defect data. The data set is presented in a CSV file. Similar to Example 1, this data set has column headers that represent different properties of the entities, and there are no metadata in the data set. The metadata are provided on the web page from which the data set originated. The entities are assumed to be source code files because some of the values represent the conventions for Java source code files.

In this data set, some properties of the entities contain values that are neither measurements nor identifiers for the entities. The property version is one such example because it contains values 1.3 in all rows. This shows that we need to be able to model this type of value that potentially can be values for context of data set.

The metadata on the web page describe information about the publication that uses this data set. In the publication, there is a discussion of the properties of the entity that correspond to the column headers in the data set. The discussion provides a description for 20 of the 24 column headers. We assume that these 20 column headers are metrics because they were explicitly described as Chidamber and Kemerer (CK) object-oriented metrics.

There are four column headers that are not described anywhere, and two of them are the same with different values in the data set. We do not know whether these two column headers refer to the same properties or two different properties. However, we assume that they are different properties because the values are different. This shows that other researchers using the same data sets can potentially misinterpret what the properties of the entity means.

To model this data set, we need to capture what the column headers actually represent. This requires the interpretation of information from the metadata (i.e., the publication). This shows that important information about a data set can sometimes be found external to the data set. It is good that such information is available, but it raises the possibility that another user will not know of the existence of the external information.
2.2.2 Artificial data sets

In previous section, we observed the real data sets from two data repositories. The observations of the real data sets show that some similar aspects of data appear in different places within data sets. For example, the properties of the entities in the CSV data set appear in the first row, whereas the properties of the entities in the ARFF data set appear in the rows after the heading @attribute. This indicates that some similar aspects of data can be easily identified across the different formats of data sets. This suggests that we could create a standard model to describe a structure of data set.

While we observed the real data sets from the two public data repositories, it is possible that there are other data sets from other data repositories that have different structures. Therefore, in this section, we created five examples of artificial data set to show the possible ways that data may be presented in order to expand the set of formats to consider when developing the standard model.

2.2.2.1 Example Artificial 1

Figure 2.7 shows a minimal data set that contains a single value with no other information. In this example, the value is assumed to be a measurement value because it is an integer and looks like a measurement. We created this example to show that it is possible for a data set to contain one measurement value.

Figure 2.7: Artificial data set 1.

2.2.2.2 Example Artificial 2

Figure 2.8 shows a data set that contains a collection of elements in tabular format. In this example, the column headers File and Token are assumed to be the properties of the entities. The property File is associated with values that represent the source code files. The values that represent the source code files are assumed to be the identifiers that are used to identify the entities because the values appear to be unique and they look like the conventions for Java source code files. Thus, the entities in this data set are assumed to be Java source code files.

The property Token is associated with values that are integers. We assume that these values are measurement values because they look like measurements for metrics. Thus, the property Token is assumed to be an identifier for the metric. We created this example to demonstrate the conventions for elements in the data set—that is, property of the entity, identifier for the metric, measurement values and identifiers that identify the
entities. We anticipate that these elements may appear in other places in data sets. Hence, Figures 2.9 and 2.10 present two example data sets to illustrate these elements in two different structures.

<table>
<thead>
<tr>
<th>File</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>One.java</td>
<td>32</td>
</tr>
<tr>
<td>Two.java</td>
<td>17</td>
</tr>
<tr>
<td>Three.java</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 2.8: Artificial data set 2.

2.2.2.3 Example Artificial 3

Figure 2.9 shows the collection of elements of the data set in Example Artificial 2, which is presented as one element per row. Each row is separated by a line break. As mentioned earlier, this example was created to demonstrate that the same elements in Figure 2.8 appear in different rows in this data set.

Although the elements of the data set appear in a different way from Figure 2.8, we can make assumptions similar to Example Artificial 2 because of the conventions of the data set. We assume that the first element *File* is a property of an entity, the second element is an identifier that identifies the entity, the third element *Token* is an identifier for the metric, and the fourth element is a measurement value.

The first four rows are assumed to represent a list of related elements because the second element in the second row represents a source code file that identifies the entity being measured. This indicates that the first four elements could be related to a particular entity in the data set. We also applied this assumption to the second four rows and the third four rows in the data set.

<table>
<thead>
<tr>
<th>File</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>One.java</td>
<td>32</td>
</tr>
<tr>
<td>Token</td>
<td>32</td>
</tr>
<tr>
<td>File</td>
<td>Two.java</td>
</tr>
<tr>
<td>Token</td>
<td>17</td>
</tr>
<tr>
<td>File</td>
<td>Three.java</td>
</tr>
<tr>
<td>Token</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 2.9: Artificial data set 3.
2.2.2.4 Example Artificial 4

Figure 2.10 shows the collection of elements of the data set in Example Artificial 2, which is presented as four elements per row. The four elements are separated by commas in each row. Each row is separated by a line break. As mentioned earlier, we created this example to show the same elements of the data set in Figure 2.8 that appear in a different way from the structure of the data in Example Artificial 3.

In this example, we make the same assumptions for all elements similar to Example Artificial 3. In the first row of this data set, it is assumed that the first element *File* represents a property of an entity, the second element is an identifier that identifies the entity, the third element *Token* represents an identifier for a metric, and the fourth element is a measurement value. We also applied the same assumptions for the elements in the second row and the third row because they have a similar structure as the first row of the data set.

This example shows that the structure of the data set is different to Example Artificial 1 and Example Artificial 2, but the elements of the data set represent the same conventions of the data set. For example, each row contains an element (second element) that represents a unique value for a source code file, which allows us to assume that each row has values for a different entity. This indicates that there is a need to find a way to model every element of the data set to be independent of the different structures of the data set.

![Figure 2.10: Artificial data set 4.](image)

2.2.2.5 Example Artificial 5

Figure 2.11 shows an artificial data set that contains defect data presented in a tabular format. The structure of the data set is similar to Example 1: PROMISE Boetticher data set. This example was created to show four potential data quality issues in a data set: duplicate data, inconsistent data, missing data and incorrect data.

In this example, it is assumed that every row has values for a different entity because the column *Filename* contains unique values that can be used to differentiate the entities. In addition, these values can be assumed as identifiers for an entity because they represent the conventions for the absolute path of Java source code files in a computer. Thus, we can use these values to distinguish between the entities.

There are two rows in lines two and four with identical data. These two rows contain the same values associated with the same properties (*Phase*, *NOF*, *NPM*, *NOC*, *FV*, *nCom*, *nBug*) for the same value in the column *Filename*, which is `/DataGroup/ModuleA/gamesPanel.cpp`. This illustrates duplicate data in the data set.
2.2. EXAMPLES OF DATA SETS

Two rows in lines 14 and 16 contain the same value in the column **Filename**, which is `/DataGroup/ModuleA/board.cpp`. However, these two rows contain different values for properties (Phase, NOF, NPM, NOC, FV, nCom, nBug). This illustrates inconsistent data in the data set.

The first row that contains a value (`/DataGroup/ModuleA/games.cpp`) in the column **Filename** does not have a value associated with column NOC. This illustrates missing data. This could make it difficult to conduct replication studies based on the data set because we do not have measurement values for every metric.

There is a row (in the 15th line) that contains a value (k) in the column **NPM**. The label **NPM** is described as number of public methods in the data set that shows the value should be an integer. This illustrates incorrect data in the data set.

2.2.3 Discussion of examples of data sets

The observations for the examples of data sets are presented in order to analyse the variety of different formats. These data sets contain many types of data that appear explicitly in the data sets and external to them, including values, properties, metadata and other information.

Some aspects of data often appear in the same way. For example, the properties are relatively easy to identify because their position in the structure of data. They are also not difficult to find although they do appear in many different positions, such as in tabular format and ARFF. Another example is the values that are often appear as lists of values in rows separated by commas or whitespace in the structure of data, such as in CSV and tabular format. This indicates that some aspects of data can be easily identified across the different structures of data sets.

We observed some data sets contain labels for properties that convey something about what is being stated, whereas others require further information to be interpreted.
For example, PROMISE Ant 1.3 has the column header labelled ‘name’, which could be interpreted as the name of an entity identifier, but the column header labelled ‘rfc’ is not clear without further interpretation. This indicates that some labels of properties may require metadata for interpretation.

We also observed some data sets contain metadata that may help to understand what the properties are, or what the values mean. For example, Qualitas Weka 3.7.5 has metadata for property Method1 that is the name of the lexically first method in the clone pair, which describes information about the property. In addition, ARFF data sets contain metadata that indicate the type of values, which is data type (e.g., numeric, string). The data type is commonly appear after the properties in the structure of data. The data type allows researchers to interpret the values of property and to determine certain data quality issues (e.g., incorrect data).

Each data set format represents some elements relating to structure separation for elements corresponding to the measurements in the data set. These elements always appear as delimiters, headings or captions. Some data sets use delimiters (e.g., tab, comma or colon) to separate the values or column headers in the rows. Others use headings or captions to indicate a boundary for a group of elements in the data set. For example, ARFF data sets use the heading @data to indicate that the following rows contain values. This indicates that there are many different ways to represent elements corresponding to the structure of the data, and we need to find a standard way to describe these elements.

The observations reveal that real data sets in the PROMISE repository have most of their metadata described in locations other than the data sets. This could be because of the culture of the repository environment, in which researchers record basic information that contains only the properties and values in the data sets, while additional information is recorded in other locations. In this case, we need to find a way to capture metadata in external locations because they may describe important information that corresponds to the measurements in the data sets.

Every assumption that we need to make due to insufficient information may result in data quality issues. For example, PROMISE KC2 indicates that we cannot distinguish between duplicate data without an identifier for the entity in every record, or the entities might have the same data. This tends to suggest that we can highlight potential data quality issues with confidence if we have a standard way of understanding data sets.

2.3 Data sets in software engineering

In software engineering, data sets have been used to support many aspects of empirical research. They contain a wide range of information concerned with measurements, including product, process and resource characteristics in software development and maintenance, quality control, and assurance. Researchers use information from data sets for assessment and prediction to support rational decision making during software development and management with the aim of increasing software quality and
Data sets in software engineering often originate from data stored in software engineering tools such as version control systems, bug tracking systems and mailing lists. These data are not well prepared for research purposes. However, to provide insights into the process of software improvement, researchers extract data from the tools and integrate them using software metrics tools into a form of data sets. Researchers use these data sets in applications such as prediction and estimation models, and they publish them in data repositories.

Many data sets in repositories have similar characteristics (e.g., structures and formats), except for the way they are interpreted. In science, measurements in data sets are often recorded by applying a standard of measurement (e.g., International System of Units [SI units]). For example, a measurement that consists of numbers and a unit of measure in centimetres (e.g., 180 cm) can be interpreted as a value for length. In contrast, measurements in software engineering data sets are frequently generated by software metrics tools without applying any standard to convey the meaning of the measurements. For example, a measurement that consists of numbers for LOC (e.g., 233) can be interpreted in many different ways because there is no standard way for measuring the LOC in a software program.

We often make assumptions to understand what data sets represent due to a lack of metadata —particularly in relation to how the measurements are generated. This is because the methodology for generating measurements in data sets is embedded in algorithms of software metrics tools. The algorithm defines a set of metrics and produces measurements for the metrics. However, many algorithms of software metrics tools do not explicitly describe the methodology for generating the measurements, and in some cases, the algorithms are not available for reference. This creates a significant challenge in the interpretation of metrics and may make it difficult to reproduce measurements based on the data sets.

This research focuses on software engineering data sets because they are more likely to have problems with the interpretation of data. For example, the issue with the Ant 1.3 data set in Chapter 1 illustrates that although an identifier is used to indicate which metric is being used, without further information (i.e., metadata), we may have quality issues with misinterpretation of data. Such issues fail to clearly identify the entity being measured, and we cannot clarify how the metrics are being generated. Although we can make assumptions based on the information given in the data set, there is no standard way to confirm what is actually being measured. This indicates that other researchers using the same data sets may misinterpret what the measurements mean.

### 2.4 Software metrics tools

As mentioned in the previous section, software metrics tools are used to help researchers extract information from software engineering tools to generate data sets. These tools are programs that implement sets of software metrics definitions for a software system.
They extract the required information about the entities from the software system and provide the corresponding measurements for metrics.

For example, Understand scitools and Metrics2 are software metrics tools that are used to compute source code metrics. Source code metrics include complexity metrics (e.g., cyclomatic complexity, number of distinct paths), volume metrics (e.g., LOC, executable code, comment-to-code ratio) and object-oriented metrics (e.g., number of children, depth of inheritance tree). Most metrics measure the source code at the file level and class level. For example, a metric LOC can be analysed from the given source codes based on the levels of methods, classes, files or packages.

There is increasing concern that some existing software metrics tools implement the definitions of object-oriented software metrics differently. Lincke et al. analysed and compared 10 existing software metrics tools from commercial and free tools. They used the same software system as an input to determine whether the tools could interpret and implement object-oriented metrics in the same way. They found that most of the metrics tools produced different measurements for the same metrics, and that this had implications for the results of the analysis based on the measurements [31].

Researchers need to be aware when using data sets generated from software metrics tools because there is a possibility that the measurements in the data sets are metrics-tools-dependent. These measurements are difficult to compare when using different metrics tools. Although documentation is given in metrics tools, there are still unclear definitions of metrics, which may lead to different interpretations and implementations. Thus, these tools have more complex issues that could affect the implementation of measurements in data sets.

### 2.5 Conclusion

Overall, data sets generally contain data related to entities from some group of entities. These entities have data associated with them that relate to their properties. In addition, these data sets contain other information; some may provide the structure, while others are metadata that may provide an understanding of what the properties are or what the measurements mean.

In terms of metadata, we observed many data sets actually contain metadata that are labels for properties. These labels provide names for the properties that may help to understand the meaning of the properties. In addition, ARFF data sets contain metadata that are data type for values. The data type is useful for researcher to interpret the data correctly. This suggests that we need to classify some kinds of metadata into separate categories because we want to be able to distinguish their different aspects.

The discussion of observations indicates that data sets have different structures to represent the content of data sets. Thus, we need to determine a useful way to translate the different structures of data sets into a standard way of interpretation. Therefore, we need a framework that includes a standard model to describe the structure and concepts in a data set, as well as the relationships between each concept. This framework will also
support the need to better understand the content of data sets and communicate issues with data quality.

To enhance our understanding of the importance of data quality in data sets, we performed a literature review of data quality in empirical software engineering. We reviewed key related areas on data quality issues and discussed how to deal with poor-quality data in empirical software engineering. The next chapter presents the reviews of these research areas.
The purpose of this chapter is to analyse related research on data quality in software engineering. The chapter starts with a summary of the reviews of data quality reported in the literature, followed by a discussion of the methods used to carry out the systematic mapping study. The chapter also presents the detailed results of the mapping study, along with a discussion of key findings, any threats to the validity of the results and implications for future research. The chapter then describes data quality issues in software engineering data sets, including the definitions of data quality issues, types of data quality issues and how to deal with them.

### 3.1 Introduction

In Chapter 2, we discussed the importance of understanding the content of data sets and how data sets are structured for data analysis in empirical research. This chapter presents the related research in which the quality of data sets has been investigated; it focuses on the extent to which the research community has addressed data quality issues, particularly in software engineering data sets.

Section 3.2 presents a summary of the reviews of data quality reported in the literature. The reviews reveal a research gap where little attention has been paid to the quality of data sets in empirical research. Our research addresses this gap by carrying out a systematic mapping study to determine what research currently exists that addresses data quality issues. We also focus on the extent to which the existing research has considered the effect of data quality issues on empirical results. Section 3.3 describes the processes and key findings of the systematic mapping study. This mapping study has been published in the Proceedings of the Asia Pacific Software Engineering Conference, 2013 [27]. Section 3.4 describes the key findings from an update of the systematic mapping study for a subset of data quality studies that have recently been published. Section 3.5 describes
another research gap revealed by the mapping study that relates to the definitions of data quality issues in data sets.

3.2 Data quality research

As mentioned in Chapter 1, the quality of data sets used by researchers is critically important [1]. This has been acknowledged and assessed in the past few years because of the effect it could have on decisions and the fact that it may lead to invalid conclusions [11–13, 32–34]. Some studies have extensively analysed data sets and identified many quality issues [8, 10, 11], such as noise, missing data, redundant data and incorrect data.

Over the past decade, many studies have been published in software engineering literature, but only a few have explicitly reported on the quality of data sets. This absence of quality reporting is confirmed by two studies that conduct reviews of data quality in empirical software engineering [11, 12] and one study [1] that reports the review results [13].

In 2008, Shepperd and Liebchen performed a literature review of data quality [12]. They identified 23 studies published from 1995 to 2008 that address data quality issues in software engineering data sets. They classified the studies into six categories: data collection, manual quality checking, empirical quality analysis, automated quality checking, usage of quality metadata and special analysis techniques. They mainly assessed the extent and techniques used to manage the quality of data sets in software engineering. The authors reported that data quality had not received sufficient attention from the software engineering community, as proved by the small number of studies that explicitly reported the quality of data sets.

In 2010, Liebchen wrote a thesis on data cleaning techniques for software engineering data sets [13]. He performed an updated version of his previous systematic literature review [12] to cover a range of empirical software engineering topics. He identified 161 papers published from 1993 to 2010 and classified them into five categories: data collection, manual noise checking, automated noise checking, empirical analysis of quality and data quality metadata. He also evaluated the current evidence in data quality, highlighted the problems to be addressed and proposed techniques to resolve the quality issues.

Shepperd gave a keynote address in 2011 in which he reiterated that the quality of the data used in empirical software engineering research is critically important [1]. He summarised and highlighted the results of the extended version of Liebchen’s systematic review [13], and he proposed approaches to improve the quality of data in the research community. Shepperd also urged the research community to pay attention to data quality in data sets.

In 2013, Bosu and MacDonell performed a systematic literature review covering the period from January 2007 to September 2012 of data quality research in empirical software engineering. They investigated evidence for three elements of data quality: data collection reporting, data pre-processing and data quality issues [11]. Of the 221
identified papers based on the inclusion criteria, they reported that only a few studies (23 studies) considered these three elements, thereby confirming the results of the previous systematic literature review [12].

As noted by Bosu and MacDonell, some studies that have reported data quality issues have used data sets drawn from the PROMISE and ISBSG repositories [11]. The majority of the studies identified that data quality issues are related to data incompleteness, but they did not report the cause. Hence, Bosu and MacDonell indicate that researchers need to be aware of data incompleteness issues when using data sets from public data repositories.

These review studies show that a few researchers have reported the issue of data quality in data sets, but that no specific attention has been paid to the extent to which data quality affects empirical results. Such review studies mainly help us to recognise the data quality topics discussed in the literature, as well as the techniques used to address the data quality issues. However, these studies tend to focus on how researchers address the quality issues in data sets, rather than on how the quality of the data affects their results.

In our research, we conduct a systematic mapping study to understand which research looks at data quality and, in particular, to determine the extent to which researchers have examined data quality issues that might affect their research results. Although the general goal of the mapping study appears to be similar to previous review studies [11, 12], our aim is to identify studies that are concerned with how data quality affects empirical results.

We published a research paper describing this mapping study in 2013 [27], and the targeted review of data quality by Bosu and MacDonell [11] was also published in the same year. As mentioned earlier, Bosu and MacDonell focused on three elements of data quality (data collection reporting, data pre-processing and data quality issues), while we focused on determining the extent to which researchers have examined data quality issues that might affect their research results. In our mapping study, we formulated our own research questions, designed the search methods and applied the procedures for conducting a systematic mapping study in accordance with Peterson et al. [35]. The mapping study is discussed further in the next section.

More recently, two review studies updated their previous literature review of data quality in empirical software engineering [3, 36]. First, Bosu presented an updated version of the previous literature review as part of his thesis on data quality in empirical software engineering, which focused on the investigation of time-aware models in software effort estimation [36]. He identified 61 additional papers from the total of 282 papers published in the period from January 2007 to December 2014. In this updated version of the review, he found 31 studies that reported on all three factors: data collection reporting, data pre-processing and data quality issues. He also reported that the majority of studies that have reported data quality issues have used data sets drawn from public data repositories (PROMISE and ISBSG).

Second, Liebchen and Shepperd updated the previous review study in 2008 [12] by
collating the papers from the published review studies on data quality [11–13, 27, 36] and implementing a snowballing method. They found 283 relevant papers published between 1993 and May 2016, while only 23 papers had been identified in the previous review study (2008). This signifies that the research community has taken serious action in relation to the quality of data sets. Liebchen and Shepperd also reported that there is more research on automated techniques to detect quality issues in data sets [3].

### 3.3 Systematic mapping study

This section describes the procedures used to conduct the systematic mapping study, as well as the key findings of the study. We applied a systematic mapping technique to methodically identify, evaluate and analyse all the available evidence on data quality research applied in the empirical software engineering context, which is the focus of our research.

Systematic mapping is a methodology that aims to present an overview of the relevant research publications in a particular research area, and to classify the primary research publications in that area into defined categories [35, 37]. This mapping study uses high-level research questions to investigate issues such as the number of publications, methods used and issues addressed in the topic areas [38]. The results obtained through mapping studies provide high-level frequencies of current research trends outlined in tables or visualised in bar plots, and these can reveal potential research areas that will be suitable for performing a systematic review [35].

A systematic review focuses on a particular topic area and aggregates the primary studies according to defined research outcomes. In general, a systematic review uses specific and detailed research questions that relate to the outcomes of empirical studies. The main goal of a systematic review is to identify the best practices based on empirical evidence [35].

One important difference between systematic mappings and systematic reviews is the breadth and depth of evaluation from relevant publications [35]. A systematic mapping study typically requires less effort and includes more relevant articles than a systematic review because a mapping study conducts general evaluations (e.g., abstract, introduction and conclusion) to identify the specific context of each publication and develop the classification scheme. In contrast, a systematic review provides a detailed and focused evaluation of selected empirical publications to ensure the evaluation results are based on good-quality evidence.

Systematic mapping studies are being increasingly used in software engineering [35, 37, 39–45] because they provide a good overview of research activities and potential research gaps that can be further studied. In addition, they can help researchers and practitioners to collect, classify and aggregate outcomes from relevant studies to provide a balanced summary of research trends, open issues and areas for improvement.

We adopted the procedures for performing a systematic mapping study described by Peterson et al. [35]. The procedures consist of five steps:
1. definition of research questions (research scope)

2. conducting search (all papers)

3. screening of papers for inclusion and exclusion (relevant papers)

4. keywording of abstracts (classification scheme)

5. data extraction and mapping of publications (systematic map).

The systematic mapping process begins with the preparation of a protocol to be used as a framework for systematic mapping. The protocol, which is part of the planning stage, describes the strategies to be executed when the review is conducted. It specifies the research questions, strategy to be used for searching and storing the literature, inclusion and exclusion criteria, strategy used to classify keywords of abstracts, and strategy used to synthesise the evidence.

We conducted the review by identifying relevant literature required to answer the research questions. This involved searching for evidence using the search strategy in the protocol. Once the relevant literature was identified, we carried out the strategy for the selection of studies by applying the inclusion and exclusion criteria. After we filtered the studies based on the stated criteria, we constructed new classification schemes based on the research questions rather than using the strategy to classify the keywords of abstracts. This was because we constructed the research questions to capture the scope for our research activities in this mapping study. After we identified the classification schemes, we extracted data from the selected studies to sort them into the scheme.

While extracting the data, the frequency of studies in the scheme was calculated. The frequency of studies provides the means to visualise which categories in the scheme have been considered in past research, and thus to identify possibilities for future research. The following subsections discuss the research questions constructed for this mapping study, followed by a detailed explanation of the execution of the systematic mapping.

### 3.3.1 Definition of Systematic Mapping Research Questions

The most important part of systematic mapping is the formulation of the research questions, which will lead to identifying the scope for the research activities [38]. The formulation of research questions for this mapping study has been structured with the guide of the PICOC criteria suggested by Petticrew and Roberts [46].

There are five elements of PICOC: **Population, Intervention, Comparison, Outcomes** and **Context**. According to Petticrew and Roberts, population is the target group for the investigation, and intervention indicates the aspects of investigation. Comparison refers to the aspect of the investigation to which the intervention is being compared, outcomes determine the effect of the intervention, and context describes the environment of the investigation [46].

In this mapping study, we have not included the ‘Comparison’ part of the PICOC criteria because the aim of this study did not include finding evidence about the comparison of methods or models. Table 3.1 details the PIOC criteria used.
Table 3.1: Summary of PIOC

<table>
<thead>
<tr>
<th>Population</th>
<th>Software engineering data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Method or techniques to address data quality issues</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Quantity and type of evidence relating to data quality issues and techniques</td>
</tr>
<tr>
<td>Context</td>
<td>Review(s) of any empirical experiments or studies within the domain of empirical software engineering</td>
</tr>
</tbody>
</table>

The primary focus of this research is to provide an overview of the research involving data sets with the aim of understanding which research addresses data quality. This mapping also intend to determine the extent to which researchers have addressed data quality issues in order to evaluate the trustworthiness of their results.

In this study, we intend to answer the following three research questions (RQ):

- **RQ1**: What data quality topics are addressed in the research?
- **RQ2**: What data quality problems are addressed in the literature?
- **RQ3**: What data sets are used in studies on data quality?

### 3.3.2 Conducting a search for relevant literature

The main task in conducting a search for relevant literature is to construct search strings based on defined population, intervention, outcome and context. We adopted the following formulation of search strings suggested by Kitchenham et al [47] as follows:

1. The main search terms were determined by considering the population, intervention, outcome and context (see Table 3.2).

2. Additional search terms were determined by considering the search terms used in the systematic review on data quality in software engineering by Liebchen [13]. Note that we considered the search terms used in this systematic review because the general goal of the systematic review was the same as our mapping study, and we had access to the search terms used in the systematic review from Liebchen's thesis (see Table 3.3).

3. Synonyms of the search terms were identified (see Table 3.4).

4. Search strings were formulated using Boolean operators (OR, AND) to incorporate alternative spellings and synonyms, and to link the major terms of population, intervention and additional terms from the systematic review (see Table 3.5 and Table 3.6).

The complete search string used to conduct the search of relevant literature is illustrated in Table 3.6.
3.3. SYSTEMATIC MAPPING STUDY

Table 3.2: Search terms derived from PIOC criteria

<table>
<thead>
<tr>
<th>Population</th>
<th>Software/ data sets.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Techniques/ method / metric.</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Quality.</td>
</tr>
<tr>
<td>Context</td>
<td>Researchers, empirical studies, observations, students, questionnaires, surveys, formal experiments, case studies, software industry, developers, maintainers, testers, project managers</td>
</tr>
</tbody>
</table>

Table 3.3: Search terms from systematic review on data quality in empirical software engineering [13]

“data quality” OR “noisy data” OR “noise” OR “inconsistent pieces” OR “erroneous data” OR “clean the data” OR “data cleaning” OR “Issues Arising from the Data Collection” OR “recorded with enough consistency and completeness” OR “not very consistent”

Table 3.4: Synonyms derived from PIOC criteria

<table>
<thead>
<tr>
<th>Search terms</th>
<th>Alternative terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>Program/software program/software system/application</td>
</tr>
<tr>
<td>Data sets</td>
<td>Data.</td>
</tr>
<tr>
<td>Quality</td>
<td>Aspect/attribute/property/character/feature</td>
</tr>
<tr>
<td>Techniques</td>
<td>Method/approach.</td>
</tr>
<tr>
<td>Metric</td>
<td>Measure/measurement.</td>
</tr>
</tbody>
</table>

Table 3.5: Search Terms Obtained by Combining Similar Terms using Boolean OR

| Software                  | (software OR program OR software program OR software system OR application) |
| Data sets                 | (data sets OR data OR information) |
| Quality                   | (quality OR aspect OR attribute OR property OR character OR feature) |
| Technique                 | (technique OR method OR approach) |
| Metric                    | (metric OR measure OR measurement) |
| Data quality              | (data quality OR noisy data OR noisy data set OR noise data set OR noise dataset OR noise OR inconsistent pieces OR inconsistent value OR inconsistent data OR erroneous data OR erroneous value OR clean the data OR data cleaning OR empirical data OR not very consistent) |

Table 3.6: Search String Obtained by Concatenating Terms by Using Boolean AND

(software OR program OR system OR application) AND (technique OR method OR approach) AND (metric OR measure OR measurement) AND (data quality OR noisy data OR noisy data set OR noise data set OR noise dataset OR noise OR inconsistent pieces OR inconsistent value OR inconsistent data OR erroneous data OR erroneous value OR clean the data OR data cleaning OR empirical data OR not very consistent)

We divided our search process into two approaches—namely, primary and secondary. The primary approach was used to search for papers via online databases using predefined search strings. In the secondary approach, a manual search was conducted to
retrieve relevant papers using references of the selected papers from the primary search. Both approaches covered a wide range of publications to reduce the potential risk of missing relevant literature.

In this mapping study, we restricted our focus to research in the computer science and software engineering domain. We considered papers published within a five-year period, from 2008 to 2012, as a prior study that had performed a systematic review [12] was published in 2008. We did not consider the duration of years in the targeted review [11] published in 2013 because, as mentioned earlier, this mapping study was conducted in parallel with the targeted review.

3.3.2.1 Primary search

We conducted the primary search using the following online databases: SCOPUS, ScienceDirect, SpringerLink and ACM Digital Library. These databases were selected because they are the major digital libraries and search engines that are most frequently used in systematic literature reviews performed by software engineering researchers. These databases are subscribed to by the University of Auckland’s library under the ‘Computer Science’ subject category. Further, these databases cover peer-reviewed journals published by IEEE, ACM, Elsevier and Springer, which provide valid measures for the quality of the publications.

The results for the primary search are summarised in Table 3.7. The search results are presented in the sequence in which the databases were searched. The search strings given in the previous section were tailored to each database-specific search requirement. The details of the search strings used for each database, and when these databases were searched, can be found in Appendix A.1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Database name</th>
<th>No. of search results</th>
<th>No. of duplicates found</th>
<th>No. of relevant papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SCOPUS</td>
<td>139</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2.</td>
<td>ScienceDirect</td>
<td>33</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3.</td>
<td>SpringerLink</td>
<td>102</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4.</td>
<td>ACM Digital</td>
<td>22</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>296</td>
<td>3</td>
<td>26</td>
</tr>
</tbody>
</table>

3.3.2.2 Secondary search

For the secondary search, we performed two manual search approaches. First, we used the snowball method to identify relevant papers by referring to references of selected papers from the primary search. In particular, we reviewed all references in the selected papers from the primary search. If a paper was found to be suitable, it was added to the existing list of papers that had qualified for the selection of studies.

Second, we selected relevant papers published from 2008 to 2010 from a reference list of the previous systematic review [12] because the review search strings were similar to
our mapping study. As explained earlier, we constructed our search strings by considering the search string from the systematic review on data quality because the general goal of the review is the same as this mapping study. In this second approach, we reviewed the references of the systematic review and selected relevant papers that qualified for the selection of studies. The search results for the secondary search are summarised in Table 3.8.

<table>
<thead>
<tr>
<th>No.</th>
<th>Method</th>
<th>No. of selected papers</th>
<th>No. of duplicates found with primary search</th>
<th>No. of relevant papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Snowball</td>
<td>21</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2.</td>
<td>Papers from previous review</td>
<td>61</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>82</td>
<td>27</td>
<td>38</td>
</tr>
</tbody>
</table>

As shown in Table 3.8, the total number of relevant articles (38) was more than the total number of relevant articles (26) from the primary search results in Table 3.7. This may have occurred because of the restriction criteria for the search terms that we used in our search string in the primary search. However, combining the two approaches for the search process can reduce the problems that we might face if we conducted only a primary or secondary search.

### 3.3.3 Selection of studies

Based on the research questions, we defined two types of inclusion and exclusion criteria to filter out papers that were not relevant: basic and detailed. The basic inclusion criterion aimed to refine the initial results of the search process. The inclusion criterion was to include papers if the title and abstract were in the computer science and software engineering domain. The basic exclusion criteria were to exclude papers not listed in the computer science and software engineering domain, and to exclude duplicate papers.

We then applied the detailed inclusion and exclusion criteria to the results obtained from the basic inclusion and exclusion criteria. Each paper had to satisfy all of the following conditions in the detailed inclusion and exclusion criteria:

1. **Inclusion:**
   a) peer-reviewed papers related to analysing and evaluating some aspect of empirical software engineering
   b) papers that addressed issues with the quality of data explicitly
   c) papers that used software engineering data sets.

2. **Exclusion:**
   a) papers that described data quality but did not use software engineering data sets
b) non-peer-reviewed papers

c) papers not written in English.

We obtained an initial set of 296 papers after the execution of an automatic search for the primary search. The initial set was reduced to 95 papers after screening the titles and abstracts and excluding irrelevant and duplicate papers. We applied the detailed inclusion and exclusion criteria by reading the introduction and conclusion of each paper. We found that only 26 papers were relevant. In the results of the secondary search, we applied basic and detailed inclusion and exclusion criteria for the 82 selected papers and found 38 relevant papers. In total, 64 papers were selected for this mapping study (see Appendix A.2 for the list of included studies).

3.3.4 Developing the classification scheme

During the process of the selection of studies, we assumed that the quality of papers obtained would be ensured by the inclusion and exclusion criteria. However, the selection of studies was based on relevance but not quality. In particular, we selected relevant papers that were in accordance with the focus of the research questions and the goal of the mapping study.

To develop a new classification scheme for the selected papers, we formed four categories based on our research questions and mapping study goal to achieve more in-depth analysis. The four categories were: data quality topics (RQ1), data quality problems (RQ2), types of data sets (RQ3) and data quality evaluation. After identifying the categories, the papers were sorted into the categories and the number of studies per category were represented (see Section 3.3.6).

Prior to the start of the classification scheme development for the data quality topics, we used the existing topics from the previous systematic review [12]. However, we encountered problems while classifying the selected papers due to the lack of clarity on what the topics meant. Therefore, we adopted two topics from Liebchen and Shepperd [12] and added five new topics.

We performed a test review for a pilot selection of papers by the PhD supervisors. We verified the scheme using a selection of 18 papers. During the verification process, we encountered difficulties classifying the papers because of the vague descriptions in the data quality topics. After a discussion, we proposed a new research data life cycle to provide a consistent way of classifying the papers into the data quality topics scheme. This is further discussed in Section 3.3.6.1.

3.3.5 Data Extraction and Mapping of Studies

The data extraction process involved reading the abstracts of the relevant studies to collect the data needed for the synthesis of evidence [35]. In this step, we used an Excel spreadsheet to document the data extraction process and generate the mapping results. We designed a data extraction form to obtain relevant information from the studies (A
sample of the data extracted from one of the studies (S7) is given in Appendix A.3). The information extracted from each study included:

- full reference of the study (e.g., author, study title, type of study, year of study)
- study research information (e.g., research problem, objective of study, contribution of study and domain of study)
- necessary information required to answer the research questions of the mapping study (e.g., data quality topics, data quality problems, types of data sets, techniques and metrics used in the study).

We created an overview of the mapping results, frequency of topics and research gaps using a bubble plot diagram that consists of two x-y scatter plots with bubbles lying on the intersection of each category (see Figure 3.4 in Section 3.3.6.6). The size of each bubble is proportional to the number of publications for each pair of categories.

### 3.3.6 Results of the systematic mapping

This section presents the synthesis of evidence of the systematic mapping, beginning with the new research data life cycle and followed by a description of the findings of each research question. The overall results of this mapping study are presented as a bubble plot diagram at the end of this section.

#### 3.3.6.1 Research data life cycle

When trying to categorise the papers identified in this mapping study, we realised that data could be thought of as going through a ‘life cycle’ in its use in research, from its creation to the completion of research based on the data. We found that the phase the data goes through in this journey provides a useful way to classify the data quality topics into relevant phases of the life cycle. The research data life cycle is shown in Figure 3.1.

The five phases of research data consist of source, which is the creation of the data; raw, which is the recording of the data; refined, in which the data have been processed in some way (possibly many times); results, in which the data are in a form from which conclusions are drawn; and completion, which determines the trustworthiness of the conclusions due to the quality of the data is determined. Data quality questions arise with the transitions between each phase.

For example, consider a study that examines the relationship between class size (as measured by LOC) and whether a class contains one or more faults. Such a study may be carried out by gathering fault data from the commit logs in the source code control system of an open source software project. In this case, the source is the developer doing the commit, the collection is the use of the source code control system to make the commit, and the resulting log entries are the raw data. The data quality questions for this transition include whether a log entry was in fact made and how accurate and detailed the description is.
The study may then process the commit logs in some way (e.g., with scripts) to produce data that ascribes faults to classes based on the contents of each log entry. This process produces refined data. There may be multiple processing transitions, with each applying different techniques to change the data in some way—for example, to change its format or deal with data quality issues. In another example of processing raw data to refined data, the raw data may ascribe multiple faults to a class, but the data needed for the study are whether or not a class contains at least one fault.

Once the data have been suitably refined, an analysis technique (e.g., a machine learning technique) might be used to determine any relationship between the fault data and the size data for each class in the study. The data produced from this analysis are in the results phase. Finally, threats of validity analysis may be applied to determine the quality of the conclusions drawn from the results data.

In the above example, the raw data may not necessarily come from the source control system. They may, for example, be extracted and provided in packaged form in some repository. Similarly, refined data such as class faults data may be provided in a repository (e.g., promisedata.org) in refined form. When the study is published, the results data may be provided by the authors for peer review. That is, a given study may not necessarily follow data through all of the phases in our model; instead, it may acquire data during any phase as its starting point.

The research data life cycle is one of the contributions of this mapping study. It describes the stages a data set may go through, and it can therefore be used to clearly identify when issues with data quality may be introduced. As mentioned earlier, it also helps us to categorise the papers identified into the data quality topics that we described in the relevant phases of the research data life cycle. We will use the research data life cycle when applying the guidelines for creating data sets (see Chapter 7).
3.3. SYSTEMATIC MAPPING STUDY

3.3.6.2 Data quality topics (RQ1)

As mentioned earlier, we encountered problems when trying to apply Liebchen’s topic categories [12] due to the lack of clarity, therefore we refined these categories. We used two existing topics — data collection (DC) and special analysis techniques (SAT) — and added another five topics: manual quality detection (MQD), automated quality detection (AQD), data cleaning techniques — filtering (DCTF), data cleaning techniques — polishing (DCTP) and robust techniques with respect to data quality (RTDQ). These topics are described in Table 3.9.

Table 3.9: Data Quality Topic Description

<table>
<thead>
<tr>
<th>No.</th>
<th>Data Quality Topic Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Data collection (DC)</td>
<td>Refers to if the paper addresses or makes suggestions concerning data quality at the collection stage.</td>
</tr>
<tr>
<td>2.</td>
<td>Manual quality detection (MQD)</td>
<td>Refers to if the paper suggests how to address, or actually addresses, data quality problems using a manual data quality procedure.</td>
</tr>
<tr>
<td>3.</td>
<td>Automated quality detection (AQD)</td>
<td>Refers to if the paper suggests how to address, or actually addresses, data quality problems using an automated data quality procedure.</td>
</tr>
<tr>
<td>4.</td>
<td>Data cleaning techniques-filtering (DCTF)</td>
<td>Refers to if the paper explicitly uses or recommends data cleaning methods to filter data quality issues.</td>
</tr>
<tr>
<td>5.</td>
<td>Data cleaning techniques-polishing (DCTP)</td>
<td>Refers to if the paper explicitly uses or recommends data cleaning methods for data polishing and data scrubbing.</td>
</tr>
<tr>
<td>6.</td>
<td>Special analysis techniques (SAT)</td>
<td>Refers to if the paper explicitly uses or recommends any particular analysis techniques to handle data with quality issues.</td>
</tr>
<tr>
<td>7.</td>
<td>Robust technique with respect to data quality (RTDQ)</td>
<td>Refers to if the paper explicitly uses or recommends any particular techniques that are robust to handle the low quality of data.</td>
</tr>
</tbody>
</table>

Liebchen’s previous systematic review used manual noise checking and automated noise checking among the categories of data quality topics [13]; however, in this study, we avoided using the term ‘noise’ because it has been used inconsistently in previous research. Therefore, we used manual and automated quality detection to represent data quality detection and the data cleaning techniques of filtering and polishing to represent data cleaning techniques.

The DC category represents papers that mentioned or were concerned with data quality problems in the early stages of research or experiments. Some of the papers classified in this category are Bachmann et al. [48], Le-Do et al. [49] and Zhou et al. [50]. The MQD category represents papers that used a manual procedure to detect the quality issues in data sets. One example of the MQD category is the study carried out by Bachmann et al. in which a tool was developed to detect the missing links (between bugs and bug commits) [7]. The AQD category represents papers that used an automated
3.3. SYSTEMATIC MAPPING STUDY

procedure to detect the quality issues in data sets. One example of a paper in the AQD category is the study carried out by Le-Do et al. that proposed an approach to identify the noise in software project data [49].

The DCTF category represents papers that used or recommended data cleaning methods to remove the quality issues from data sets. An example paper categorised in this way is the study carried out by Kim et al. in which an algorithm was developed to detect and eliminate noise in defect data to improve the accuracy of defect prediction [14]. The DCTP category represents papers that used or recommended data cleaning methods to repair or correct the quality issues in data sets. For example, a study conducted by Rongxin et al. in which an automatic algorithm was developed to recover the missing links between bugs and change logs in defect data sets [51].

The SAT represents analysis techniques that are used to handle data with quality issues. For example, an analysis technique that handles the missing data in effort estimation models [5]. The RTDQ represents techniques that are robust enough to deal with low-quality data, such as data mining techniques [52].

In our mapping study, we found 64 papers that satisfied our inclusion criteria. Table 3.10 shows the distribution of papers on data quality topics obtained using the method described in section 3.3.1. Papers that addressed more than one topic (of which there were 27) were counted in every topic, so the totals add up to more than 64.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Collect</th>
<th>Process</th>
<th>Analyse</th>
<th>Evaluate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic/Year</td>
<td>DC</td>
<td>MQD</td>
<td>AQD</td>
<td>DCTF</td>
<td>DCTP</td>
</tr>
<tr>
<td>2012</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2010</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2009</td>
<td>13</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2008</td>
<td>12</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>24</td>
<td>8</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

As shown, according to the data quality topics, the number of data quality contributions decreased from 2008 to 2012. This is consistent with the result obtained in the previous study [13]. In particular, there are a high number of data quality contributions on the DC topic, which is consistent with Liebchen’s previous review [13] on DC. However, we found that the data quality contribution trend was decreasing from the second phase towards the end phase of the research data life cycle.

Although the decreasing trend of the data quality contributions was in line with the previous study [13], for a three-year range (2008-2010), a few data quality contributions were noted in recent years (2011-2012). In particular, in the evaluation phase, there was a paper on every topic (except RTDQ) in 2011. Moreover, as shown in Table 3.10, the research community contributed nine papers to the process phase between 2011 and 2012.
3.3.6.3 Data quality problems (RQ2)

A variety of terms have been used to discuss problems with data quality in empirical research. The results in Figure 3.2 show the terms used and how many of the identified papers used them. We extracted these terms directly from the relevant papers without any interpretation. We added one new category —‘Did not say’ —for studies in which data quality was mentioned, but no specific problem was identified. Papers that employed more than one term (of which there were 13) were counted in every category, so the totals add up to more than 64.

As shown in Figure 3.2, 9 terms were used to describe problems relating to data quality, but it was not always clear what the authors meant by these terms or they appeared to be used inconsistently. The papers used unclear terms that could be interpreted in many different ways. For example, if a piece of data was missing in a given data set, the terms ‘missing data’ or ‘incomplete data’ may be used interchangeably by one author, but they could indicate different problems when used by another author. This problem could also occur with terms such as ‘incorrect data’, ‘invalid data’ and ‘inaccurate data’.

In our mapping study, we found 13 papers that had used more than one term to indicate more than one problem related to data quality. It appears that the researchers were trying to solve more than one data quality problem simultaneously in order to produce more trustworthy data.

3.3.6.4 Types of data sets (RQ3)

There are many types of data sets used in empirical research. We classified the data sets into eight categories: open source, closed source, ISBSG, synthetic, NASA, PROMISE, CSBSG and ‘none’. These categories are described in detail in Table 3.11.
3.3. SYSTEMATIC MAPPING STUDY

Table 3.11: Data sets information

<table>
<thead>
<tr>
<th>No.</th>
<th>Data set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Open source</td>
<td>Refers to if the paper used a free distribution data set.</td>
</tr>
<tr>
<td>2.</td>
<td>Closed source</td>
<td>Refers to if the paper used a commercial and industry distribution data set.</td>
</tr>
<tr>
<td>3.</td>
<td>ISBSG</td>
<td>Refers to if the paper used data sets from the International Software Benchmarking Standards Group (ISBSG) repository.</td>
</tr>
<tr>
<td>4.</td>
<td>Synthetic</td>
<td>Refers to if the paper used artificial data sets that were created to meet certain conditions that may not be found in the real data sets.</td>
</tr>
<tr>
<td>5.</td>
<td>NASA</td>
<td>Refers to if the paper used data sets from the NASA Metrics Data Program repository (a public repository).</td>
</tr>
<tr>
<td>6.</td>
<td>PROMISE</td>
<td>Refers to if the paper used data sets from the PROMISE repository (a public repository).</td>
</tr>
<tr>
<td>7.</td>
<td>CSBSG</td>
<td>Refers to if the paper used data sets from the China Software Benchmarking Standards Group (CSBSG) repository.</td>
</tr>
</tbody>
</table>

Figure 3.3 shows the data sets used and how many of the relevant papers used these. The category ‘none’ refers to papers that did not describe any specific data sets. Papers that used more than one type of data set (of which there were eight) were counted in every type of data set, so the totals add up to more than 64.

As shown in Figure 3.3, open source was the most frequently used type of data set in among the selected papers. Six papers used Apache [34], three papers used Mozilla [7] and nine papers used Eclipse [53]. In total, 30 papers used data sets from the following: ISBSG (15 papers) [54], CSBSG (one paper)[4], PROMISE (five papers) [54] and NASA (nine papers) [8]. The result for the PROMISE data set, which was only used by five of the papers, was unexpected as this was inconsistent with the result from the previous study.
This could be due to a specific search strategy and the different focus of the targeted review [11].

The remaining 14 papers used closed-source data sets that belonged to a commercial body or organisation in the industry. For example, one study used bank and stock data sets from financial companies [54]. Six papers used synthetic data sets, which refer to in-house software development projects for the simulation environment [55].

### 3.3.6.5 Overall results

As mentioned in the introduction, the overall aim of this mapping study was to explore the extent to which researchers were concerned that the quality of their data had affected their empirical results. To achieve this goal, we evaluated all of the papers and classified them according to the type of data quality evaluation. We derived eight types of evaluation from the papers we reviewed. The results are summarised in Table 3.12. We used ‘subset’ data, which refers to a selection of data sets from the original source of the data sets.

<table>
<thead>
<tr>
<th>Type</th>
<th>Data Quality Evaluation</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Create a way to determine the quality of the data</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>Create a way to improve the data quality</td>
<td>17</td>
</tr>
<tr>
<td>III</td>
<td>Create a way to resist low-quality data</td>
<td>2</td>
</tr>
<tr>
<td>IV</td>
<td>Use a technique to improve data quality</td>
<td>12</td>
</tr>
<tr>
<td>V</td>
<td>Use a robust technique with respect to low-quality data</td>
<td>3</td>
</tr>
<tr>
<td>VI</td>
<td>Use subset data to improve data quality</td>
<td>11</td>
</tr>
<tr>
<td>VII</td>
<td>Use a technique to determine the quality of data and show concern about the results</td>
<td>5</td>
</tr>
<tr>
<td>VIII</td>
<td>Mentioned data quality</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>64</strong></td>
</tr>
</tbody>
</table>

As shown, the majority of papers attempted to develop solutions to improve the data quality. There were seven papers in Type I, which indicated the researchers’ interest in determining the quality of the data. The seven papers in Type VIII mentioned the importance of data quality without taking further action. Four types of data quality evaluation —Types IV, V, VI and VII —used techniques to improve the quality of the data. In total, in 31 papers, serious action had been taken regarding the quality of the data sets and this had influenced the behaviour of the results.

### 3.3.6.6 Mapping

To present an overview of data quality in software engineering research, Figure 3.4 presents a map that shows the distribution of papers according to the data quality topics, data quality problems and types of data sets. The size and number in each bubble represents the number of papers classified to the pair of categories corresponding to the bubble coordinates. We grouped the relevant papers in more than one element for each category because they may make multiple contributions. Therefore, this will reflect that the total paper count does not equal the absolute total of 64.
As discussed in Section 3.3.6.1, we categorised the papers identified in our mapping study into data quality topics that listed in the phases of research data life cycle in Figure 3.1. The results of the mapping in Figure 3.4 indicate that most research efforts involved the data collection topic, which is in the collection phase of the research data life cycle. This seems to show that most researchers tended to describe how data were collected with respect to the data quality problems associated with the data collection.

Another area that has been widely explored is the process phase, in which a considerable amount of research effort is spent to improve the quality of the data. The most investigated area in the process phase is the manual quality detection topic, where researchers examine problems relating to data quality in a manual way. In addition, the data cleaning techniques for filtering topic appears to have received more attention than other topics in the process phase.

A small research effort has been made to analyse and evaluate the phases of the research data life cycle to increase data quality with analysis and robust techniques. These areas do not seem to have been explored by many researchers.

3.3.7 Discussion of the Systematic Mapping Findings

This section presents the systematic mapping findings according to the three research questions: data quality topics (RQ1), data quality problems (RQ2) and type of data sets (RQ3). We compare the evidence in this mapping study with findings reported in previous literature. Finally, we discuss the threats to the validity of systematic mapping and the
implications for research.

3.3.7.1 Studies addressing data quality

This mapping study sets out to determine how confident we can be that research results are not affected by poor-quality data. If the quality of the data set is poor, the results cannot be trusted. We found that 31 papers had taken serious action on the quality of data sets, which influenced the behaviour of the results obtained. This finding is consistent with prior findings that only a small proportion of research has considered data quality in software engineering [12], [11]. This indicates that researchers still do not pay enough attention to reporting how trustworthy their data are.

This mapping study also provides a refined version of existing data quality topics adapted from Liebchen’s systematic review [12]. We adopted two topics from that review and added five new topics in this mapping study. We refined the adopted topics due to the unclear terms and descriptions used in the review. We classified studies according to the refined list for data quality topics and found that 12 studies applied data cleaning techniques for filtering and polishing data sets. It could be that few researchers have reported the techniques used in cleaning poor-quality data for analysis in their research.

3.3.7.2 Unclear and inconsistent terminologies of data quality issues

Another important finding is that many studies used unclear and inconsistent terminologies to describe problems with their data sets. The variation of study topics and the specific context of problems may have resulted in researchers using terminologies interchangeably to indicate different problems. This makes it difficult to clearly report the quality of data sets, and researchers may have many ways to interpret problems related to data quality in their research. There is a need to adopt consistent terminology so data quality issues can be clearly identified.

3.3.7.3 Different sources of data sets

Our findings also highlight that studies used data sets from many different sources, including organisations and data repositories. Thirty studies used data sets from different data repositories. These data sets may have a variety of formats and structures, and every format has its own way of representing the structure of the data. They also use specific terminologies to describe the content of the data. Thus, the variation of formats and the specific terminologies used in data sets may influence researchers’ interpretation and understanding of the content of data sets prior to describing the problems related to the quality of the data.

3.3.7.4 Threats to validity of the systematic mapping results

Several factors need to be taken into account when generalising the results of the systematic mapping. We identified three factors for threats to the validity of the systematic mapping results.
The first factor is publication bias during the selection of papers in the mapping study. We cannot ensure that all relevant papers were selected throughout the searching and selection process. The search string in the searching process may have included inadequate search terms relating to data quality. To mitigate this threat, we developed a systematic mapping protocol, verified it with the PhD supervisors and followed the protocol in a consistent way. In addition, we extended the potential of papers selected through the snowball method and followed references from a similar study [13].

The second factor is the uncertain description of data quality topics. We encountered difficulties classifying the papers using the previous study [12] topics because of the vague descriptions. To minimise this threat, we refined the topics and developed a new research data life cycle to consistently classify the papers into the data quality topics.

The third factor is misclassification of papers in the data quality topics. This mapping study has the potential to classify the papers incorrectly in the data quality topics. To minimise this threat, we implemented a pilot test to validate the classification for the data quality topics. The pilot test involved the PhD supervisors to correctly classify a selection of 18 papers into the data quality topics.

### 3.3.8 Implications for research

Based on the findings of the mapping study, we found that there is a lack of clear and consistent terminology regarding data quality in the research literature, particularly with respect to the types of quality problems a data set might have. The research community may interpret data quality issues differently. This suggests that further reviews need to be conducted to investigate the terminologies used in the definitions of data quality issues in the literature.

We also found that researchers used data sets from many different sources, including organisations and data repositories. These data sets may have many different formats and structures in describing their content. This could be evidence to support our findings in Chapter 2 that there is a need to develop a standard way to describe the structure of data sets.

### 3.4 Recently published data quality studies not included in the systematic mapping study

As the original systematic mapping study included studies published between 2008 and 2012, this section presents the key results from the systematic mapping study conducted for the recently published data quality studies dating from January 2013 to October 2016. We applied the same procedures to conduct the systematic mapping study as described in sections 3.3.1 to 3.3.5 to the recently published studies. We found that the results of the mapping study for the recently published studies were particularly notable for the distribution of data quality topics per year and the types of data sets used in the selected papers. We describe these results further in the following paragraph.
3.4. RECENTLY PUBLISHED DATA QUALITY STUDIES NOT INCLUDED IN THE SYSTEMATIC MAPPING STUDY

As described in section 3.3.6.2, the original mapping study identified 64 papers published between 2008 and 2012 that satisfied our inclusion criteria. In the mapping study of the recently published studies, we identified 38 papers that satisfied these same inclusion criteria. Table 3.13 shows the distribution of papers on data quality topics for the papers published per year from 2013 to 2016. As in Table 3.9, we use acronyms to represent the seven data quality topics. Papers that addressed more than one topic were counted in every topic that is addressed.

Table 3.13: Distribution of Papers on Data Quality Topics (2013-2016)

<table>
<thead>
<tr>
<th>Topic/Year</th>
<th>Collect</th>
<th>Process</th>
<th>Analyse</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>5 1</td>
<td>3 1</td>
<td>0 6</td>
<td>0 16</td>
</tr>
<tr>
<td>2014</td>
<td>4 4</td>
<td>0 1</td>
<td>6 0</td>
<td>15</td>
</tr>
<tr>
<td>2015</td>
<td>2 1</td>
<td>1 0</td>
<td>2 0</td>
<td>6</td>
</tr>
<tr>
<td>2016</td>
<td>1 1</td>
<td>0 2</td>
<td>1 4</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>12 7</td>
<td>4 2</td>
<td>1 14</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3.5 shows the distribution of the data quality topics for the papers published during two specified periods (2008-2012 and 2013-2016). The total number of papers published on the SAT topic increased in the 2013-2016 period with 14 papers compared to nine papers in the 2008-2012 period. In contrast, the total number of papers published on the DC topic decreased in the 2013-2016 period with 12 papers compared to 36 papers in the 2008-2012 period. This demonstrates that, in recent research, serious action has been taken with regard to developing techniques to handle data with quality issues.

Figure 3.5: Distribution of Papers on Data Quality Topics (2008-2016)

Figure 3.6 shows the types of data sets used in the published papers for two specified periods: 2008-2012 and 2013-2016. The types of data sets for both periods are similar, though the number of papers that used these types of data sets were different. As shown in the 2013-2016 period, according to the types of data set, the PROMISE data set was
3.5. REVIEW OF DATA QUALITY ISSUES

The most frequently used with 15 papers, followed by the closed source data set with nine papers. These results differ from the number of published papers for the 2008-2012 period, in which the PROMISE data set was used by five papers and closed source data sets were used by 21 papers. This indicates that the recent papers use data sets from public data repositories more widely in their analysis than the research from the earlier period.

![Frequency of data sets types used in the papers (2008-2016)](image)

Figure 3.6: Frequency of data sets types used in the papers (2008-2016)

Table 3.14 shows the distribution of papers based on data quality evaluation for the two periods: 2008-2012 and 2013-2016. The results for the 2013-2016 period show that the majority of papers gave consideration to data quality by mentioning the importance of data quality in data sets or reporting the quality issues in data sets. We found that the total number of papers that used techniques to address and improve the quality of data sets for the 2013-2016 period (22 papers) was lower than the total number of papers for the 2008-2012 period (31 papers). These papers were categorised as one of four types: IV, V, VI and VI.

The results of the mapping study for the recently published studies on data quality did not change the main patterns of findings from the original mapping study (2008-2012). The findings from our studies show that the research community is still not paying sufficient attention to the quality of data sets. This is clear in the small number of papers in which serious action was taken relating to the quality of the data set that influenced the empirical results in the recently published papers from the 2013-2016 period. In addition, most of these papers provided additional evidence regarding the absence of efforts made by the research community to assess and improve the quality of data used in research.

3.5 Review of data quality issues

As mentioned in the mapping study, we found a lack of clear and consistent terminology in the reporting of data quality issues in the literature. To investigate this gap, a further review of the data quality issues in the literature is conducted in this section. We start by
3.5. REVIEW OF DATA QUALITY ISSUES

Table 3.14: Distribution of papers based on data quality evaluation

<table>
<thead>
<tr>
<th>Type</th>
<th>Data Quality Evaluation</th>
<th>No. of papers (2008-2012)</th>
<th>No. of papers (2013-2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Create a way to determine the quality of the data</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>II</td>
<td>Create a way to improve data quality</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>III</td>
<td>Create a way to resist low-quality data</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>IV</td>
<td>Use a technique to improve data quality</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>V</td>
<td>Use a robust technique with respect to low-quality data</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>VI</td>
<td>Use subset data to improve data quality</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>VII</td>
<td>Use a technique to determine the quality of data and show concern about the results</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>VIII</td>
<td>Mention data quality</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>64</td>
<td>38</td>
</tr>
</tbody>
</table>

discussing the terminology used in the definitions of the data quality issues in detail. We then describe the dimensions of the data quality issues, followed by a presentation of the examples of data quality issues and a description of how they occur in the data sets. Next, we discuss studies that consider data quality in their research to ensure the credibility of the results. Finally, we present some approaches that deal with poor-quality data in data sets, including a review of the current data quality model in software engineering.

3.5.1 Definitions of data quality issues

Data quality issues in a data set are unexpected conditions that present obstacles to researchers using the data set. Some data quality issues may be introduced when the data are being collected from the source of data and during the integration of the processed data from software engineering tools. This involves the raw and refined phases in the research data life cycle as presented in Figure 3.1 in Section 3.3.6.1.

Many studies in the literature have used different definitions of data quality issues to describe quality issues with their data sets [1, 8, 9, 11, 56–59]. Table 3.15 presents some of the definitions of data quality issues from the literature.

Table 3.15 shows that the definitions of data quality issues vary depending on the context of use and the research domain. Some definitions of data quality issues are defined in a general way, such as ‘wrong information’ [57] and ‘data corruption’ [58]. Others are defined in the specific context of data, such as ‘records’ [56] and ‘cases’ [9].

In terms of different research or domain contexts, there are two definitions for duplicate data in Table 3.15. The first definition is for defect prediction [8], and the second definition is for a database [56]. As shown, both definitions are completely different and use different terms, such as ‘points’, ‘records’, ‘value’ and ‘field’. Although
### Table 3.15: Definitions of data quality issues

<table>
<thead>
<tr>
<th>Sources</th>
<th>Terms</th>
<th>Domain (or Context)</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gray et al. [8]</td>
<td>Duplicate data</td>
<td>Defect prediction</td>
<td>Data points (or instances) that appear more than once within a data set</td>
</tr>
<tr>
<td>Galhardas et al. [56]</td>
<td>Exact duplicate data</td>
<td>Database</td>
<td>Different records have the same value in a field (or a combination of fields) for which duplicate values are not allowed.</td>
</tr>
<tr>
<td>Shepperd et al. [9]</td>
<td>Identical cases</td>
<td>Defect prediction</td>
<td>Two or more cases contain identical values for all features including class label</td>
</tr>
<tr>
<td>Manago and Kodratoff [57]</td>
<td>Noise</td>
<td>Machine learning</td>
<td>Wrong information, incomplete information or unreliable information</td>
</tr>
<tr>
<td>Yoon and Le-Do [58]</td>
<td>Noise</td>
<td>Effort estimation</td>
<td>General data corruptions that may cause a negative effect on the performance of a model built on the historical data.</td>
</tr>
<tr>
<td>Shepperd et al. [1]</td>
<td>Noisy data</td>
<td>Software engineering</td>
<td>A recorded value of data that is not the same value with its true value</td>
</tr>
<tr>
<td>Rahm and Do [59]</td>
<td>Missing value</td>
<td>Database</td>
<td>Unavailable values during data entry</td>
</tr>
<tr>
<td>Bosu and Mac-Donell [11]</td>
<td>Missing data</td>
<td>Empirical software engineering</td>
<td>Not able to be found because a value is present but not in its expected place, or is not present or included when it is expected.</td>
</tr>
</tbody>
</table>

These terms appear to be similar, we do not know whether they refer to the same thing. Further, it is not clear whether these two definitions refer to the same quality issue.

The definition for ‘duplicate data’ in a database by Galhardas [56] uses terms that can be interpreted in different ways. For example, the term ‘record’ is not necessarily the same as a row or a record in a database. It can also refer to a collection of values in a data set or a basic data structure in programming languages. Every term in the definitions of data quality issues may have its own meaning in many circumstances, such as ‘database’, ‘data mining’ and ‘programming languages’.

Some definitions of data quality issues in Table 3.15 are proposed by taxonomies of data quality issues. These taxonomies act as references to identify data quality issues in a specific context, such as database [59] and financial data [60]. Some taxonomies provide classification for general data quality issues, but none of them are constructed to be used in the specific context of data sets. However, some terminologies from these taxonomies can be adapted in constructing definitions for data quality issues in a data set.
For example, Bosu and MacDonell proposed a taxonomy of data quality challenges for empirical software engineering [61]. They classified data quality challenges into three general areas: accuracy, relevance and provenance. In accuracy, they included definitions for some data quality issues and discussed how they have been identified in empirical research. These quality issues are: noise, outliers, data incompleteness, data inconsistency and redundant data. The explanation for data quality issues in this taxonomy is particularly useful in helping us to explore further common terminologies and definitions that can be used for constructing definitions of data quality issues in software engineering.

Overall, there is evidence to indicate that there are no standard definitions for data quality issues in software engineering data sets. Further, there is no clear terminology in the definitions used for data quality issues. Our review of the definitions suggests that there is a need to determine a standard way to construct clear definitions for data quality issues. This will enable researchers to apply a common interpretation of data in describing the quality issues in data sets.

3.5.2 Examples of data quality issues in software engineering data sets

Defect prediction and effort estimation are two active research areas that often generate software engineering data sets to be used by the research community. These data sets are useful for making accurate predictions and estimations to improve the quality of software. However, many researchers have found data quality issues in these data sets, such as missing data, noisy data, redundant data and inconsistent data [7, 8, 14, 21].

Software defect prediction uses change logs and bug reports to provide useful information about the evolution of software projects. Software developers are responsible for managing the change logs and bug reports in the software configuration management activities. However, it is possible for developers to make mistakes when they write keywords or do not write any specific keywords in the change logs that link to the bug reports. This will cause quality issues such as missing data if the number of linked bugs does not match the number of total fixed bugs reported by the developers.

For example, Bachmann et al. analysed five open source data sets and found missing data in the change logs and bug reports [7]. They discovered that both primary sources were untraceable because of missing links in bug reports or empty messages in the change logs. Another similar study by Bachmann and Bernstein investigated correlations between process data quality and characteristics and software product quality using six open source and two closed source data sets [34]. They highlighted that Eclipse has the biggest problem with empty commit messages, as 20% of the total commit messages were empty due to missing information related to the fixed bugs. It seems that these two studies used the term ‘missing data’ in reporting the issues with missing information either in change logs or bug reports.

In another example of defect prediction, Kim et al. proposed a method to measure the effect of noise in defect prediction models [14]. Before implementing the method, they performed a preliminary study to investigate the level of noise in the Eclipse SWT
component. They found that 32% were unlinked bugs in Eclipse 3.0 and 21% were unlinked bugs in Eclipse 3.1 that did not exist in the Concurrent Version System (CVS) logs. They identified the unlinked bugs found in the Eclipse SWT component as ‘noisy data’. This indicates that the authors used the term ‘noisy data’ to refer to missing information to link the bugs with the change logs.

In terms of quality issues in bug reports, Sureka et al. and Rongxin et al. reported that bug reports frequently contain source code segments and system error messages that are not in natural language text [62, 63]. Sureka et al. conducted a study on fault localisation based on an information retrieval model. They used the term ‘noisy text’ to indicate data quality issues such as spelling, grammar and sentence errors that were found in the bug title and description from two open source data sets. For example, they found a bug title that used a misspelled word ‘Managment’ in a bug ID JBAS-1552 from JBoss [62].

Rongxin et al. also found that the format of change logs varies in describing bug references whether across different projects or within the same project from two open source data sets. For example, they used the term ‘irregular bug reference format’ to indicate the quality issues related with the regular expressions used in change logs such as ‘solve problem 681’, ‘Fixed for #239’, ‘see #149’ and ‘for Issue 14’ in ZXing and OpenIntents project [63]. We speculate that the two studies by Sureka et al. and Rongxin et al. described the same kinds of quality issues that related to errors in the bug reports.

We noticed that the examples of studies that described data quality issues in defect prediction do not provide explicit definitions for them. This may explain why some studies have used different terms to describe the same quality issues. In particular, Kim et al. used the term ‘noisy data’ and Bachmann et al. used the term ‘missing data’ to describe the same quality issue—that is, missing information either in change logs or bug reports. In terms of incorrect data reported in bug reports or change logs, Sureka et al. used the term ‘noisy text’ and Rongxin et al. used the term ‘irregular bug reference format’. These examples provide evidence that we need to have not only standard definitions to describe data quality issues, but also standard terms to indicate data quality issues in data sets.

Another important research area is software effort estimation. This research area uses software project data that contains data related to software development to estimate the effort and cost of software projects. A project manager estimates the effort and cost of a new project using historical project data. Generally, historical project data contain quality issues such as ‘noise’, ‘inconsistent data’ and ‘missing data’ that would affect the performance of the software effort estimation [49].

For example, Liebchen and Shepperd analysed the productivity of a large data set and found inconsistent data relating to the project’s size attribute [32]. Project size was recorded in LOC or function points (FPs) or both in the data set. This indicates that the data set has two attributes to represent the size of projects, which makes it difficult to compare the actual size of the projects. However, the authors have selected projects that used FPs for their final analysis because they aimed to measure the variation of programming languages and coding styles [32].
Kitchenham and Mendes identified a number of problems with studies that compared different estimation models [21]. One problem that attracted our attention was missing values in the data sets. They also mentioned that empirical studies often use two or three data sets to validate the new estimation techniques, and they failed to justify the rationale for using the data sets. This tends to create a validation risk because no one can be sure whether the selected data sets are the most appropriate and reliable. Kitchenham and Mendes suggested that researchers need to justify their selection of data sets to support the validation of the proposed techniques [21].

This section provided examples of data quality issues in defect data and effort data to demonstrate how quality issues occur in data sets and how they might affect data analysis in software engineering research. It is important for researchers to be aware of the types of quality issues that could occur because they may affect their techniques, thereby leading to questionable results in the prediction and estimation models. These examples of data quality issues also provide evidence that there are no standard definitions for data quality issues in data sets.

### 3.5.3 Dimension of data quality issues

In the previous section, we described examples of data quality issues in data sets that might affect data analysis in software engineering research. Similar kinds of data quality issues also occur in other research areas, such as information systems, management and statistics. In some of these areas, data quality is defined by dimensions of data quality that are appropriate for specific contexts. In this section, we describe the dimensions of data quality that exist in the research area related to software engineering.

Over the past decades, the quality of data in information systems has been measured along many dimensions [25, 26, 64–66]. Wand and Wang have conducted a study to define data quality dimensions according to ontological foundations. They analyse data quality from two views: information systems and direct observation. The results of the study identify and define four dimensions of data quality that determine whether the data are complete, meaningful, unambiguous and correct [64].

Wang and Strong conducted an empirical study to understand data quality dimensions from the point of view of the user using the data. They identify 15 dimensions of data quality that are grouped into four categories: intrinsic (dimensions: accuracy, objectivity, believability and reputation), contextual (dimensions: value-added, relevancy, timeliness, completeness and appropriate amount of data), representational (dimensions: interpretability, ease of understanding, representational consistency and concise representation) and accessibility (dimensions: accessibility and access security) [65]. This study indicates that the dimensions are useful in order to highlight aspects of data collection, storage and use that have an impact on user perceptions of quality.

Redman identifies 27 dimensions of data quality from the perspective of data modelling and classifies them into three categories of dimensions: 1) data model (e.g., content, level of detail, composition); 2) data values (e.g., accuracy, completeness, currency); 3) data presentation (e.g., appropriateness and interpretability). Redman also groups data
quality issues into four categories: issues with data views (e.g., relevancy, granularity and level of detail), issues with data values (e.g., accuracy, consistency, currency and completeness), issues with the presentation of data (e.g., the appropriateness of the format and ease of interpretation) and other issues, such as security, privacy and ownership [66].

Batini et al. analyse the literature on data quality for the assessment and improvement of information systems and report on 12 methodologies, more than 50 dimensions of data quality and 100 relevant metrics. For the dimensions of data quality, they indicate that the most relevant dimensions are accuracy, completeness and consistency. The study proposed definitions of these dimensions: accuracy is ‘the closeness between a value \( v \) and a value \( \hat{v} \)’, considered as the correct representation of the real-life phenomenon that \( v \) aims to represent’, completeness is ‘the extent to which data are of sufficient breadth, depth, and scope for the task at hand’ and consistency is ‘the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file’.

Batini et al. also mention some other dimensions that have been proposed in the literature: currency, timeliness, interpretability, volatility and accessibility [25, 67]. The study proposes definitions of these dimensions: currency ‘concerns how promptly data are updated’, timeliness ‘expresses how current data are for the task at hand’, volatility ‘characterizes the frequency with which data vary in time’ and accessibility ‘measures the ability of the user to access the data from his or her own culture, physical status/functions, and technologies available’ [25].

For interpretability, the study indicates that this dimension ‘concerns the documentation and metadata that are available to correctly interpret the meaning and properties of data sources’. The study also describes the types of documentation, such as ‘the conceptual schema of the file or database’, ‘a set of metadata for cross-domain information resource description’, ‘a certificate contains measures description’ and ‘information on the history and provenance of data’ [25].

In 2014, Valverde et al. conducted a study in which a data quality model was applied to experiments in software engineering, they selected six dimensions of data quality proposed by Wang and Strong, Batini et al. and Scannapieco et al. [25, 68, 69]. The six dimensions of data quality are accuracy, completeness, consistency, uniqueness, representation and interpretability. The study summarises the definition of each selected dimension and proposes the following quality factors [70]:

(i) Accuracy: ‘Specifies how accurate and valid the data is. It indicates if a correct association between the Information System (IS) states and the real world objects exist. Three quality factors are proposed: semantic accuracy, syntactic accuracy and precision’.

(ii) Completeness: ‘Specifies if the IS contains all the important data, with the required scope and depth. It indicates the IS capacity to represent all the significant states of the reality through two factors: coverage and density’.
(iii) Consistency: ‘Specifies if the semantic rules are satisfied in the IS. An inconsistency exists when more than one state in the IS is associated with the same object in the real world. There are different kinds of integrity restrictions: domain restrictions and intra-relation restrictions’.

(iv) Uniqueness: ‘Specifies the duplication level of the data through two factors: duplication occurs when the same entity is duplicated exactly, while contradiction occurs when the entity is duplicated with contradictions’.

(v) Representation: ‘Considers the consistent and concise representation of the data in the IS, and the extent in which data is always represented in the same format and structure. Two quality factors are proposed: data format and data structure’.

(vi) Interpretability. ‘Refers to the documentation and metadata available in order to correctly interpret the meaning and properties of the IS. Two quality factors are proposed: ease of understanding and metadata’.

In a recent study by Batini and Scannapieco, the dimensions of data quality were refined and grouped into eight clusters to indicate some of the similar dimensions. The clusters of similar dimensions are accuracy, completeness, consistency, redundancy, readability, accessibility, usefulness and trust. The study constructs definitions of each cluster according to the relevant dimensions [26].

As described in previous sections, much of the research in the area of software engineering has been concerned with quality issues, such as noise, incompleteness, redundancy and inconsistency [13, 14, 21, 49, 61]. For example, Liebchen et al. conduct experiments to measure the effect of noise on a predictive model and to evaluate the accuracy of techniques for handling noise in data sets [33]. Bosu and MacDonell identify a range of quality issues (noise, incompleteness, redundancy, inconsistency and outliers) and classify them into three categories: accuracy, relevance and provenance [61]. Some of these quality issues are related to the dimensions of data quality described in the information system (e.g., the issue with data values described by Redman [66] and the dimensions proposed by Batini et al. [25]).

Our research aims to identify the quality issues related to the interpretation of data and the common quality issues (incorrect data, missing data, inconsistent data and duplicate data) that frequently appear in data sets from public data repositories. To capture the quality issues related to the interpretation of data, the most appropriate data quality dimension is interpretability. We adapted the definition of interpretability proposed by Valverde et al. [70] to better fit our research context, that is, it refers to the metadata in data sets to correctly interpret their meaning. Further, to capture the common quality issues, we have selected the appropriate data quality dimensions and the definitions proposed by Valverde et al. [70]. The data quality dimensions are accuracy, completeness, consistency and uniqueness.
3.5. REVIEW OF DATA QUALITY ISSUES

3.5.4 Data quality studies

Thus far we have discussed related work on the definitions of data quality issues, examples of data quality issues in the data sets and dimensions of data quality issues. In this section, we discuss studies that take data quality seriously in research based on data sets. We discuss studies that have identified and eliminated data with the quality issues from their analysis to ensure the credibility of their research results [8, 49, 71, 72]. We also discuss studies that show initiatives to classify information about data sets in public data repositories. Most of the studies have been considered in our mapping study and these studies used data sets from data repositories that provide public access to the research community.

For example, Do et al. and Li et al. conducted a study on improving the accuracy of software effort estimation [49, 71]. Both studies used software project data from the Desharnais and ISBSG Telecom data sets, and they used pre-processing techniques to filter out the quality issues. They found missing values in four out of 81 projects’ data in the data sets. They excluded the four projects to ensure accuracy of the estimation model. The results showed a significant improvement for the estimation model without the noisy projects.

In another example, Koru et al. [72] tested the theory of relative defect proneness in a number of data sets from the NASA MDP in the PROMISE repository. They found 4,841 data points with zero LOC in the MC1 data set. Generally, if a module contains zero lines of code and zero operands and operators, it should be an empty module containing zero lines of code. Koru et al. considered this to be a data quality issue because the value for LOC was not a valid value. Thus, they have excluded the MC1 data set from their experiments and the results supported the theory of defect proneness.

A number of studies have reported that many data sets from the NASA MDP had data quality issues either from the original NASA MDP repository or the PROMISE repository [8–10, 73, 74]. Some of the studies mentioned the data quality issues in the data sets without further explanation and excluded these data sets from their experiments. For example, Kaminsky and Boetticher [74] used a KC2 data set and found redundant data in the KC2 data set. They excluded these data from the experiments and the results were statistically significant.

In another example of a study that excluded data sets from the experiments, Boetticher [73] removed the duplicate tuples and invalid tuples (e.g. LOC value is 1.1) from five NASA data sets in their experiments. Boetticher claimed that the results of the experiments without duplicate tuples decreased significantly for the performance of the classification model compared to the results of the experiments with the duplicate tuples. This indicates that the existence of duplicate tuples in the data sets may influence the results of the research.

The studies presented above have provided evidence that data quality issues in data sets may affect research results. The studies have demonstrated that research results improve by excluding data with quality issues. Thus, the results are more reliable than studies that use data sets with quality issues in their analysis.
3.5. REVIEW OF DATA QUALITY ISSUES

A number of studies have used data sets from the NASA MDP in their analysis without knowing the existence of data quality issues [17, 75, 76]. This is because the NASA repository does not explicitly publicise quality issues in their data sets. In addition, these studies did not conduct any pre-processing techniques to filter out the data quality issues. Coincidentally, some studies produced results that were positively good, but the results were questionable due to the quality issues. The results of these studies would have been more promising if the authors were aware of the quality issues and had taken data quality issues seriously in the analysis.

For example, Zhang investigated the relationship between LOC and defects using the data sets from Eclipse and NASA MDP. The author made no attempt to identify quality issues in the data sets before implementing the experiments. However, he mentioned that the validity of the results depended on the quality of the data sets. If the quality of the data sets is poor, their results may be invalid [17]. Thus, the author believed that the data quality issues may have affected the results; however, no action was taken to consider the quality of the data sets.

In recent years, few studies surveyed public data repositories and categorised them according to the content of data sets [10, 22]. Rodriguez et al. surveyed 12 public data repositories and classified the repositories into five main categories: the type of information stored; public or private availability of the dataset; existence of single project or multi-project data; type of content; and the format of data storage. They also classify the data quality issues that are faced by researchers when using these repositories. The quality issues are classified into issues related to data extraction, replicability and machine learning (e.g., missing values and inconsistencies, outliers, redundant and irrelevant attributes). The results presented that some repositories do have similar content in terms of the format of data storage and type of information stored.

In another similar study, Cheikhi and Abran presented a structured overview for the content of data sets in two public data repositories: ISBSG and PROMISE [22]. The purpose of this study is to make it easier for researchers to understand the information provided in the data repositories and able to use the data sets in modelling more efficiently. They classified the content of data sets into eight categories: the topics addressed, the source of the data sets, the year in which the data set was originally donated, the availability, the attribute description, the type of software project used by the data set, the number of attributes, and the number of instances. The results presented some data sets do contained similar kind of information (e.g., the availability, the attribute description). The results in Cheikhi and Abran study are consistent with the results from Rodriguez et al. [10] in terms of some repositories do contain similar content of data sets. This shows that it is possible for our research to design a standard way of describing the content of data sets for data repositories.

Considering all of this evidence, it appears that the research community needs to seriously consider quality issues in data sets because they affect research results. This is supported by the findings of our mapping study, in which a few studies used a technique to determine the quality of the data and concern about the results. Our review of data
quality studies suggests that there is a need to identify quality issues in data sets because some studies' results have improved by excluding poor-quality data in their analysis. Further, few studies that surveyed public data repositories presented some similar kinds of information in the data sets can be described in a standard way.

3.5.5 Dealing with poor quality of data sets

Motivations from the systematic review performed by Liebchen and Shepperd [12], and the targeted review performed by Bosu and MacDonell [11], have attracted a wider audience on approaches to improve the quality of data that influence empirical results. These approaches have focused generally on identifying quality issues as well as dealing with and assessing the quality of data. In this section, we present approaches to deal with data quality issues in software engineering or try to improve the quality of data in the future.

Cartwright et al. examined two imputation techniques for dealing with the problem of missing data in software engineering [77]. To assess the usefulness of software engineering data that contained the problem of missing data, they applied the imputation technique to two commercial software project data sets. In the analysis, they found that the k-NN imputation method generated the best results and may have some practical utility for the effort prediction model. However, Cartwright et al. pointed out that the software engineering data were still not fully accurate, even when the imputation techniques were applied. Thus, this raises the question of the usefulness of any prediction model that relies on such data. The authors suggested that the research community needs to observe the quality of data collection before using the data for analysis [77].

Mockus conducted a study about how to deal with missing data in software engineering [5]. He highlighted the need to determine the mechanism to identify missing data and to add observations that may clarify why the values are missing. He applied several missing data techniques to three data sets from a case study and found that different conclusions may be reached depending on the particular techniques used. He suggested that future research needs to choose suitable methods to deal with missing data in order to obtain promising empirical results.

Bachman et al. conducted a study on data quality using software engineering data and highlighted the need to ensure accurate data quality at its origin (i.e., from the data source) [7, 34]. The authors enhanced a technique to prepare software engineering process data and defined a data quality framework that included data quality and characteristics' metrics to evaluate the quality of software engineering data. They prepared software engineering process data and analysed data from six open source and two closed source software projects. However, they still found data quality issues in all of the software projects that may have significant effects on the research results. As noted by Bachman et al. [7], the data source needs to provide better data quality to ensure promising research results. However, there are limits to how far the data source can manage this issue due to the complexity of software engineering processes. The findings would have been more useful if they included a pre-processing technique for process
data quality before preparing the software engineering data.

The ISBSG repository has established data quality metadata criteria that aim to analyse specific data quality criteria for each submitted data set. These criteria describe the perceived quality of each data set and use it to filter out low-quality data from the analysis. Low-quality data are identified as data sets that have many missing values.

Marta et al. conducted a study that investigated the influence of data quality metadata criteria on the behaviour of empirical results [78]. The analysis started by applying the data selection criteria to 5,052 projects contained in the ISBSG data set Release 11. The data selection criteria screening focuses on the problem of missingness for the evolution of the projects. Some of the data selection criteria include ‘project implementation date known’, ‘development team effort known’, ‘effort across the whole life cycle’ and ‘unadjusted function points known’. After the data selection criteria screening, they considered 830 projects for the data quality metadata criteria screening. As mentioned earlier, the data quality metadata criteria screening focus on issues related to missing values. Of 830 projects, 262 were considered for further analysis in their study. Marta et al. noticed the data quality metadata criteria and the data selection criteria concerns were redundant because both criteria focused on the same problem (i.e., missingness). They noted that the analysis of the study might have little effect. However, one question that needs to be asked is whether the data selection criteria or data quality metadata criteria are important variables to use to determine which data sets are more reliable. In fact, ISBSG needs to create awareness to show how the data quality metadata criteria may lead to promising empirical results.

In 2011, Gray et al. conducted a study that proposed a data cleaning method to clean 13 NASA data sets that had quality issues [8]. For example, they found two identical attributes that had the same values but different labels in the KC4 data set. The labels were number of lines and loc total, and these labels were poorly defined in the NASA repository. This indicates an issue with insufficient metadata to interpret the meaning of the attributes. If there were complete metadata that explicitly described the meaning of the attributes, then we could interpret whether the two identical attributes represented two different metrics or the same metrics. In addition, Gray et al. found seven data sets in the NASA data sets that contained missing values in a decision density attribute [8]. This attribute had an empty value because of a division by zero error that was generated by two attributes: condition count and decision count. The seven data sets were created using the same software metrics tool because they had the same structure of attributes. This error may have occurred because the software metrics tool did not capture the complete conditional statement to derive a meaningful value for the attributes of the seven data sets. The error might not have occurred if the software metrics tool had captured the complete metadata for the attributes in the seven data sets. Gray et al. concluded that the proposed cleaning method was useful to ensure the NASA data sets were suitable for research based on data sets [8]. They also highlighted that any experiments based on the data sets without the pre-processing method to clean them may have led to erroneous results.
Shepperd et al. [9] responded to a study by Gray et al. [8], who mentioned the quality issues in the 13 NASA data sets that were currently in use by researchers. Shepperd et al. investigated two versions of the NASA data sets: one from the NASA repository and the second from the PROMISE repository, and they identified the different types of data quality issues. They categorised the data quality issues into five categories: identical features, constant features, features with missing values, features with conflicting values and features with implausible values. They developed a pre-processing algorithm to address the data quality issues in the two versions of the data sets, and they published new versions of the data sets in the PROMISE repository. In their detailed investigation, Shepperd et al. indicated that the two versions of the data sets had a small difference due to different pre-processing and version control issues, and that the difference may influence the variance of the research results [9]. They also urged the research community to pay attention to the metadata of data sets instead of the performance of computational methods because the research results greatly depend on the meaning of the data sets. The research results would be more meaningful if the metadata were captured in detail using a standard way of documentation to allow researchers to apply a common interpretation of the data sets. We believe that one possible standard way could be a metamodel.

This section has reviewed a number of studies that have explicitly considered that the quality of data sets may influence research results. Some studies have developed new methods to improve the quality of data sets, while others have applied existing statistical methods to deal with quality issues in data sets. We noticed that despite the discussion in Chapter 2 of the importance of metadata, this concern rarely came up in the reviewed studies. Only three studies indirectly mentioned metadata but did not take serious action. This indicates a gap in data quality for software engineering research, and this thesis will attempt to fill the gap by considering the metadata in assessing the quality of data sets.

3.5.6 Review of current data quality model in software engineering

In a recent study, Valverde et al. proposed a data quality model and a systematic approach to analyse and improve data quality in software engineering experiments [70]. The data quality model and the systematic approach are designed based on a data quality metamodel proposed by Etcheverry et al. [79]. The quality model comprises the following data quality concepts: dimensions, factors and metrics. The data quality dimensions represent a general classification of data quality issues, data quality factors represent particular aspects of the dimensions and data quality metrics are instruments used to assess the data quality factors and indicate the presence or absence of data quality issues (see Figure 3.7).

As described earlier, in the Valverde et al. study, the data quality dimensions are selected from data quality research areas (e.g., information systems) [25, 68, 80] that are applicable to software engineering experiments data. The dimensions are accuracy, completeness, consistency, uniqueness, representation and interpretability. The study proposes 14 data quality factors based on the dimensions, for example, the accuracy dimension contains three data quality factors: syntactic accuracy, semantic accuracy and
precision. Based on the quality model, the study defines 21 data quality metrics that can be used in experiments. The data quality model and an example of its application is shown in Figure 3.7.

Figure 3.7: Data quality model and its application (Valverde et al. [70]).

The purpose of the data quality model is to support the identification and assessment of quality problems associated with the collection of data from software engineering experiments. Valverde et al. classify the quality problems into three categories: data errors, questionable values and improvement opportunities. Data errors ‘correspond to errors in the data that (whenever possible) need correction’. Questionable values refer to situations in which ‘it is not possible to assure if the data quality problem corresponds to a real data error (examples of these are outliers)’. Improvement opportunities ‘correspond to suggestions of aspects that could be improved in order to prevent the occurrence of a data quality problem in the future’ [70].

Valverde et al. conduct two controlled experiments that compare the effort required to develop a web application by applying either a traditional development approach or a model-driven development approach. The data collected consist of experimental material (e.g., documentation, experiment context and conceptual model) and the data from questionnaires. All these data are recorded in spreadsheets.

The study applies 16 of the 21 data quality metrics that are defined in the data quality model. The overall results of the experiments show that data quality problems are present in 9 metrics of the base experiment and 10 metrics of the replication experiment. The study proposes preventive actions for each data quality problem present in the metrics to avoid its occurrence in future experiments. The study concludes that the proposed quality model was able to assess the quality of the experimental data used and that the identified data quality problems could affect the experimental results [70].

While the purpose of the Valverde et al. study is similar to our research, that is, to evaluate the quality of data in software engineering, the dimensions of the studies are distinct because the problems that these studies are trying to solve are different. The Valverde et al. study focuses on six main data quality dimensions (accuracy, completeness, consistency, uniqueness, representation and interpretability) to identify quality issues associated with the data collected in software engineering experiments, while
our research focuses on interpretability as the primary dimension because it aims to identify quality issues related to the interpretation of data. Apart from the interpretability, our research consider another four data quality dimensions (accuracy, completeness, consistency, redundancy) as the secondary data quality dimensions to assist researchers in identifying common quality issues that frequently appear in data sets.

While the Valverde et al. study uses data from software engineering experiments that are similar to the software engineering data sets used in this study, they differ in terms of the format and structure from the data used in this study, as described in Chapter 2. Valverde et al. state that the experimental data are recorded in spreadsheets, but they do not specify how to interpret the data in the context of the spreadsheets. We could assume that some data may be structured in tabular format and some data in other structures (e.g., plain text for the questionnaire data) because the experiments contain many different kinds of data.

The studies differ further in that Valverde et al. present a data quality model that allows researchers to predefine metrics for the identification of quality problems in software engineering experiment data (these metrics can be measured using a measurement method defined by the authors). The data quality assessment framework undertaken in our research, however, goes beyond the quality model as it allow researchers to determine whether a data set contains sufficient information to facilitate the correct interpretation of data for analysis in empirical research. The framework consists of two main processes: the modelling of the data set (described in Chapter 5) and the quality assessment (described in Chapter 6).

Finally, the approach proposed by Valverde et al. does not support the independent evaluation of the quality of data as it requires experimenters (who conduct the experiments) and a data quality analyst to perform the approach, while the data quality assessment framework proposed in our research provides a systematic approach that allows researchers to evaluate the quality of data sets independently. Such a systematic approach should be beneficial in terms of the facilitation of replication, as our research considers data sets from public data repositories.

3.6 Discussion and conclusions

The studies presented in this chapter have provided evidence that the research community appears to have recognised the critical role played by data quality in software engineering. In section 3.2, we described the systematic literature review studies that have informed the research community regarding the importance of data quality in software engineering. We then described the systematic mapping study that we carried out and presented the results based on the synthesis of research on data quality in software engineering in section 3.3. In the mapping study, we found that researchers have used unclear and inconsistent terminology to describe data quality issues, which can be interpreted in many different ways.

To illustrate the potential of this finding from the mapping study, we discussed
3.6. DISCUSSION AND CONCLUSIONS

the definitions of data quality issues from the literature and presented a number of studies in which these quality issues were found in section 3.5. These studies considered quality issues such as noise, missing data, duplicate data or inconsistent data. However, we noticed that one particular quality issue had not received much attention from researchers: data was not able to be interpreted correctly due to insufficient metadata. This quality issue may lead to incorrect conclusions if researchers misinterpret the data.

Our review of the definitions of data quality issues indicated that there are no standard definitions of data quality issues in software engineering data sets, thus suggesting a need to construct a formal definition of data quality issues. The formal definition will require a standard terminology to specify the quality issues across the different structures and formats of data sets. This standard terminology can be represented using a standard model.

In section 3.5.3, we described a number of information system studies that used data quality dimensions to measure the quality of data. Some studies define data quality dimensions that are appropriate for the context of their studies. Other studies propose data quality dimensions based on data quality research in information systems. In section 3.5.4, we discussed a number of studies that take data quality seriously to ensure the reliability of research results. Some of these studies eliminated quality issues in data sets in order to establish meaningful results; however, a few studies did not use pre-processing techniques to eliminate quality issues from their analysis and these issues would have affected their research results.

Section 3.5.5 consisted of a discussion of the studies that have addressed data quality issues. A few examples were presented to show the practicality of the approaches in dealing with and assessing the quality of data. It appears that researchers tend to focus on the effect of poor data quality on empirical results. This is undoubtedly a critical issue because if the data are flawed, the results will be questionable. In section 3.5.6, we discussed a study that proposes a data quality model and systematic approach to the analysis and evaluation of the quality of data that were collected from a software engineering experiment. The study is concerned with solving quality problems related to the data collected from their experiments; they focus on six data quality dimensions, which differ from those used in our research.

The ultimate goal of our research is to evaluate the quality of data sets. In order to make an accurate evaluation, we need a standard way to identify quality issues in data sets. As discussed earlier in this chapter, the research community has often used unclear and inconsistent terminology, not only in reporting quality issues, but also in the definitions of data quality issues. In Chapter 2, we found that existing data sets have different formats and structures, which create challenges in terms of understanding what the data sets are actually intended to represent. The inconsistent use of terminology and the varying formats and structure of data affect the interpretation of data sets and make it difficult to report clearly on the quality of the data.

Therefore, our research aims to provide a standard way to better understand and interpret data sets and to identify any potential problems a data set may have. We want
to develop a framework that incorporates a metamodel to enable researchers to apply a common interpretation to the description of the structure of data sets, as well as a quality assessment process to evaluate the quality of data sets. This framework will assist researchers in understanding the quality of the data sets used in their empirical research. The next chapter describes the design of metamodel and the process for modeling data sets as the first step in developing a framework for data quality assessment.
This chapter describes the full specifications of the metamodel for modelling data sets. It provides details on the definition of data sets and their potential elements. A metamodel is presented to describe the structure and concepts in a data set, as well as the relationships between each concept. This chapter also introduces the notion of an extended data set to describe information relevant to data sets. This is followed by a discussion of the application of the metamodel for modelling data sets from data repositories.

4.1 Introduction

Previous chapters described why researchers should care about the quality of data sets in software engineering research. Chapter 1 discussed the importance of data quality in data sets and the problem to be investigated. Chapter 2 described the different structures and formats of data sets. Chapter 3 presented a review of related research on data quality issues in software engineering research, which indicated that there were no standard definitions for data quality issues and little research available considering the importance of metadata in software engineering data sets.

These earlier findings provided insight we draw on to inform the development of a framework for data quality assessment. The framework covers two main processes: modelling the data sets and evaluating the quality of the data sets. This framework is designed to help researchers identify whether a data set contains sufficient information for analysis, and will also allow researchers to apply a common standard to evaluate the quality of data sets.

One way to evaluate the quality of a data set is to identify the kinds of potential quality issues it may have. Thus, researchers are required to be familiar with and understand the different kinds of data quality issues that can arise. However, as mentioned in Chapter 3, many definitions of data quality issues are unclear as a result of the variety of terminology used in these definitions. Therefore, a consistent interpretation of terminology that
clarifies the meaning of ambiguous data quality terms is required.

Using a formal definition is a suitable approach to clearly describe data quality issues in data sets. As a first step in the construction of the formal definition of data quality issues, we need to create a common agreement about the terminology and concepts in data sets, which should be defined clearly and consistently in order to describe data sets accurately.

While real data sets are shown to have many different formats and structures (as described in Chapter 2), there are some aspects of data that are commonly present across the varying structures of data sets. Therefore, it is possible to use a standard terminology to refer to these common aspects.

The need to use a standard terminology in the formal definition of data quality issues has motivated us to design a metamodel that describes the structure and concepts in a data set, as well as the relationships between each concept. Each concept in the metamodel is defined using standard terminology, which allows researchers to specify data quality issues consistently using these standard definitions. In addition, the metamodel allows researchers to gain a shared understanding of the content of data sets. The metamodel can be used as a basis for modelling data sets, particularly data sets that have quality issues.

In section 4.2 of this chapter, we review related work on conceptual and ontological modelling approaches to identify the approaches that are best suited to modelling data sets with quality issues. None of the reviewed approaches have considered data sets with quality issues, which justifies the need for the metamodel that is presented in section 4.3. We describe the process for modelling data sets in section 4.4. Finally, in section 4.5, we present the application of the metamodel to a number of existing data sets from the data repositories to demonstrate its utility.

4.2 Related work on conceptual and ontological modelling of data

Many software engineering studies on the approaches to the conceptual and ontological modelling of data and metadata have been published. Such approaches include the structural model for measurement [81, 82], the conceptual data model [83] and measurement ontology [84, 85]. Some approaches have attempted to improve the software measurement process by developing data models to capture the definitions and relationships involved in the measurement process [81–83]. Other approaches have proposed ontologies to harmonise the different software measurement proposals and standards [84, 85] by presenting a set of common concepts used in software measurement. Some of these approaches can also be used to define and model data sets in software projects.

Although these existing approaches are mature enough to describe the structure and concepts for the purposes of understanding the data, they assume that the data (e.g., the values and properties) stored in data sets are correct. However, some data sets do have quality issues [8, 10] and there are no existing models that deal with these kinds of data.
4.2. RELATED WORK ON CONCEPTUAL AND ONTOLOGICAL MODELLING OF DATA

sets. As mentioned above, this thesis proposes a metamodel that can be used to model data sets with quality issues. In the following subsections, we review five of the existing data modelling approaches and a data analysis tool.

4.2.1 Framework for software measurement validation

In 1995, Kitchenham et al. proposed a framework for validating software measurement [81] that focused on identifying elements of software measurement and their properties. The framework established theoretical and empirical methods to validate the properties of the elements of the measurements. The authors designed a measurement structure model that consisted of entity, attribute, value, unit, measurement instrument and scale type. It also defined a measurement protocol to ensure the consistent and repeatable measurement of a specific attribute on a specific entity.

Kitchenham et al. discussed the importance of the entity population model to handle data analysis or data interpretation in empirical research [81]. The entity population model allows researchers to make assumptions about the measurements by referring to the values of specific attributes for specific entities in the data sets. This indicates that the population information is essential to be able to interpret the measurements in the data sets.

Although Kitchenham et al. designed a measurement structure model that includes the elements of software measurement and highlights the importance of the entity population model for data analysis, their framework does not mention any concerns with modelling data sets that have quality issues. In particular, their proposed measurement structure model does not include elements that indicate the structure of a data set, such as a record. The recognition of this kind of element is essential, not only to indicate the specific element in the structure of data set, but also to help researchers to clearly specify data quality issues.

4.2.2 Model of software measurement data

In 2001, Kitchenham et al. extended their view of modelling data by proposing a method to model software data sets that captures the definitions and relationships among software measures [82]. The aim of this approach was to provide a standard structure to define software measures. The approach also aimed to allow researchers to model complex data sets that include values related to entities from different levels of granularity (such as values related to actual recorded value, estimated recorded value and target recorded value).

The method is illustrated using an Entity Relationship (ER) model, which consists of three main components: generic domain, development model domain and project domain. The generic and development model domains define metadata that are necessary to facilitate more complete and accurate reporting of software data sets. The generic domain defines attributes, units and scale types. The development model domain provides the relationships between the measures and the entities to provide content for measurement.
The project domain defines the actual collection of data that represent the recorded values collected from real software projects [82].

Although the ER model is able to model complex data sets that include the validation of values (e.g., invalid values or implausible values), it does not take account of data sets with quality issues, such as duplicate data. These kinds of quality issues require validation of specific aspects of data (e.g., records) in the structure of data sets. The ER model would be more relevant to our research if it were able to model data sets with quality issues.

4.2.3 Model for software measurement (MOSME)

A broader perspective of the measurement process was adopted by Chirinois et al. who introduced a data model to define software measures, and provide the required elements and their relationships, to achieve repeatable and reproducible measurement values [83]. It classifies the elements of data into three parts: the algebraic system (entity, entity type, attribute), measurement (measure, unit, scale type, project, target, actual and context of use) and the numerical system (numerical relation, numerical system, number).

The data model for software measurement (MOSME) aims to improve the software measurement process by providing software measure definitions to obtain the measurement values consistently during software development. It can be used to define and model data sets from several software projects. It also provides definition of elements of measurement, particularly the metadata, such as counting rule, conditions and context of use.

Chirinois et al. highlight the importance of capturing a well defined measurement process at the data collection stage to ensure that all elements involved are accurately recorded [83]. However, MOSME does not contain definitions of elements that are relevant to the structure of data sets (e.g., value, column header, record). In order to perform data analysis, researchers not only need to know the elements in the measurement process, but they also need to understand the complete structure of data sets. MOSME would have been a more convincing model if they had included the definitions of elements to describe the structure of data sets.

4.2.4 Ontology for software metrics and indicators

An early example of research into ontology in software measurement was conducted by Martin and Oslna. They proposed an ontology for software metrics and indicators that aimed to measure the quality of web applications, particularly for cataloguing web systems [84]. The ontology focuses on metric and indicator knowledge that is specified as specific concepts, attributes and relationships.

In general, an ontology clarifies the knowledge structure that defines the domain concepts and the relationships between them. As noted by Martin and Oslna, the sources of knowledge for the proposed ontology came from the literature, including their own work in metrics, and evaluation processes and method [84]. It is therefore apparent that
the ontology for software metrics and indicators was developed based on the knowledge belonging to the domain of the problem.

While the ontology describes the domain concepts and relationships for software metrics and indicators using a conceptual model, it does not model the structure of the data. If the structure of the data is not modelled properly, there is a possibility that the data is interpreted incorrectly.

4.2.5 Software measurement ontology (SMO)

A similar approach to ontological modelling is taken by Garcia et al. [85–87] who developed a software measurement ontology (SMO) containing a set of four sub-ontologies that define and conceptualise measurement concepts and quality, as provided by the IEEE Std. 610.12: ‘Standard Glossary of Software Engineering Terminolog’, IEEE Std. 1061-1998: ‘IEEE Standard for a Software Quality Measures Methodolog’, and ISO/IEC 15939: ‘Software engineering - Software measurement process’. The four sub-ontologies are: (a) Software measurement characterisation and objectives: to provide context and goals of measurement; (b) Software measures: describes the measure terminology; (c) Measurement approach: describes the ways to obtain measurement results; (d) Measurement: describes concepts related to the measurement process.

The SMO provides practitioners and researchers a standard terminology of measurement concepts to carry out the measurement process and record the results in a consistent way. It also illustrates the relationships between each concept. For example, the measurement sub-ontology defines a measurement (an action) as ‘a set of operations having the object of determining the values of a measurement result for a given attribute of an entity, using a measurement approach’ [85]. The measurement sub-ontology contains six concepts: entity, attribute, measurement, measurement approach, measure and measurement result.

Although the SMO contains the necessary measurement concepts, this ontology does not have elements that describe the structure of the data set. In particular, the SMO does not model metadata in the data set and there is, therefore, a potential problem with the interpretation of the data.

4.2.6 KEEL (Knowledge Extraction based on Evolutionary Learning) tool

Alcala-Fernandez et al. introduced a software tool for the analysis of evolutionary algorithms to solve the different kinds of problems in data mining research [88] (such as regression, classification and unsupervised learning). They named this software tool KEEL (Knowledge Extraction based on Evolutionary Learning). KEEL offers three modules for research and education: data management, design of experiments and educational experiments.

The data management module provides a set of features that can be used mainly to export and import a data set in other formats to the KEEL format in order to customise the structure of the data and to apply partitions to the data set. This module reduces the technical work and allows researchers to transform the existing format of a data set.
into the KEEL format to be used in the analysis of evolutionary algorithms. The design of experiments module defines the desired experimentation over selected data sets and provides options for data customisation. The educational experiments module allows researchers to learn the process of a certain model in designing the experiments for data mining problems.

The KEEL tool provides a feature for data preparation in the data management module. This feature aims to produce good quality data for analysis by allowing users of the tool to clean the data, transform the data and reduce the size of the data. However, this feature does not provide a way to capture and describe the metadata of the data set because it is developed specifically for data mining algorithms. There is therefore a possibility that users of the tool could interpret the data incorrectly.

4.3 Designing a dataset metamodel

In this section, we start by providing a precise definition of a measurement data set and measurement metadata. We then present a dataset metamodel to describe the structure and concepts in a data set, and the relationships between each concept. Each concept is defined using standard terminology. We introduce the notion of an extended data set to describe the information that is relevant to a data set that is distributed in locations other than the data set file.

4.3.1 Definitions of measurement data set and measurement metadata

Before presenting the definition of a measurement data set and measurement metadata, it is necessary to clarify exactly what is meant by measurement. In this thesis, we adopt the definition of measurement suggested by Fenton and Pfleeger [89]:

*The process by which numbers or symbols are assigned to attributes of entities in the real world in such a way so as to describe them according to clearly defined rules.*

The definition of measurement represents a process that captures information about the attributes of entities. From this definition, we define measurement values as values that are obtained through the process of measurement.

We provide the following definitions of a measurement data set and measurement metadata.

- A measurement data set is a collection of elements that include one or more measurement values.
- Measurement metadata is a set of elements that describe the entities and the metrics of the data set.

As described above, a data set consists of a collection of elements. An element can be a token or a collection of tokens. A token is a sequence of non-whitespace characters, and it is not necessary for a token to be delimited by a whitespace character.
4.3. DESIGNING A DATASET METAMODEL

A classic example of elements in a real data set is shown in section 2.2.1.3 Example 3: PROMISE KC2. The 10th line in this data set shows ‘% 1.loc :numeric % McCabe’s line count of code’. To determine the elements, the first step is to identify the tokens.

In this example, there are three reasonable interpretations: (1) six tokens and one collection of tokens which are ‘%, 1., loc, :, numeric, %, McCabe’s line count of code’; (2) five tokens and one collection of tokens which are ‘%, 1., loc, :, numeric, %, McCabe’s line count of code’; or (3) 11 tokens which are ‘%, 1., loc, :, numeric, %, McCabe’s, line, count, of, code’. Alternatively, we could identify every character as a token; however, most researchers do not interpret the tokens in this way. It is therefore apparent that the step to determine the tokens in a data set may require experience and can lead to more than one interpretation.

The next step is to determine the elements in this example that are appropriate for our metamodel. In the first interpretation, we view six tokens and one collection of tokens as consisting of seven elements, while in the second interpretation, we view five tokens and one collection of tokens as consisting of six elements. Both interpretations could be appropriate for our metamodel because we may classify some of the elements (e.g., ‘%, :’) into the same category of element in the metamodel (the dataset category is discussed in the next subsection).

For easy access to the elements in a data set, we define a data set in a single file. However, some real data sets consist of elements in multiple locations, such as a web page or a publication, rather than in a single file (as described in Section 2.2.1). We define this kind of data set as an extended data set in order to refer to all of the elements of the data set in all locations (the extended data set is discussed in Section 4.3.3.).

4.3.2 Dataset metamodel

We are interested in data sets that contain data pertaining to entities drawn from some populations. Every entity has data associated with it. We call this data the ‘characteristics’ of the entity, except in Chapter 2 where they were identified as ‘properties of the entity’ to represent the column headers for every example of data set.

In this chapter, we divide the characteristics into attributes and properties. Attributes are characteristics associated with metrics (e.g., length of a program that measure with the number of lines of code). Properties are characteristics that are not associated with metrics (e.g., name of a source code file).

Data sets may contain other data that do not come from attributes and properties. Some aspects of the data set may be there simply to provide structure (e.g., a string that indicates the beginning of a section). Some may help to organise the values (e.g., column headings that effectively provide names for the characteristics). Other aspects may help to explain what the characteristics are, or what the values mean (e.g., descriptions that provide more detailed explanations of the characteristics than just the column headings). Such aspects may also provide detailed information about the entities involved, or the population the sample comes from.
To model a data set, we needed to identify all the concepts mentioned above. Some of the concepts are abstract (e.g., an attribute) and so do not explicitly appear in a data set, but there are parts of the data set that may correspond to those concepts. Other concepts do appear explicitly, but we wanted to be able to distinguish their different aspects (e.g., some values are measurement values while others are not).

Figure 4.1 shows our metamodel for a data set, using Unified Modelling Language (UML) notation (see Appendix B for large size). The metamodel contains three levels: physical structure, dataset category and dataset concepts. It captures the features of the above discussion of data sets in more precise detail.

A data set consists of things. We have called these things elements. We have divided an element into four categories because we want to distinguish the different aspects of a data set for better interpretation of data. We have used standard terminology for the four categories of element, which are value, label, metadata, and ancillary.

We provide precise definitions for the physical structure as shown in Table 4.1.

<table>
<thead>
<tr>
<th>Element category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>An element that is used to represent a characteristic.</td>
</tr>
<tr>
<td>Value</td>
<td>An element that is recorded about a characteristic.</td>
</tr>
<tr>
<td>Metadata</td>
<td>A set of elements that represent a description of the characteristics.</td>
</tr>
<tr>
<td>Ancillary</td>
<td>A set of elements that are not label, value or metadata.</td>
</tr>
</tbody>
</table>

Value is an element that is recorded about a characteristic of entity and label is an element that is used to represent a characteristic of entity. Metadata contain descriptions for the characteristics of entities (attributes and properties) and ancillary contains elements that are not value, label or metadata.
The following sections (4.3.2.1 to 4.3.2.5) elaborate the design of the metamodel and demonstrate how physical structure, dataset category and dataset concept are used in the metamodel to describe the structure of data sets.

4.3.2.1 Physical structure: Value

We introduced the notion of value in the metamodel; value is a common aspect that appears in all examples of datasets described in Chapter 2. As described earlier, value is an element that is recorded about a characteristic of entity. We classified value as measurement value, record identifier or ancillary value because we wanted to distinguish the different categories of value that correspond to different characteristics of entities (attributes and properties).

Measurement values correspond to the attributes of entities, and record identifiers, as well as ancillary values, correspond to properties of entities. We chose these categories of values because they are useful for describing the particular data quality issues illustrated in Chapter 5.

Before describing each category of value, we introduce the concept of record as an element, for which we identified a need in Example 1: PROMISE Boetticher; Example 2: PROMISE Datatrieve; Example Artificial 3; and Example Artificial 4. These examples show that some rows in the data sets contained values for a different entity. Therefore, we define a record as a list of values associated with a particular entity. The value for the record can be measurement value, record identifier or ancillary value.

Measurement values are values that are obtained through the process of measurement (as described earlier in Section 4.3.1). The measurement values are values for the attributes of entities that are associated with metrics. We can see that each example of a data set given in Chapter 2 contains measurement values because, by definition, a data set contains a collection of elements that include measurement values.

Record identifiers are values for the properties that represent the entities. Some data sets present the record identifiers as values that are used to identify the different entities being measured (e.g., see Example Artificial 5). These values must be unique so that they distinguish one entity from another. Other data sets present the record identifiers as values that represent the entities; however, in these cases, the entities cannot be determined. These values are used to differentiate the records (e.g., data from an anonymous survey contain values that represent identifiers for the entity). We identified the need for the record identifiers in Example 4: Qualitas Weka 3.7.5, Example Artificial 2 and Example Artificial 5.

Ancillary values are values for the properties of entities that are neither record identifiers nor measurement values. We classified values that are not relevant for assessing the quality of data sets as ancillary value. We can see the need for the ancillary value in Example 1: PROMISE Boetticher and Example 5: PROMISE Ant 1.3.
4.3. DESIGNING A DATASET METAMODEL

4.3.2.2 Physical structure: Label

We identified the need for label in the metamodel in some examples of data sets in Chapter 2 (e.g., Example 1: PROMISE Boetticher; Example 4: Qualitas Weka 3.7.5; Example 5: Ant 1.3; Example Artificial 2 and Example Artificial 5). As described earlier, label is used to represent a characteristic of entity. It is associated with a value. The labels are also used to organise the values by providing names for the characteristics of entity.

We classified labels as metric label, entity label or ancillary label because we wanted to distinguish the different characteristics of entities that are attributes and properties. These categories of label commonly appear in data sets. We chose these categories because they are useful to convey certain information about the corresponding values and because they help us describe some of the data quality issues.

The metric label is used to represent a metric that measures the attributes of entities. The metric label is associated with measurement value. We can see the need for metric label in all examples of data sets given in Chapter 2. The entity label is used to represent a property whose values distinguish the different entities. The entity label is associated with the record identifier. We identified the need for the entity label in Example 4: Qualitas Weka 3.7.5, Example Artificial 2 and Example Artificial 5. The ancillary label is used to represent a property that is neither a metric label nor an entity label and is associated with the ancillary value. We can see the need for the the ancillary label in Example 1: PROMISE Boetticher and Example 5: PROMISE Ant 1.3.

4.3.2.3 Physical structure: Metadata

One of the key aspects of data in our metamodel is metadata (for which we identified the need in Example 2: PROMISE Datatrieve; Example 3: PROMISE KC2; Example 4: Qualitas Weka 3.7.5; and Example Artificial 5). The purpose of metadata is to provide adequate, correct and relevant information about the characteristics of entities to researchers to facilitate interpretation of data. We classified metadata as metric metadata, entity metadata, data type metadata or ancillary metadata because we wanted to differentiate the different descriptions for characteristics of entities that commonly appear in data sets. We chose these categories because they assist in identifying some data quality issues.

In particular, metric metadata describe the metrics used to measure an attribute of an entity, whereas entity metadata describe the properties whose values are used to distinguish between entities. These two categories of metadata are used to determine which entities have been measured and exactly which metrics have been used. We can see the need for metric metadata and entity metadata in Example 4: Qualitas Weka 3.7.5 and Example Artificial 5.

While the entity label and the metric label represent elements that relate to the metrics and entities, metric metadata and entity metadata are described separately in the metamodel. This is because we can see from the examples of data sets given in Chapter 2 that metric label and entity label always appear in a data set; however, metric metadata and entity metadata can appear in external locations.
4.3. DESIGNING A DATASET METAMODEL

Regarding data type metadata, we can see the need for them in some data sets, particularly ARFF data sets (Example 2: PROMISE Datatrieve and Examples 3: PROMISE KC2). Data type metadata are used to determine values that are legal for the measurement values. They describe explicitly the data type of measurement values (e.g., numeric or string) to allow researchers to interpret the measurement values correctly in data sets.

In some data sets, the data type metadata describe additional information to support the interpretation of measurement values. For example, if a data set contains values for the rank of students’ performance, the data type metadata should describe the data type of the values (e.g., numeric), and the values should be rank ordered. They should also describe particular information related to the rank, for example, the criteria of rank such as ‘the same value for a rank is allowed’ or ‘the values are in an ordered sequence’. The additional information in the data type metadata is useful not only to assess the validity of measurement values but also to assist in interpreting the measurement values correctly.

Ancillary metadata contain descriptions that are not metric metadata, or entity metadata or data type metadata. We classified metadata that are not relevant for the purposes of assessing the quality of datasets as ancillary metadata. We identified the need for the ancillary metadata in Examples 3: PROMISE KC2.

4.3.2.4 Physical structure: Ancillary

We classified a set of elements that are not value, label nor metadata as ancillary. We considered that these elements were not relevant for the purposes of assessing the quality of a data set. The ancillary elements often appear to indicate the structural separation of elements in a data set (e.g., a per cent sign (%) to indicate a comment, or a comma (,) to separate an element) and to describe the context of the data set (e.g., the tool version used to generate the data file, and the date of data set creation). We can see the need for the ancillary elements in Example 2: PROMISE Datatrieve; Example 3: PROMISE KC2; Example 4: Qualitas Weka 3.7.5; Example Artificial 3; and Example Artificial 4.

In Table 4.2, we provide precise definitions for the dataset category mentioned above.

4.3.2.5 Dataset concepts

The metamodel in Figure 4.1 also illustrates five concepts that do not appear explicitly in a data set. These concepts are labelled dataset concepts. The dataset concepts are metric, entity, attribute, property and population. We adopt the definition for metric provided by the IEEE Standard Glossary of Software Engineering Terminology [90] to enable a common understanding throughout this thesis. We provide precise definitions for dataset concepts, as shown in Table 4.3.

4.3.3 Extended data sets

The elements of a data set may exist in multiple locations. For example, a data set may consist of a file containing the measurement values, while information (metadata) about those measurement values is described in external locations, such as a separate file, web
### Table 4.2: Definitions of dataset category elements

<table>
<thead>
<tr>
<th>Element category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement value</td>
<td>A value that is obtained through the process of measurement</td>
</tr>
<tr>
<td>Metric label</td>
<td>A label that represents a metric that measures an attribute of an entity</td>
</tr>
<tr>
<td>Record</td>
<td>A list of values that are associated with a particular entity</td>
</tr>
<tr>
<td>Record identifier</td>
<td>A value in a record that represents an entity</td>
</tr>
<tr>
<td>Entity label</td>
<td>A label that represents a property whose values distinguish between entities</td>
</tr>
<tr>
<td>Ancillary value</td>
<td>A value that is neither a measurement value nor a record identifier</td>
</tr>
<tr>
<td>Ancillary label</td>
<td>A label that is neither a metric label nor an entity label</td>
</tr>
<tr>
<td>Metric metadata</td>
<td>Metadata that describe the metric used to measure an attribute of an entity</td>
</tr>
<tr>
<td>Data type metadata</td>
<td>Metadata that describe explicitly the data type of a measurement value</td>
</tr>
<tr>
<td>Entity metadata</td>
<td>Metadata that describe the property whose values distinguish between entities</td>
</tr>
<tr>
<td>Ancillary metadata</td>
<td>Metadata that are not metric metadata or entity metadata or data type metadata</td>
</tr>
</tbody>
</table>

### Table 4.3: Definitions of dataset concept elements

<table>
<thead>
<tr>
<th>Element category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>‘A quantitative measure of the degree to which an item possesses a given quality attribute’ [90]</td>
</tr>
<tr>
<td>Attribute</td>
<td>A characteristic of an entity that is associated with a metric</td>
</tr>
<tr>
<td>Property</td>
<td>A characteristic of an entity that is not associated with a metric</td>
</tr>
<tr>
<td>Entity</td>
<td>A thing that can be measured by its characteristics</td>
</tr>
<tr>
<td>Population</td>
<td>A group of entities that describe the same characteristics of the entity</td>
</tr>
</tbody>
</table>

Page or publication. Some of the external locations can be seen in the examples of real data sets that are described in Section 2.2.1.

Data sets that contain elements described in external locations may raise a risk to anyone using them of which they will not be aware. For example, if the metadata to interpret the measurement values are described in the external files, then any update that is made to the files may not be reflected in the actual data set. This could also happen if the metadata are described in web pages. There is, therefore, an issue with the synchronisation of data that involves more than one location.

If the metadata are described in publications, researchers may not be aware that the metadata exist because they would need to spend a lot of effort reviewing the contents of the publication to find the relevant metadata that describe the measurement values in the data set. Moreover, sometimes, the metadata for measurement values are not explicitly described in the publication making it difficult to interpret the data set correctly, particularly the measurement values.

In this thesis, we introduce the notion of an extended data set to indicate all of the
n necessary elements of a data set required to interpret the measurement values in all locations (a single file and external locations). The reason for an extended data set is that we found in some of the real data sets in Chapter 2 that metadata to interpret the measurement values was contained not only in the single file, but also in external locations. Hence, we considered the measurement data set as a subset of the extended data set (see Figure 4.2) because the measurement data set contains all of the necessary elements to interpret the measurement values.

Figure 4.2: An overview of an extended data set, measurement data set and measurement metadata.

We present an overview of the extended data set, measurement data set and measurement metadata in Figure 4.2. There are five regions: region E represents an extended data set that contains both the measurement data set and the measurement metadata; region A and region C represent a measurement data set; within region A, region D represents the measurement values; region C represents the measurement metadata and region B represents the measurement metadata that exist in external locations. If region C is empty, then there are no measurement metadata in the measurement data set. If region B and C are empty, then there are no measurement metadata in the extended data set.

As shown in Figure 4.2, measurement metadata are represented by region B and region C. Both regions contain entity metadata and metric metadata, but not the entity label or the metric label. As mentioned earlier, although the entity label and the metric label represent elements that relate to the metrics and the entities, we separate them from the measurement metadata because we describe them separately in our metamodel.
For that reason, the entity label and the metric label exist only in region A.

Ideally, the important regions shown in Figure 4.2 are region A and region C, the measurement data set that contains all of the elements necessary to correctly interpret the data set. In particular, this means that the measurement data set contains sufficient metadata to uniquely identify which entity is being measured and which metric is being used. However, if the information explaining the measurement values is in the extended data set (e.g., region B), then there is a potential for quality issues. If researchers only received the measurement data set, they may overlook the information in the extended data sets. Therefore, in order to encounter the fewest potential quality issues, the information should be placed in the measurement data set.

While we believe the ideal way to describe the measurement metadata is in the measurement data set, majority of data sets from public repositories describe some of the necessary metadata in external locations. In particular, the measurement metadata is commonly described at the location that we go to access the data set file, such as a web page. The web pages of some data sets contain the necessary metadata to facilitate the interpretation of data sets. In this case, our data quality framework will take into consideration the extended data set in evaluating the quality of data sets.

4.4 Modelling a data set: The process

The process of modelling a data set involved creating a model of data sets by applying the dataset metamodel using a formal procedure. This process was developed as an approach to understand what a data set is intended to represent. In addition, it helps to identify what information may be missing in a data set.

Figure 4.3 provides a diagram for modelling data sets in the data quality framework. Modelling a data set is the first process in our data quality framework and requires a data set as a primary input. If the metadata of the data set is described on a web page, we consider the web page to be a secondary input.

We used the dataset metamodel to construct a formal procedure to model a data set. The formal procedure is broken into two steps: (1) identification: to identify the physical structure elements; (2) interpretation: to classify the physical structure elements into the dataset category elements. The output from the modelling of the data set process is the model of the data set.

4.4.1 Identifying the physical structure elements

The identification step is intended to identify the physical structure elements in the data set. As described in Section 4.3, the physical structure elements are label, value, metadata and ancillary. The formal procedure to identify these elements is illustrated in Table 4.4.

Before performing the formal procedure, we prepared the necessary content of the data set for the modelling in Excel spreadsheets. First, we extracted the entire contents of the data set from the data set file into the Dataset sheet. Then, if the metadata existed on the web page of the data set, we extracted the metadata from the web page of the data
set into the WebpageDataset sheet. Finally, we performed the formal procedure outlined in Table 4.4.

Table 4.4: Procedure for identifying the physical structure elements.

<table>
<thead>
<tr>
<th>No.</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Identify label. Colour the label blue.</td>
</tr>
<tr>
<td>2.</td>
<td>Identify values. Colour the values light green.</td>
</tr>
<tr>
<td>3.</td>
<td>Identify elements that represent metadata. Colour the metadata elements dark grey.</td>
</tr>
<tr>
<td>4.</td>
<td>Identify elements that represent ancillary. Colour the ancillary elements purple.</td>
</tr>
<tr>
<td>5.</td>
<td>Check that all elements in the Dataset sheet have been coloured. Then, look in the WebpageDataset sheet. Identify elements that represent metadata. Colour the metadata elements dark grey.</td>
</tr>
</tbody>
</table>

In the formal procedure, we used colour codes to specify the individual elements of the data set according to the physical structure elements defined in the dataset metamodel. Each physical structure element was assigned a specific colour. Figure 4.4 illustrates an example of the results for the identification step in the process of modelling the data set. As can be seen in Figure 4.4, every element in the data set has been coloured according to the legend for Dataset Elements.

In this thesis, our definitions of the elements of data sets are focused on a single file; however, these definitions can be generalised to include the elements in external locations (as described in Section 4.3). With regard to the external locations, we took into account the metadata that existed on the web page of data sets in our data quality framework. In the process of modelling the data set, we considered the web page of the data set to be the secondary input. If the metadata existed on the web page of the data set, we modelled this metadata in the same way as we modelled the metadata in the
4.4. MODELLING A DATA SET: THE PROCESS

Figure 4.4: An example of the results of the identification of the physical structure elements.

single file. This method allows researchers to apply the same interpretation to metadata that are located in a single file or on the web page of the data set.

4.4.2 Interpreting the physical structure elements

The interpretation step is to classify the physical structure elements into the dataset category elements of the data set. The dataset category elements are record, ancillary value, record identifier, measurement value, metric label, entity label, ancillary label, metric metadata, entity metadata, data type metadata and ancillary metadata. The formal procedure to classify the physical structure elements is illustrated in Table 4.5.

Similarly to the procedure for the identification step, we used colour codes in the interpretation step procedure to classify each physical structure element of the data set into the dataset category elements. We eliminated the ancillary elements in this step because these elements are not relevant to data quality assessment. We modelled all the dataset category elements, including records.

When modelling the records, we needed to ensure that each record identifier is uniquely represented the entity being measured. If two or more records have the same measurement values associated with the same metric label for the same record identifier, there would have been a potential for quality issues in the data set. In this case, to indicate the quality issues, we needed to model those records that had the same record identifier in a different way. It was important to consider this case carefully at this point to avoid the unlikely situation that all the values would be part of the same record. We describe the procedure to consider this case in No. 11 of the interpretation steps in Table 4.5.

We illustrate an example of the results for the second step in Figure 4.5. This example also represents the output of the data set modelling process, which is a model of the data set.
Table 4.5: Procedure for interpreting the physical structure elements.

<table>
<thead>
<tr>
<th>No.</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent <em>measurement values</em>. Colour the <em>measurement values</em> green.</td>
</tr>
<tr>
<td>2.</td>
<td>In the <strong>Dataset</strong> sheet, look for labels (already coloured blue) that correspond to <em>metric labels</em>. Check whether the labels are associated with the <em>measurement values</em>. If yes, they are <em>metric labels</em>. If no, check whether there are metadata (already coloured with dark grey) that describe a metric and are associated with the labels. If yes, the labels are <em>metric labels</em>. Colour the <em>metric labels</em> red.</td>
</tr>
<tr>
<td>3.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent <em>metric metadata</em>. Colour the <em>metric metadata</em> dark blue.</td>
</tr>
<tr>
<td>4.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent <em>record identifiers</em>. Colour the <em>record identifiers</em> light blue.</td>
</tr>
<tr>
<td>5.</td>
<td>In the <strong>Dataset</strong> sheet, look for labels (already coloured blue) that represent <em>entity labels</em>. Check whether the labels are associated with the <em>record identifiers</em>. If yes, they are <em>entity labels</em>. If no, check whether there are metadata (already coloured dark grey) that describe an entity and are associated with the labels. If yes, the labels are <em>entity labels</em>. Colour the <em>entity labels</em> aqua.</td>
</tr>
<tr>
<td>6.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent <em>entity metadata</em>. Colour the <em>entity metadata</em> pink.</td>
</tr>
<tr>
<td>7.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent <em>ancillary values</em>. Colour the <em>ancillary values</em> blue.</td>
</tr>
<tr>
<td>8.</td>
<td>In the <strong>Dataset</strong> sheet, look for labels (already coloured blue) that represent <em>ancillary labels</em>. Check whether the labels are associated with the <em>ancillary values</em>. If yes, they are <em>ancillary labels</em>. If no, check whether there are metadata (already coloured dark grey) that describe a property that is not a metric label or entity label and are associated with the label. If yes, the labels are <em>ancillary labels</em>. Colour the <em>ancillary labels</em> purple.</td>
</tr>
<tr>
<td>9.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent <em>ancillary metadata</em>. Colour the <em>ancillary metadata</em> light grey.</td>
</tr>
<tr>
<td>10.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metric metadata (already coloured dark blue) that represent <em>data type metadata</em>. If the <em>data type metadata</em> is present, draw a brown rectangle to highlight the <em>data type metadata</em>. Next, look for metadata (already coloured dark grey) that represent <em>data type metadata</em>. Draw a brown rectangle to highlight the <em>data type metadata</em>.</td>
</tr>
<tr>
<td>11.</td>
<td>In the <strong>Dataset</strong> sheet, identify records. Draw an orange rectangle to highlight the records. In each record, look for record identifier (already coloured light blue). If the <em>record identifier</em> is present, ensure each record has a unique <em>record identifier</em>. If there are two or more records with the same record identifiers, then these records potentially have quality issues. Draw a red rectangle to highlight these records.</td>
</tr>
</tbody>
</table>

In Figure 4.5, we can see that all of the elements in the data set were modelled according to the colour codes in the legend for *Dataset Elements*. This example contains typical tabular data and population data. We modelled the population data using the same procedure for which we modelled the tabular data. For example, we modelled
4.5 MODELLING THE EXISTING DATA SETS

Files and Modules as metric labels and the values associated with these metric labels as measurement values. In this example, we modelled the measurement values (183, 3) as a record for the population data because they represent a list of values for a particular entity. The record for the population data is indicated by a broken orange rectangle in Figure 4.5.

4.5 Modelling the existing data sets

We evaluate the use of our dataset metamodel by applying it to existing data sets from the PROMISE repository. In the initial stage, the PROMISE repository contains 17 data sets, which are from the NASA Metrics Data Program. This repository has since expanded to include 96 data sets (last accessed in May, 2014). In this 2014 version of the PROMISE repository, these data sets are grouped into five categories: defect prediction, effort prediction, text mining, model-based software engineering and general.

We selected 87 of 96 data sets from PROMISE that had similar structures to the examples of the data sets used in Chapter 2. In total, we applied our dataset metamodel to 92 data sets, including the examples of the data sets given in Chapter 2. The remaining (5 of 96) data sets from PROMISE are not in an appropriate format to be included in our study. These data sets are Modis, Spe, Quantitative, Abacus2013 and Generics. Some of these data sets provide unfamiliar format of files that difficult to understand and others require a specific kind of software to be able to read the data sets. Therefore, we excluded these data sets in this evaluation.

As mentioned earlier in the previous section, the process of modelling data sets starts by identifying the physical structure elements. The next step is to classify the physical structure elements into the dataset category elements based on the metadata provided in the data set. Below, we present in more precise detail the results of the dataset modelling process for five data sets from Chapter 2.
4.5. MODELLING THE EXISTING DATA SETS

4.5.1 Example 1: PROMISE Boetticher

Before modelling this data set, we decided to model the column headers and the values associated with the column headers as elements. We italicised the column headers in Figure 4.6 to clearly indicate that we modelled the column headers as labels. We also modelled the elements associated with the column headers as values.

Based on our assumption in Section 2.2.1, we then modelled some of the labels as **metric labels** because they were associated with values that are integers and symbols, and look like **measurement values**. For example, we modelled the labels *Comp Sc Undergrad Courses*, *Gender*, *Highest Degree* as **metric label** because these labels was associated with values that we assumed to be **measurement values**.

As mentioned in Chapter 2, values in this data set were organised into rows that contain information about the entities. We modelled every row as a **record** because we assumed that every row contained a list of values associated with a particular entity. This assumption allowed us to apply the definition of a record given in Section 4.3.2. However, we had no information to confirm that every record was a different entity. As it was not clear whether or not every row contained values for a different entity, there was a risk of misinterpretation. If there had been metadata to show that every row contained values for a different entity, this may reduced the risk of misinterpretation.

Although this data set had a potential risk of misinterpretation, we were able to model the data set using the data modelling procedure with some assumptions based on the common conventions we have seen in many other data sets. We illustrate part of the model of the PROMISE Boetticher data set as the results from the data set modelling process in Figure 4.7.

![Figure 4.6: PROMISE Boetticher data set.](image)

4.5.2 Example 2: PROMISE Datatrieve

In this data set, we determined the elements by applying the definition of an element given in Section 4.3.1. In Figure 4.8, we have highlighted in bold the elements that we assumed to be labels after the heading @attribute. This is because these elements...
4.5. MODELLING THE EXISTING DATA SETS

looked similar to the column header in the tabular data. In the second step of the data set modelling process, we modelled every label as a *metric label* because we assumed the values after the heading *@data are measurement values*. We modelled the elements (e.g., numeric) after the *metric labels as data type metadata* because we assumed these elements describe the data type for the measurement values.

In Chapter 2, we described some of the metadata of this data set located on the data set’s web page (as shown in Figure 4.9). We assumed that the attribute information on the web page of the data set presented the metadata for the labels in the data set because the description contained similar labels to the data set. For example, the element *LOC60* in the attribute information on the web page appear to be similar with the label *LOC6_0* in the data set. In this case, we modelled the metadata (attribute information) as *metric*
4.5. MODELLING THE EXISTING DATA SETS

metadata because they describe the metric used to measure the entity in the data set.

Figure 4.9: PROMISE Datatrieve metadata on the web page of the data set.

As described in Chapter 2, there are elements that provide the structural separation for the data set. We have highlighted these elements (@relation, @attribute, @data) in italics in Figure 4.8. We modelled these elements as ancillary in the first step of the data set modelling process because they represent elements that are not label, value or metadata and are not relevant for assessing the quality of the data set. In the second step of the data set modelling process, we eliminated the ancillary elements because we focused on modelling the elements for the dataset category.

We modelled every row after the heading @data as a record because we assumed that each row represented a list of values for a particular entity. This illustrates that, although the values of the data set were organised in a different structure to that of Example 1, we were able to model the record by applying the definition of record given in Section 4.3.2. We illustrate part of the model of the PROMISE Datatrieve data set in Figure 4.10 and the results of the modelling of the metadata in Figure 4.11.

4.5.3 Example 3 : PROMISE KC2

The structure of the KC2 data set is the same as that of the previous example, and it allowed us to model all the same elements in a similar way. In this data set, we wanted to focus on the use of metadata, record and ancillary.

Just as in Example 2, we determined the elements by applying the definition of element given in Section 5.3.1 (note that we determined the elements in the 10th line in this data set in Section 5.3.1). We modelled the elements that were used to provide the structure of the data set (e.g., %, 1.) as ancillary elements because these elements were not label, value or metadata; nor were they relevant for assessing the quality of the data set. We also modelled the elements that described the data set creation information (e.g., the time the dataset file was created, and the author of the dataset file) as ancillary elements because they described information that is not relevant for the purpose of data quality assessment.

We noticed that the population data (e.g., number of instances) represented the properties of the entities corresponding to the measurements. Therefore, we modelled the population data in the same way as the tabular data —by applying the definition of population in Section 5.3.2 because we considered a population to be another kind
4.5. MODELLING THE EXISTING DATA SETS

Figure 4.10: PROMISE Datatrieve data set.

Figure 4.11: PROMISE Datatrieve metadata on the web page of the data set

of entity. For example, the population data contain a label (Number of instances) and a value (522). We modelled the Number of instances as the metric label and the associated value (522) as the measurement value.

For the information about the properties of entities (e.g., attribute information), we modelled the description of the properties of entities as metric metadata because we assumed that the description represented the meaning for each metric label in the data set. We also modelled some of the elements (e.g., %, 1.) in the information about the properties of entities as ancillary elements because they provided the structure of the data set.

We observed that after the heading @data, every row contained a list of values separated by commas. We assumed that these values corresponded to different properties of entities that were organised in rows after the @attribute. In a similar way to Example 2, we modelled every row as a record because it contained a list of values for a different
4.5. MODELLING THE EXISTING DATA SETS

While modelling the records, we noticed there were two rows (the lines start with (9, 2, 0, 1...)) that contained identical data. If an ID or a record identifier is given for every row, it will be obvious that every record is different and indicates the value of a different entity. In this particular case, we assumed that these rows belonged to different entities and we modelled them as different records. However, they may actually have been duplicate records. This highlights the value of a record identifier because it is able to confirm whether or not these rows belong to different entities.

We illustrate a part of the model of the KC2 data set in Figure 4.13. We eliminated ancillary elements in this model because we wanted to focus on the elements for the dataset category in the dataset metamodel. These elements are metric label, measurement values, metric metadata, data type metadata and records.

4.5.4 Example 4 : Qualitas Weka 3.7.5

The structure of the Weka data set was different from the other examples, but the elements modelled in the data set were the same as in the previous examples. In this example, we focused on the use of the entity metadata, entity label, record identifier,
4.5. MODELLING THE EXISTING DATA SETS

As mentioned in Chapter 2, this data set contained two groups of data for different kinds of entities. The two groups of data were clone pairs and clusters. We modelled the two groups of data as two populations by applying our definition of population from Section 4.3.2.

In the clone pairs population, each row represented a clone pair that contained values for a particular entity. We assumed that each clone pair contained a pair of values that is regarded as a set of values that uniquely represents the entity. The set of values consisted of values associated with Method1 and values associated with Method2. We modelled the set of values as a record identifier because it illustrated unique values that represented the entities. In addition, we modelled Method1 and Method2 as entity labels because the description for Method1 and Method2 described the property of an entity whose values distinguish between entities. For that reason, we modelled the descriptions for Method1 and Method2 as entity labels.
and Method2 as the entity metadata.

In the same clone pairs population, we assumed that File1 and File2 contained values for properties that were neither measurement values nor record identifiers. This is because the descriptions for File1 and File2 provided metadata relating to the data set, but they were neither metric metadata nor entity metadata. Hence, we modelled the values associated with File1 and File2 as ancillary values, File1 and File2 as ancillary labels and the descriptions for File1 and File2 as ancillary metadata.

In Figure 4.14, we have highlighted in italics the elements below the heading Global values to illustrate the population data in the data set. We modelled this population data in the same way as the typical tabular data. For example, we modelled the label (Files) as metric label, the value (1006) as measurement value and the description (Number of files analysed) as metric metadata. We illustrate a part of the model of the Weka 3.7.5 data set in Figure 4.15.

4.5.5 Example 5: PROMISE Ant 1.3

The structure of the Ant 1.3 data set is similar to that of Example 1, which had column headers and values, except that every column header and value was separated by a comma (,). We determined column headers, values and commas (,) as elements. We have highlighted in bold the column headers in Figure 4.16, and we modelled them as labels. In the second step of the data set modelling process, we modelled some of the labels as metric labels and the associated values as measurement values, just as in Example 1,
and we focused on the use of \textit{ancillary value} and \textit{ancillary label} because this data set contained some values that were neither measurements nor identifiers for entities.

Based on our assumption in Chapter 2, some of the labels contained values that were neither measurements nor identifiers for different entities. We modelled these values as \textit{ancillary values} because they were neither \textit{measurement values} nor \textit{record identifiers}. For example, it is clear that the values for the label \textit{name} (the first column header) represent values that are neither entity labels nor metric labels. Therefore, we modelled the values as \textit{ancillary values} and the associated label as \textit{ancillary label}.

While modelling the data set, we noticed that there were two labels that were the same, but contained two different values. The labels in question are \textit{name} in the first column header and the third column header, and the two associated values represent \textit{name of dataset} and \textit{name of source code file}. Based on the two different values, we assumed that the labels represented two different properties of entities. As mentioned...
above, we modelled the label name in the first column as ancillary label because it was associated with values that are neither measurement values nor record identifiers. For the label name in the third column, we modelled this as entity label because it was associated with the values that we assumed to be record identifiers. This indicates that the process of modelling the data set was able to solve the problem of the two labels that were the same, but contained two different values. We illustrate a part of the model of the PROMISE Ant 1.3 data set in Figure 4.17.

4.5.6 Summary of results of modelling the data sets

We present a matrix for the results of modelling the data sets in Figure 4.18 (See Appendix D). The columns of the matrix consist of four physical structure elements and 11 dataset category elements of the metamodel. We excluded the dataset concept elements in the matrix because they represent abstract concepts that do not appear explicitly in a data set. The rows contain 92 data sets, including the examples of real data sets in Chapter 2. We group a number of data sets that have the same structure and column headers in a single row; for example, the Ant 1.3 data set has the same structure and column headers as 32 data sets. We use the symbol (√) to show the elements that are present and the symbol (–) to show those that are not present in the data sets.

For metadata, we use x/y that indicates x as the number of elements (metric metadata, entity metadata, data type metadata and ancillary metadata) and y as the number of labels (metric label, entity label, ancillary label) that are present in the data set. For example, PROMISE AM1 contains 4 metric metadata, 12 data type metadata, 1 entity metadata and 13 ancillary metadata for 18 labels in the data set.

The results of modelling the data sets show the contents of the data sets. From the data in Figure 4.18, it is apparent that every data set contained values and ancillary for the physical structure elements. In particular, five of the 92 data sets (PROMISE NrpClassic(5)) contained no other physical structure elements except values and ancillary. An additional 43 data sets contained labels without metadata and the remaining data sets (44) contained labels with metadata.

In terms of dataset category elements, every data set contained measurement values and records. Of the 92 data sets, 87 contained metric labels, and 57 contained entity labels and record identifiers. Regarding metadata in particular, 18 of the 92 data sets contained metric metadata, 8 data sets contained entity metadata, 43 data sets contained data type metadata and 14 data sets contained ancillary metadata.
4.5. MODELLING THE EXISTING DATA SETS

In Figure 4.18, we can see that five data sets (PROMISE AM1, Qualitas Weka 3.7.5, PROMISE nasara93, PROMISE Cosmic and PROMISE ISBSG 10) contained all the physical structure and dataset category elements. This indicates that some existing data sets contained all the elements for physical structure and dataset category from our metamodel.

The results of modelling the data sets reveal where the data sets could be improved by indicating the dataset category elements that are not present in the model of the data sets.
In particular, with regard to metadata, many data sets did not contain metric metadata (74 data sets), entity metadata (84 data sets) and data type metadata (49 data sets). These categories of metadata are essential to support the correct interpretation of data. This indicates that the authors of data sets should provide these metadata to reduce the potential risk of misinterpretation in their data sets.

4.5.7 Discussion of the results of modelling the data sets

We found that most data sets contained values and labels for the physical structure of the data. It seems possible that this was because of the nature of the real data sets, which typically have measurements with multiple columns. Moreover, the column headers used to distinguish these columns were identified as labels, mainly for metrics and entities. In addition, although the metadata were not available for these data sets, this does not mean that all the values could be misinterpreted. In some cases, it could be that the labels were sufficient to uniquely identify the metrics and the entities.

However, we noticed that due to the absence of some dataset category elements in the data sets, the process of modelling the data set may have risked misinterpretation. For example, PROMISE KC2 indicated that without a record identifier in every record, we would be unable to distinguish between duplicate data or those entities that might have the same data. In this particular case, we found a potential risk of misinterpretation by modelling the data set.

The overall result indicates that very few data sets in the PROMISE repository actually contained metadata to describe what the labels mean. In particular, we mentioned that the Ant 1.3 data set example in Chapter 1 illustrated an issue with the interpretation of the data because no information was available to explain what the label LOC means in the data set. In Chapter 4, we modelled the Ant 1.3 data set using our metamodel. We found that the Ant 1.3 data set did not contain metric metadata and entity metadata. Consequently, there is a potential for other researchers using the same kinds of data sets from the PROMISE repository to misinterpret the meaning of the measurement.

4.6 Conclusions

This chapter has detailed the metamodel for describing the structure of a data set. The metamodel provides a standard terminology for describing a data set and classifies the elements of the data set into three levels: physical structure, dataset category and dataset concept. Further, the metamodel describes the relationship between each level.

The practical use of this newly designed metamodel in modelling data sets has been illustrated by applying it to the existing data sets. The process of modelling the data sets has been demonstrated by applying the metamodel using formal procedures. It was found that the process may raise the potential risk of misinterpretation because of the absence of dataset category elements in data sets.

The metamodel was designed because we found that there is no clear existing definition of data quality issues in the literature described in Chapter 3. We have attempted to
construct a formal definition of data quality issues that has a clear and precise meaning (see Chapter 5). We used the standard terminology from the metamodel to specify the quality issues in the formal definition. However, the usefulness of the formal definition of data quality issues can only be judged by evaluating it with a study (see Chapter 7).

The metamodel forms the foundation for the framework to evaluate the quality of data sets. Further discussion on the framework for data quality follows in Chapter 5, which concentrates on the development of a quality assessment process. Also presented in Chapter 5 are some examples of how the formal definition of data quality issues can be used to identify the quality issues in a data set.
This chapter presents the second part of a framework for data quality assessment: the quality assessment process. The quality assessment process introduces formal definitions of data quality issues (to identify the quality issues) and an evaluation scale (to evaluate the quality of metadata). The chapter also discusses the use of this framework.

5.1 Introduction

In Chapter 4, we discussed the part of the data quality assessment framework that pertains to the dataset metamodel and the process for modelling the data set. Chapter 5 presents the remaining part of the data quality assessment framework—the quality assessment process. The aim of the data quality assessment framework is to determine whether a data set contains sufficient information to facilitate the correct interpretation of data for analysis in empirical research. The framework may identify when there is uncertainty as to which entities have been measured and which metrics have been used. The framework may also help researchers to identify not only common data quality issues (e.g., missing and duplicate data), but also quality issues related to the interpretation of data.

In Figure 5.1, we present the data quality framework to highlight the two processes: the modelling of the data set and quality assessment. The first process (described in detail in Chapter 4) aims to model data sets using a formal procedure based on the dataset metamodel. The second process, which is discussed in this chapter, aims to evaluate the quality of data sets. There are four steps in the quality assessment process: (1) assess the model of the data set, (2) identify data quality issues, (3) evaluate the metadata of the data set, and (4) prepare an assessment report.

The framework contains two components in the knowledge platform: the dataset metamodel and formal definitions of data quality issues. As described earlier, the dataset metamodel is used in the formal procedure for modelling the data sets described in
Chapter 4. In section 5.2, we introduce the formal definitions of data quality issues to clearly describe data quality issues in data sets for the quality assessment process.

The remainder of this chapter is structured as follows. Section 5.3 presents the quality assessment process and describes in detail each step of the quality assessment process. Section 5.4 describes the results of the quality assessment process. Section 5.5 shows the application of the quality assessment process to existing data sets from the public data repositories to demonstrate its use.

5.2 Formal definition of data quality issues

In Chapter 3, we discussed the need for a standard way to describe data quality issues in data sets. In this chapter, we present a formal definition that clearly and rigorously describe each quality issue. Besides rigour, the formal definition is precise and consistent because we use the standard terminology from the dataset metamodel, allowing researchers to apply a common interpretation of the data set to identify the quality issues.

We use the seven dataset category elements from the dataset metamodel (see Chapter 4) to construct the formal definitions of data quality issues (in this section) and the formal procedures to identify the quality issues (in section 5.3). These seven dataset category elements are important for identifying the structure of a data set that may have quality issues. For convenience, we repeat the definitions of the dataset category elements:

(i) Measurement value: A value that is obtained through the process of measurement.

(ii) Metric label: A label that represents a metric that measures an attribute of an entity.

(iii) Record: A list of values that are associated with a particular entity.

(iv) Record identifier: A value in a record that represents an entity.
5.2. FORMAL DEFINITION OF DATA QUALITY ISSUES

(v) Entity label: A label that represents a property whose values distinguish between entities.

(vi) Metric metadata: Metadata that describe the metric used to measure an attribute of an entity.

(vii) Entity metadata: Metadata that describe the property whose values distinguish between entities.

Apart from the need for the formal definitions of data quality issues, we need to use a clear and consistent terminology to indicate quality issues in data sets. We consider the terminology that is used to indicate some common quality issues reported in our systematic mapping study (discussed in Chapter 3) and some quality issues presented in a targeted review on data quality [11]. We selected four terms that indicate the common quality issues, which are duplicate data, inconsistent data, missing data and incorrect data. These four terms are in line with the proposed quality issues from the accuracy category (redundancy, inconsistency, incompleteness and noise) in the taxonomy of data quality presented by Bosu and MacDonell [61].

We identified the need for quality issues related to interpretation of data from the observation of data sets (described in Chapter 2), which showed that the existing data sets (from public data repositories) contain insufficient metadata. In particular, the data sets contain insufficient metadata for the metrics and entities that are required to interpret the measurements. Because of this, we considered four metadata-related quality issues with metrics and entities in our quality assessment process. The four metadata-related quality issues are incomplete metadata for a metric, incomplete metadata for an entity, imprecise metadata for a metric and imprecise metadata for an entity.

We present the formal definitions for common data quality issues and metadata-related quality issues in the following subsections.

5.2.1 Common data quality issues

In this section, we construct formal definitions for four common quality issues: duplicate data, inconsistent data, missing data and incorrect data. Following the formal definition, each data quality issue is illustrated with an example. We reproduce artificial data set 5 from Chapter 2 in Figures 5.2 and 5.3. This example was purposely created to illustrate the four quality issues we are considering (described in Chapter 2).

First, we present the formal definitions for duplicate data and inconsistent data, and illustrate with examples (see Figure 5.2). We modelled the element Filename as an entity label and elements NOF, NPM, NOC, FV, nCom and nBug as metric labels. The definitions are as follows:

(i) Duplicate data

a) Definition: Two or more records that have the same measurement values associated with the same metric for the same entity.
b) Example: There are two identical records containing the entity label File-name \`C:\DataGroup\ModuleA\gamesPanel.cpp\` and metric label NPM 2110. This illustrates duplicate data because they have the same measurement values (2110) associated with the same metric (NPM) for the same entity (\`C:\DataGroup\ModuleA\gamesPanel.cpp\`). The two identical records are highlighted in blue in Figure 5.2.

(ii) Inconsistent data

a) Definition: Two or more records that have different measurement values associated with the same metric for the same entity.

b) Example: There are two records containing the same entity label Filename \`C:\DataGroup\ModuleA\board.cpp\` with two different values (2115 and 1220) for the same metric label (NPM). These records (highlighted in orange in Figure 5.2) illustrate inconsistent data because they have different measurement values associated with the same metric (NPM) for the same entity (\`C:\DataGroup\ModuleA\board.cpp\`).

Second, we present the formal definitions for missing data and incorrect data and illustrate with examples in Figure 5.3. Similarly to Figure 5.2, we modelled the element Filename as an entity label and the other elements as metric labels. The definitions are as follows:

(i) Missing data

a) Definition: A record that does not have a measurement value for a given metric.
5.2. FORMAL DEFINITION OF DATA QUALITY ISSUES

b) Example: There is a record with an empty string or null value for metric label NOC. In this data set, it is not valid to have a null value for a metric. This illustrates missing data because there is no measurement value associated with a metric label (NOC). This record is highlighted in yellow in Figure 5.3.

(ii) Incorrect data

a) Definition: A record that has an invalid measurement value for a given metric.

b) Example: There is a record that contains value k for metric label NPM (highlighted in green in Figure 5.3). The metric metadata for the metric label NPM describes the number of public methods, indicating that any value for metric label NPM should be an integer — a measurement value k is invalid for this metric label (NPM) and hence this represents incorrect data.

Figure 5.3: An illustration of missing and incorrect data, based on artificial data set 5

<table>
<thead>
<tr>
<th>Filename</th>
<th>PHase</th>
<th>NOF</th>
<th>NPM</th>
<th>NOC</th>
<th>nFv</th>
<th>nCam</th>
<th>nBug</th>
</tr>
</thead>
<tbody>
<tr>
<td>./DataGroup/ModuleA/Game.c.cpp</td>
<td>1</td>
<td>45</td>
<td>2003</td>
<td>89</td>
<td>66</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>./DataGroup/ModuleA/GamePanel.cpp</td>
<td>3</td>
<td>85</td>
<td>2110</td>
<td>100</td>
<td>45</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/panelSync.cpp</td>
<td>3</td>
<td>87</td>
<td>2227</td>
<td>15</td>
<td>16</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/panelBreak.cpp</td>
<td>3</td>
<td>85</td>
<td>2110</td>
<td>100</td>
<td>45</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionFacade.cpp</td>
<td>2</td>
<td>85</td>
<td>1551</td>
<td>10</td>
<td>15</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionSequence.cpp</td>
<td>1</td>
<td>78</td>
<td>1385</td>
<td>10</td>
<td>14</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionFeedback.cpp</td>
<td>1</td>
<td>12</td>
<td>1201</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/HeadShell.cpp</td>
<td>2</td>
<td>85</td>
<td>1011</td>
<td>22</td>
<td>36</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionFacade.cpp</td>
<td>1</td>
<td>12</td>
<td>1225</td>
<td>55</td>
<td>85</td>
<td>53</td>
<td>5</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionSync.cpp</td>
<td>1</td>
<td>42</td>
<td>1420</td>
<td>81</td>
<td>32</td>
<td>81</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionInterface.cpp</td>
<td>2</td>
<td>32</td>
<td>1325</td>
<td>85</td>
<td>42</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionContent.cpp</td>
<td>2</td>
<td>15</td>
<td>1521</td>
<td>12</td>
<td>32</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/ActionFacade.cpp</td>
<td>2</td>
<td>89</td>
<td>1144</td>
<td>42</td>
<td>25</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/HeadShell.cpp</td>
<td>1</td>
<td>87</td>
<td>1220</td>
<td>88</td>
<td>89</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/panelBreak.cpp</td>
<td>3</td>
<td>85</td>
<td>k</td>
<td>32</td>
<td>22</td>
<td>85</td>
<td>5</td>
</tr>
<tr>
<td>./DataGroup/ModuleA/panelContent.cpp</td>
<td>3</td>
<td>85</td>
<td>2115</td>
<td>100</td>
<td>44</td>
<td>22</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 5.3: An illustration of missing and incorrect data, based on artificial data set 5

5.2.2 Metadata-related quality issues

Metadata-related quality issues are quality issues that relate to interpretation of data. We considered these quality issues to indicate whether or not a data set contains sufficient metadata for interpretation. This is essential to assist researchers in determining data sets that are appropriate for analysis in research.

To apply a common interpretation of data, we used the standard terminology from the dataset metamodel (e.g., entity metadata, metric metadata, metric label) in the formal definitions for these metadata-related quality issues. We formally defined each of these metadata-related quality issues: incomplete metadata for a metric, incomplete metadata for an entity, imprecise metadata for a metric and imprecise metadata for an entity.
5.2. FORMAL DEFINITION OF DATA QUALITY ISSUES

We reproduce artificial data set 5 from Chapter 2 in Figure 5.4 to present the formal definitions for incomplete metadata for a metric and incomplete metadata for an entity. We modelled the element Filename as an entity label and elements NOF, NPM, NOC, FV, nCom and nBug as metric labels in Figure 5.4. The definitions are as follows:

(i) Incomplete metadata for a metric.

a) Definition: A metric label that does not have metric metadata.

b) Example: The metric label nBug does not have a description for metric metadata. This label is highlighted in blue in Figure 5.4.

(ii) Incomplete metadata for an entity.

a) Definition: An entity label that does not have entity metadata.

b) Example: The entity label Filename does not have a description for entity metadata. This label is highlighted in purple in Figure 5.4.

To provide examples of data quality issues for imprecise metadata for metrics and entities, we created a new artificial data set (Figure 5.5). This example contained source-code data for a software project. The metadata were described in the data set. We modelled the elements associated with ModNm as record identifiers because they represent entities. We assumed the element associated with the record identifiers was an entity label. We assumed all the elements that looked like column headers, except ModNm, were metric labels, because the values associated with the column headers looked like measurement values.

Formal definitions for imprecise metadata for metrics and for entities, along with examples (illustrated in purple in Figure 5.5), are as follows:

<table>
<thead>
<tr>
<th>File</th>
<th>Phase</th>
<th>NOF</th>
<th>NPM</th>
<th>NOC</th>
<th>FV</th>
<th>nCom</th>
<th>nBug</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataGroup/ModulA/games.cpp</td>
<td>1</td>
<td>45</td>
<td>2001</td>
<td>89</td>
<td>66</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>DataGroup/ModulA/gamesPanel.cpp</td>
<td>3</td>
<td>85</td>
<td>2110</td>
<td>100</td>
<td>45</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelSync.cpp</td>
<td>3</td>
<td>87</td>
<td>2227</td>
<td>15</td>
<td>55</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/gametPanel.cpp</td>
<td>3</td>
<td>85</td>
<td>2110</td>
<td>100</td>
<td>45</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelFacade.cpp</td>
<td>2</td>
<td>85</td>
<td>1551</td>
<td>10</td>
<td>55</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionSequence.cpp</td>
<td>1</td>
<td>78</td>
<td>1285</td>
<td>10</td>
<td>54</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionShell.cpp</td>
<td>1</td>
<td>36</td>
<td>1201</td>
<td>15</td>
<td>55</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionInterface.cpp</td>
<td>2</td>
<td>85</td>
<td>1011</td>
<td>22</td>
<td>36</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionSet.cpp</td>
<td>1</td>
<td>12</td>
<td>1225</td>
<td>55</td>
<td>85</td>
<td>55</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionFacade.cpp</td>
<td>1</td>
<td>42</td>
<td>1420</td>
<td>81</td>
<td>12</td>
<td>81</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/ActionFacade.cpp</td>
<td>2</td>
<td>32</td>
<td>1235</td>
<td>85</td>
<td>42</td>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelFacade.cpp</td>
<td>2</td>
<td>15</td>
<td>1521</td>
<td>12</td>
<td>32</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelFacade.cpp</td>
<td>2</td>
<td>89</td>
<td>1144</td>
<td>42</td>
<td>15</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelFacade.cpp</td>
<td>2</td>
<td>89</td>
<td>1144</td>
<td>42</td>
<td>15</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>DataGroup/ModulA/panelFacade.cpp</td>
<td>2</td>
<td>89</td>
<td>1144</td>
<td>42</td>
<td>15</td>
<td>36</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 5.4: An illustration of incomplete metadata for metrics and for entities, based on artificial data set 5.
(i) Imprecise metadata for a metric.

a) Definition: The metric metadata do not explicitly describe the metric used to measure the attribute of the entity.

b) Example: The metric metadata for LOC (making the assumption that LOC is the metric label) do not explicitly describe the kind of LOC (lines of code) being measured (e.g., physical lines or executable lines).

(ii) Imprecise metadata for an entity.

a) Definition: The entity metadata do not explicitly describe the property of entity whose values distinguish between entities.

b) Example: The entity metadata for ModNm (making the assumption that ModNm is the entity label) do not explicitly describe the property of the entity (e.g., module name or module number).

Figure 5.5: Illustration of imprecise metadata for metrics and for entities, using artificial data set 6.

<table>
<thead>
<tr>
<th>ModNm</th>
<th>NOF</th>
<th>NPM</th>
<th>LOC</th>
<th>nCom</th>
<th>nBug</th>
</tr>
</thead>
<tbody>
<tr>
<td>/DataGroup/ModuleA01</td>
<td>15</td>
<td>1991</td>
<td>190</td>
<td>66</td>
<td>11</td>
</tr>
<tr>
<td>/DataGroup/ModuleB01</td>
<td>45</td>
<td>1980</td>
<td>230</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>/DataGroup/ModuleC02</td>
<td>57</td>
<td>2001</td>
<td>151</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>/DataGroup/ModuleD02</td>
<td>25</td>
<td>1289</td>
<td>103</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>/DataGroup/ModuleE03</td>
<td>25</td>
<td>1672</td>
<td>102</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>/DataGroup/ModuleF03</td>
<td>39</td>
<td>1091</td>
<td>44</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>/DataGroup/ModuleG04</td>
<td>45</td>
<td>2718</td>
<td>39</td>
<td>85</td>
<td>5</td>
</tr>
<tr>
<td>/DataGroup/ModuleH04</td>
<td>19</td>
<td>2388</td>
<td>110</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>/DataGroup/ModuleI05</td>
<td>49</td>
<td>2918</td>
<td>19</td>
<td>42</td>
<td>2</td>
</tr>
<tr>
<td>/DataGroup/ModuleJ05</td>
<td>27</td>
<td>1260</td>
<td>78</td>
<td>24</td>
<td>22</td>
</tr>
</tbody>
</table>

5.3 Quality assessment

As discussed in Chapter 3, a number of researchers have proposed approaches to deal with data quality issues and to improve the quality of data sets [5, 7, 34, 70, 77]. Some approaches have attempted to improve the quality of data sets by developing new methods. Others approaches have used statistical methods to deal with quality issues in data sets. However, none of these approaches provides solutions that determine whether data sets have sufficient information for interpretation in evaluating their quality.

The purpose of the quality assessment process is to understand the quality of a data set better and to identify potential quality issues in the data set. We designed four steps in the quality assessment process to focus on determining the degree to which the data set
can be interpreted correctly. Further, this quality assessment process allows researchers to draw a conclusion about whether the data set has sufficient metadata to support correct interpretation for analysis in empirical research.

Figure 5.6 shows a diagram of the quality assessment process. This assessment process requires a model of the data set (output from the dataset modelling process) and a formal definition of data quality issues. In the following sections we describe how we assessed model of the data set to determine the actual content of the data set and then identified the data quality issues by applying the formal definitions of data quality issues. Next, we present an evaluation of the level of completeness and accessibility of the metadata. Finally, we present the end of the process, which was to prepare an assessment report in the form of a data quality report.

![Figure 5.6: The quality assessment process](image)

5.3.1 Assessing the data set

The first step in quality assessment is determining the contents of the data set. For this, we needed a model of the data set that illustrated the identified dataset category elements (as described earlier). Some model of data set may illustrate few dataset category elements. This could be because of insufficient information in the metadata of the data set. For example, a model of a data set that does not illustrate the metric metadata might result from a lack of information describing the metric used to measure the entity.

To assess a data set, we started by identifying the dataset category elements that are present and not present in the model of the data set — an absence of dataset category elements may raise problems for identifying data quality issues. For example, if the record identifier is not present, we would not be able to distinguish between records, and hence
would be unable to identify quality issues in the form of duplicate data or inconsistent data.

After the dataset category elements that are present and not present in the model of the data set had been identified, we determined information about the entity from the model of the data set. This information was required to identify the entity being measured and to facilitate procedures to identify quality issues.

To facilitate this, we have set out a formal procedure and accompanying advice in Table 5.1.

Table 5.1: Step 1: Assess the data set

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the elements of dataset category that are present and not present in the model of the data set.</td>
<td>Search for dataset category elements that are present and not present in the model of the data set. List them.</td>
</tr>
<tr>
<td>2.</td>
<td>Describe the entities.</td>
<td>If record identifiers and entity label are present in the model of the data set, find entity metadata. If found, determine the entities. Describe them. If not found, make an assumption for the entities based on the record identifiers and entity label. Describe the entities.</td>
</tr>
</tbody>
</table>

To demonstrate the first step of the quality assessment process, we use Example 5, the PROMISE Ant 1.3 data set, from Chapter 2. This data set contains elements that we assumed were labels and values. In Chapter 4, we modelled this data set using our dataset metamodel and produced the model of the Ant 1.3 data set as shown in Figure 5.7.

![Figure 5.7: A model of Ant 1.3 data set](image)

Applying the formal procedure listed in Table 5.1 to the Ant 1.3 data set gave the following outcomes (Figure 5.7):

(a) The model of the Ant 1.3 data set presented seven dataset category elements: entity label, record identifier, metric label, measurement value, ancillary label, ancillary value and record.

(b) The dataset category elements that were not present in the Ant 1.3 data set model were metric metadata, entity metadata and ancillary metadata.
5.3. QUALITY ASSESSMENT

There was no entity metadata in the data set. We assumed the record identifiers represented the conventions for class names in the Ant project. We assumed the entities were the classes of Ant project.

Every assumption we made to assess the model of the data set was based on information about the data set. For example, in Figure 5.7, we made justifiable assumptions based on the dataset category elements and structure of the Ant 1.3 data set. It is possible that the assumptions made by researchers about data sets are wrong. Incorrect assumptions may lead to incorrect interpretation of the data set, raising the risk of data quality issues. This may influence the second step in the quality assessment process, which is to identify data quality issues.

5.3.2 Identifying data quality issues

The second step of quality assessment aims to identify data quality issues. In this step, we set out a formal procedure to identify quality issues and provide notes on its use (see Table 5.2). We used the standard terminology from the dataset metamodel (e.g., the dataset category elements) to allow a common interpretation. For example, we used the element record in the advice column to identify common data quality issues (e.g., duplicate data, inconsistent data and missing data).

The formal procedure and advice given in Table 5.2 contain objective and subjective assessment to identify the quality issues. The objective assessment requires researchers to determine the specific dataset category elements and use them to make decisions when identifying the quality issues. For example, researchers need to determine the record identifiers, measurement values and metric labels in the data set to make decision when identifying duplicate data or inconsistent data. The data quality issues that can be evaluated objectively are duplicate data, inconsistent data, missing data, incorrect data, incomplete metadata for a metric and incomplete metadata for an entity.

The subjective assessment requires researchers to interpret the specific dataset category elements to make decisions in identifying the quality issues. The interpretation often depends on the researchers’ experience and personal preferences. For example, to identify imprecise metadata for a metric, researchers need to examine the degree of metric metadata that describe the metric. In this case, novice researchers may find it more challenging to examine the degree of metadata than expert researchers. The data quality issues that can be evaluated subjectively are imprecise metadata for a metric and imprecise metadata for an entity.

In Section 5.2, we set out eight formal definitions for data quality issues. For some of these, it was not possible to identify problems in the absence of dataset category elements. To explore this further for the purpose of quality assessment, we classified data quality issues into two categories: identified data quality issues and unidentified data quality issues. We describe these two categories as follows:

(a) Identified data quality issue: The quality issue can be identified based on the dataset category elements present in the data set.
Table 5.2: Step 2: Identify data quality issues

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Identify duplicate data</td>
<td>Search for identical record identifiers in the records. If not found, go to procedure No. 3. If found, check whether the measurement values for all metric labels are identical. If this is the case, then these records are duplicate data. List the record identifiers for these records.</td>
</tr>
<tr>
<td>2.</td>
<td>Identify inconsistent data</td>
<td>Search for identical record identifiers in the records. If not found, go to the next procedure. If found, check whether the measurement values for all metric labels are identical. If this is not the case, then these records are inconsistent data. List the record identifiers for these records.</td>
</tr>
<tr>
<td>3.</td>
<td>Identify missing data</td>
<td>Search for records that have either an empty string or null value for a given label. If not found, go to the next procedure. If found, list the record identifiers. If there is no record identifier, list the associated label.</td>
</tr>
<tr>
<td>4.</td>
<td>Identify incorrect data</td>
<td>Search for records that have implausible measurement values for a given metric label. If not found, go to the next procedure. If found, find data type metadata and check whether the implausible measurement values represent the legal measurement values for the metric label. If this is not the case, then these records contain incorrect data. For each record that contains incorrect data, list the record identifier. If there is no record identifier, list the associated label.</td>
</tr>
<tr>
<td>5.</td>
<td>Identify incomplete metadata for a metric</td>
<td>Search for metric labels that do not have metric metadata. If not found, go to the next procedure. If found, list the metric labels.</td>
</tr>
<tr>
<td>6.</td>
<td>Identify incomplete metadata for an entity</td>
<td>Search for entity labels that do not have entity metadata. If not found, go to the next procedure. If found, list the entity labels.</td>
</tr>
<tr>
<td>7.</td>
<td>Identify imprecise metadata for a metric</td>
<td>Search for metric metadata. If found, check whether the metric metadata explicitly describe the metric used to measure the attribute of an entity. If this is not the case, then this metric metadata contains imprecise metadata for the metric. List the metric labels. If metric metadata is not found, this data set has a risk of misinterpretation.</td>
</tr>
<tr>
<td>8.</td>
<td>Identify imprecise metadata for an entity</td>
<td>Search for entity metadata. If found, check whether the entity metadata explicitly describe the property whose values distinguish between entities. If this is not the case, then this entity metadata contains imprecise metadata for the entity. List the entity label. If entity metadata is not found, this data set has a risk of misinterpretation.</td>
</tr>
</tbody>
</table>

(b) Unidentified data quality issue: The quality issue cannot be identified because of the absence of some dataset category elements.

With regard to the absence of metadata for metric and entity, there were two metadata-related quality issues that could not be identified: imprecise metadata for a metric and for
5.3. QUALITY ASSESSMENT

an entity. Although we classified them as unidentified data quality issues, they actually indicated insufficient metadata in data sets, which may have led to misinterpreting the data.

To demonstrate the second step of the quality assessment process, we again used the Ant 1.3 data set (see Figure 5.7) and applied the formal procedure described in Table 5.2. We found the following data quality issues:

(i) Identified data quality issues

(a) Incomplete metadata for a metric: The first row contains metric labels (wmc, dit, noc, cbo, rfc, lcom, ca, ce, npm, lcom) that do not have metric metadata.

(b) Incomplete metadata for an entity: The first row contains an entity label (name) in the third column that does not have entity metadata.

(ii) Unidentified data quality issues

(a) Imprecise metadata for an entity: This data set does not contain entity metadata.

(b) Imprecise metadata for a metric: This data set does not contain metric metadata.

5.3.3 Evaluating the metadata of data sets

For this step, we introduce an evaluation scale to assess the completeness and availability of metadata. The scale describes different characteristics of metadata and their usefulness, using the number of stars to indicate the level of completeness and availability. The scale consists of five star levels as shown in Table 5.3.

<table>
<thead>
<tr>
<th>Metadata scale</th>
<th>Criteria</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Stars</td>
<td>Metadata are not visible except for labels available to represent the property or attribute of entity.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>1 Star</td>
<td>Metadata are visible, but do not explicitly describe the meaning of the property or attribute of entity.</td>
<td>In the extended data set (a single file and/or external locations)</td>
</tr>
<tr>
<td>2 Stars</td>
<td>Metadata are visible, explicitly describe the meaning of the property or attribute of entity, but are not precise.</td>
<td>In the extended data set (a single file and/or external locations)</td>
</tr>
<tr>
<td>3 Stars</td>
<td>Metadata are visible, explicitly describe the meaning of the property or attribute of the entity, and are precise.</td>
<td>In the extended data set (a single file and/or external locations)</td>
</tr>
<tr>
<td>4 Stars</td>
<td>Metadata are visible, explicitly describe the meaning of the property or attribute of the entity, and are precise.</td>
<td>In the data set</td>
</tr>
</tbody>
</table>

Table 5.3: An evaluation scale of metadata
In this scale, 0 stars indicates that the data set does not contain any metadata except for the labels (e.g., metric label is CCN). In the case where there is also no label, the rating of the metadata are also considered to be 0 stars. One star indicates that the metadata are available in the extended data set but the meaning of the labels is not explicitly described—an example in the current case would be the metadata for the label CCN that read `cyclomatic (code)`.

Two stars indicate that the metadata are available in the extended data set but the metadata are not precise—an example in this research would be the metadata for the label CCN that read `a count of the number of cycles in the program flow control graph`. A 2-star rating implies that researchers are able to use the data set, but will have to make some assumptions.

Three stars indicate that the metadata are available, sufficient, precise and accessible in the extended data set. An example from the current study would be the metadata for the metric label CCN that read `for each module to be e - n + 2, where e and n are the number of edges and nodes in the control flow graph`, which would be considered 3 stars. Four-star metadata are equivalent to 3-star metadata but are also accessible within the data set itself. In our example, 4-star metadata for the metric label CCN would be the same as for the 3-star case, but in addition, the metadata would be accessible in the data set.

Both 3-star and 4-star metadata allow researchers to reuse the data set with confidence because they do not need to make any assumptions to interpret the meaning of the data set. To qualify for these two ratings, the metadata should be described explicitly, including the conventions that the researchers used when creating the data sets. For example, the metadata for record identification in the data set should describe a record that contains a list of values associated with a particular entity.

Except for 4 stars, each star rating indicates that metadata are available in the extended data set, which includes all locations that include a single file and the external locations (as described in Chapter 5). When rating metadata in this study, we took into account the exact location of the externally located metadata. For example, if the metadata were located on the web page of the data set, we modelled the metadata as secondary inputs to the modelling process, because we consider web pages to be as only one stage further from where we accessed the data set.

If we were to allow more than one stage for the accessibility of metadata, we might encounter problems ensuring a consistent interpretation from the external locations with the actual data file. For example, in some cases the label (e.g., identifier) for the metric found in the actual data file differs from that found in related metadata stored in an external locations. This makes it difficult to relate the metric in the data set to its equivalent in the metadata.

To execute this third step in the quality assessment process, we set out a formal procedure with advice that contains step-by-step instructions to evaluate the entity and metric metadata of the data set. This formal procedure and advice contain subjective evaluation that depends on the researcher’s experience to evaluate the metadata. The
5.4. RESULTS OF QUALITY ASSESSMENT

The procedure starts by evaluating each metadata individually. The next step is to evaluate the overall rating of the metadata, which is assigned the same rating as the lowest of its components. The procedure is shown in Table 5.4.

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the entity metadata. Evaluate.</td>
<td>Find the entity metadata that describe the entity label. Determine whether the entity metadata explicitly describe the meaning of the property of the entity. Identify the location of metadata. Assign and list a rating.</td>
</tr>
<tr>
<td>2.</td>
<td>List the metric metadata. Evaluate.</td>
<td>Find the metric metadata that describe the metric label. Determine whether the metric metadata explicitly describe the meaning of the attribute of the entity. Identify the location of metadata. Assign and list a rating.</td>
</tr>
<tr>
<td>3.</td>
<td>Evaluate the overall metadata rating. List.</td>
<td>Review the ratings for entity and metric metadata. Assign the minimum rating to indicate the overall rating of the metadata. List.</td>
</tr>
</tbody>
</table>

When implementing the procedure in Table 5.4 to evaluate the metadata of the Ant 1.3 data set, we observed that there were no metadata available for either the entity label or the metric label. Thus, the overall rating of the metadata for the Ant 1.3 data set was 0 stars. The complete quality assessment for the Ant 1.3 data set is shown in Appendix F.

In this research, we used the rating of the metadata as a guide to improve the quality of metadata. For example, once a data set has been assessed and assigned a rating, researchers should strive to improve the quality of the metadata in terms of completeness and availability as much as possible to ensure the data set has sufficient metadata to be usable for analysis in future research. Guidelines for providing good-quality data sets are discussed further in Chapter 6.

5.3.4 Preparing the assessment report

The assessment report was intended to be a standalone report documenting outcomes, observations and suggestions to improve the quality of the data set. To prepare the assessment report, we set out a formal procedure to consolidate and review the outcomes from previous steps (Table 5.5). (The complete formal procedure for the quality assessment process is included in Appendix E.)

5.4 Results of Quality Assessment

We produced two reports on the results of the quality assessment process: a summary report and a full report. The summary report presents an overview of results, and the full report presents the results in detail and provides suggestions to improve the quality of the data set.
Table 5.5: Step 4: Prepare the assessment report

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedures</th>
<th>Type of report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the dataset category elements that are present in the model of the data set.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>2.</td>
<td>List the dataset category elements that are not present in the model of the data set. Describe them.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>3.</td>
<td>For each dataset category elements that are not present in the model of the data set, describe suggestions to improve the quality of the data set.</td>
<td>Full</td>
</tr>
<tr>
<td>4.</td>
<td>List identified and unidentified data quality issues. Describe them.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>5.</td>
<td>Describe the overall results for the metadata evaluation scale.</td>
<td>Summary</td>
</tr>
<tr>
<td>6.</td>
<td>Describe the rating of each element of the metadata, and the overall rating.</td>
<td>Full</td>
</tr>
</tbody>
</table>

5.4.1 Summary report

The summary report provides a brief description of each of the outcomes of the steps in the quality assessment process. It describes the list of data set elements (dataset category and dataset concept) that were present and those that were missing from the model of the data set, and it lists both identified and unidentified data quality issues. It also describes the overall rating for the metadata evaluation scale. We illustrate this with an example summary report for the Ant 1.3 data set in Figure 5.8 (see Appendix F for more detail).

![Figure 5.8: An example summary report for results of the quality assessment](image-url)
5.4. RESULTS OF QUALITY ASSESSMENT

5.4.2 Full report

The full report provides a detailed description of each of the outcomes of the quality assessment process. Similarly to the summary report, the full report describes which data set elements (dataset category and dataset concept) existed in the model of the data set and which did not, along with both the identified and unidentified data quality issues. The descriptions of existent and non-existent data set elements contain detailed observations for the quality assessment. The full report also provides suggestions to improve the quality of the data set.

With regard to the metadata evaluation scale, the full report presents the outcomes for each element. The final rating was produced by selecting the minimum rating if one metadata element was missing. We illustrate with an example of a full report for the Ant 1.3 data set in Figure 5.9 (see Appendix G for more detail).

![Assessment results: Full report](image)

<table>
<thead>
<tr>
<th>Assessment results: Full report</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
</tr>
<tr>
<td>Elements of dataset category that are present in the model of data set</td>
</tr>
<tr>
<td>Elements of dataset category that are not present in the data set</td>
</tr>
<tr>
<td><strong>Element of dataset category</strong></td>
</tr>
<tr>
<td>Entity metadata</td>
</tr>
<tr>
<td>Metric metadata</td>
</tr>
<tr>
<td>Ancillary metadata</td>
</tr>
<tr>
<td><strong>Result for identifying data quality issues</strong></td>
</tr>
<tr>
<td>Identified data quality issues</td>
</tr>
<tr>
<td>Underidentified data quality issues</td>
</tr>
<tr>
<td><strong>Result for evaluation scale of metadata</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric label</th>
<th>Scale of metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 star</td>
</tr>
<tr>
<td>2</td>
<td>0 star</td>
</tr>
<tr>
<td>3</td>
<td>0 star</td>
</tr>
<tr>
<td>4</td>
<td>0 star</td>
</tr>
<tr>
<td>5</td>
<td>0 star</td>
</tr>
<tr>
<td>6</td>
<td>0 star</td>
</tr>
<tr>
<td>7</td>
<td>0 star</td>
</tr>
<tr>
<td>8</td>
<td>0 star</td>
</tr>
<tr>
<td>9</td>
<td>0 star</td>
</tr>
<tr>
<td>10</td>
<td>0 star</td>
</tr>
<tr>
<td>11</td>
<td>0 star</td>
</tr>
<tr>
<td>12</td>
<td>0 star</td>
</tr>
<tr>
<td>13</td>
<td>0 star</td>
</tr>
<tr>
<td>14</td>
<td>0 star</td>
</tr>
<tr>
<td>15</td>
<td>0 star</td>
</tr>
<tr>
<td>16</td>
<td>0 star</td>
</tr>
<tr>
<td>17</td>
<td>0 star</td>
</tr>
<tr>
<td>18</td>
<td>0 star</td>
</tr>
<tr>
<td>19</td>
<td>0 star</td>
</tr>
<tr>
<td>20</td>
<td>0 star</td>
</tr>
<tr>
<td>21</td>
<td>0 star</td>
</tr>
</tbody>
</table>

Figure 5.9: An example full report for results of the quality assessment
5.5 Applying quality assessment to the data sets

In this section, we evaluate the use of our quality assessment process by applying it to the existing data sets from the public data repositories. Similarly to Section 4.5, we selected 87 data sets that had similar structures to the examples of the data sets used in Chapter 2. In total, we applied our quality assessment process to 92 data sets, including the examples of the data sets given in Chapter 2.

We present a matrix for the results of applying the quality assessment process to the existing data sets in Figure 5.10 (See Appendix G). The columns of the matrix consist of step 1, step 2 and step 3. Step 1 presents 11 dataset category elements to indicate the elements that were present and not present in the model of the data set. For simplicity in presenting the matrix, we have excluded procedure 2 in step 1 of quality assessment process. Step 2 presents eight data quality issues. Step 3 provides the star rating for metric metadata and entity metadata. We have excluded step 4 of the quality assessment process in the matrix because it describes the procedure to prepare the quality assessment reports of data sets. The rows contain 92 data sets, including the examples of real data sets given in Chapter 2. In this matrix, we grouped a number of data sets that had the same structure and column headers in a single row, for example, the Ant 1.3 data set had the same structure and column headers as 32 data sets. We have used colours to indicate relations between step 1 to step 2 and step 3.

The results of step 1 show that many data sets contained dataset category elements that are required for identifying data quality issues in step 2. We found 57 of 92 data sets contained record identifier, metric label and measurement value, which are required for identifying common data quality issues (e.g., duplicate data and inconsistent data). In addition, we found 18 data sets contained metric metadata and 8 data sets contained entity metadata (as described in Chapter 5). These metric and entity metadata are required for identifying metadata-related quality issues (e.g., incomplete metadata for a metric and imprecise metadata for an entity).

In step 2, we found a number of data quality issues in the data sets as shown in Figure 5.10. The data quality issues are: duplicate data (1 data set), missing data (8 data sets), incorrect data (5 data sets), incomplete metadata for a metric (69 data sets), incomplete metadata for an entity (79 data sets), imprecise metadata for a metric (12 data sets) and imprecise metadata for an entity (3 data sets). In particular, regarding incomplete metadata for a metric and for an entity, there were no metadata for all the metric labels and entity labels.

However, we could not identify some of the common data quality issues because of the absence of some dataset category elements in step 1. The unidentified common data quality issues were duplicate data and inconsistent data for 35 data sets that did not have record identifiers, incorrect data for 48 data sets that did not have metric metadata and data type metadata, and missing data for 5 data sets (Nrp1) that did not have other dataset category elements except measurement values. In Figure 5.10, we have used green to highlight the dataset category elements that are not present in the data sets (step 1) and the unidentified data quality issues (step 2).
### Table 5.6: Specific Dataset Category Elements that are Required to Identify Some of the Data Quality Issues

<table>
<thead>
<tr>
<th>No.</th>
<th>Measurement data set (count of similar data sets)</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
<th>Element present</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PROMISE AMI</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>QualityWeb 2.7.5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>PROMISE NaCl</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>PROMISE 25020</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>PROMISE Coarse</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>PROMISE NaClII</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>PROMISE Nc</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>12</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>13</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>14</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>16</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>17</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>19</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>21</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>22</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>23</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>24</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>25</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>26</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>27</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>28</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>29</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>30</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>31</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>32</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>33</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>34</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>35</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>36</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>37</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>38</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>40</td>
<td>PROMISE 8N</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

**Legend:**
- **X:** Element present in the data set.
- **-:** Element present in the data set.
- **N:** Number of (element) present in the data set. Total number of labels in the data set.
- **E:** The quality issue is identified in the data set.
- **U:** The quality issue is unclassified due to the absence of specific dataset category elements.
- **A (B):** A is counting for (category) metadata and B is number of category labels.
- **R:** Risk of misinterpretation is identified in the data set.

Figure 5.10: Matrix of quality assessment process.

Table 5.6 lists the specific dataset category elements that are required to identify some of the data quality issues.

Regarding metadata-related quality issues, we also could not identify imprecise metadata for an entity and for a metric in the data sets because of the absence of metadata. However, these data sets may have carried risks of misinterpretation. In particular, we identified 80 data sets that may have carried risks of misinterpretation because of the absence of entity metadata and 71 data sets that may have carried risks of misinterpretation because of the absence of metric metadata. We indicate these data sets...
Table 5.6: Dataset category elements required to identify the quality issues

<table>
<thead>
<tr>
<th>Element category</th>
<th>Data quality issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record identifier</td>
<td>Duplicate data and inconsistent data</td>
</tr>
<tr>
<td>Metric metadata</td>
<td>Imprecise metadata for a metric</td>
</tr>
<tr>
<td>Entity metadata</td>
<td>Imprecise metadata for an entity</td>
</tr>
<tr>
<td>Metric metadata and data type metadata</td>
<td>Incorrect data</td>
</tr>
</tbody>
</table>

as R for risk of misinterpretation in Figure 5.10.

In step 3, we found 70 data sets were assigned 0 stars for rating metric metadata and entity metadata because of the absence of metric and entity metadata in the data sets. This can be seen in Figure 5.10 where we have used the same colour (orange and red) to highlight the dataset category elements that were not present in data sets (step 1) and the rating for metric metadata and entity metadata (step 3).

Further, we found six out of 18 data sets that contained metric metadata were assigned a rating of 2 stars and the remaining data sets were assigned a rating of 1 star. In particular, two of six data sets that were assigned with a rating of 2 stars had a few of their metric metadata were assigned with a rating of 1 star. For entity metadata, we found four of eight data sets were assigned a rating of 2 stars and the remaining data sets were assigned a rating of 1 star. This shows that although eight data sets contained both metric metadata and entity metadata, and 10 data sets contained only metric metadata, few data sets actually contained adequate metadata to support the interpretation.

In summary, these results reveal that some of the dataset category elements were essential for evaluating the quality of data sets. These dataset category elements were record identifier, measurement value, metric label, metric metadata, entity metadata and data type metadata. Some common data quality issues (e.g., duplicate data and inconsistent data) could not be identified because of the absence of some of these dataset category elements (e.g., record identifiers). In particular, with regard to metadata-related quality issues, although some of these could not be identified in the data sets, these data sets carried risks of misinterpretation because of the absence of metadata. Further, the metadata rating of 0 stars is given to the data sets that did not have metric metadata and entity metadata.

5.5.1 Discussion of the results for applying the quality assessment process

In this section, we discuss in more precise detail the results of applying the quality assessment process for three data sets from the matrix of quality assessment. The three data sets are PROMISE AM1, KC2 and NrpClassic. We chose these three data sets because they indicated interesting results.

In the PROMISE AM1 dataset, we did not find any data quality issues or risk of misinterpretation because all the data set category elements were present. In terms of the rating for metric metadata, we rated two metric labels with 1 star because the metadata
were present and another two metric labels with 2 stars because the metadata described explicitly the meaning of the metrics. For entity metadata, we rated one entity label with 2 stars because the metadata described explicitly the meaning of the property. This indicates that the PROMISE AM1 data set may have presented a reasonable quality of data because of the existence of all the dataset category elements that could support the correct interpretation of data.

As shown in the matrix in Figure 5.10, we found that the PROMISE KC2 data set contained three data quality issues, which were incorrect data, incomplete metadata for entity and imprecise metadata for metric. In the case of incorrect data, we found a measurement value ‘1.1’ for the metric label ‘loc’ had a data type ‘numeric’; however, the metric metadata for the metric label ‘loc’ did not describe the data type for the measurement value. We identified the imprecise metadata for metric because the metric metadata for ‘loc’ describe ‘McCabe’s line count of code’, which did not describe explicitly the metric used to measure the entity. We also identified the incomplete metadata for entity in addition to a risk of misinterpretation in this data set because of the absence of entity metadata. However, we could not apply the formal definitions for duplicate data and inconsistent data because of the absence of record identifiers in the data set.

Regarding the overall rating for metric metadata, we rated the 19 metric labels with 1 star because the metric metadata were present. Although this data set contained few data quality issues and a risk of misinterpretation, it contained some metric metadata that may help researchers to interpret the data.

The PROMISE NrpClassic data set contained only measurement values and records for the data set category elements, as shown in the matrix of quality assessment. In this data set, we could not identify the four common data quality issues (e.g., duplicate data, inconsistent data) because of the absence of the specific elements to identify these quality issues. Further, we identified more than one indicator for risk of misinterpretation because of the absence of metric metadata and entity metadata.

Although the overall rating of metadata for the NrpClassic data set (0 stars) was the same as another 65 data sets, it contained more unidentified data quality issues and indicators for risks of misinterpretation than the 65 other data sets because of the absence of dataset category elements. This indicates that the author of a data set should provide the essential dataset category (e.g., record identifier, metric metadata and entity metadata) elements to assist researchers in interpreting the data for identifying data quality issues as well as for evaluating the metadata.

5.6 Conclusion

This chapter has described in detail the framework for evaluating the quality of data sets; specifically, it described the quality assessment process. The quality assessment process focuses on determining whether a data set includes sufficient information to be usable for analysis in empirical research. This assessment process was designed to assess the quality of a data set based on the model of the data set. The chapter has also presented
formal definitions for data quality issues and developed a metadata evaluation scale to facilitate the assessment process.

The uses of our quality assessment process are manifold: to assess the data set, to identify data quality issues, to evaluate the metadata and to prepare the assessment report. We captured the results of the quality assessment and proposed suggestions for data quality improvement in two data quality reports: a summary report and a full report. The practical use of this quality assessment process was illustrated with reference to the existing data sets. Further discussion on suggestions to improve the quality of data sets and guidelines for creating data sets are presented in Chapter 6.
GUIDELINES FOR CREATING GOOD-QUALITY DATA SETS

This chapter presents the guidelines for creating good-quality data sets that have been established in our data quality framework. The guidelines provide details on the contents and metadata of data sets, as well as the relationships between the elements in a data set. The chapter describes the components used to construct the guidelines and also describes how to use the guidelines to facilitate data sharing in research.

6.1 Introduction

In Chapter 1, we discussed the fact that a primary concern in data quality research is the poor quality data that exist in data sets from public data repositories. This issue arises as a result of data quality issues in data sets, such as issues relate to the accuracy of data (e.g., missing data, duplicate data and incorrect data) and issues relate to the interpretation of data (e.g., misinterpretation of the data). In Chapter 2, we found that most researchers create and share data sets in public data repositories without providing adequate metadata. In particular, they often do not describe metadata for the metrics and entities that are required to interpret the measurements. In Chapter 4, we evaluated data sets in public data repositories with our dataset metamodel and found that very few data sets actually contain adequate metadata to describe the meaning of the properties of the entity. A standard way of creating a data set with adequate metadata to facilitate the correct interpretation of data is therefore required.

We presented a data quality framework that incorporates a dataset metamodel and a quality assessment process in Chapter 5. The dataset metamodel provides standard terminology to describe the structure of data sets, while the quality assessment process provides formal procedures to evaluate the quality of data sets. In this chapter, we draw a set of essential terminology and procedures from the framework that should be considered when creating data sets.

The need to provide a standard way to create a data set that contains adequate
metadata has motivated us to construct these guidelines for the creation of good-quality data sets. The guidelines aim to help researchers (particularly the authors of data sets) to prepare data sets that are well organised and documented, contain adequate metadata and are easily accessible in order to facilitate research that reuses the data sets. In this chapter, we provide guidelines to describe adequate metadata that contain the intended meaning of the data and provide a context for interpreting the data.

Before presenting the guidelines, we describe the components that we use to construct the guidelines in section 6.2. A complete set of the guidelines for creating data sets are discussed in detail in section 6.3 and a description of how to use the guidelines is presented in section 6.4.

### 6.2 Constructing the guidelines

In this section, we describe the components used to construct the guidelines for the creation of data sets. The components come mainly from the framework, the dataset metamodel and the quality assessment process. We have also included some of the practices gathered throughout our research in the guidelines.

First, we define the contents of the data set by applying the essential standard terminology from the dataset metamodel, the *dataset category* elements, to the guidelines. These elements are *record*, *record identifier*, *measurement value*, *ancillary value*, *metric label*, *entity label*, *ancillary label*, *metric metadata*, *entity metadata*, *data type metadata* and *ancillary metadata*. For the purposes of convenience, the dataset metamodel is shown in Figure 6.1 and the *dataset category* elements are highlighted in blue.

![Figure 6.1: Dataset metamodel](image)

Second, we extract some of the formal procedures from the framework, particularly the quality assessment process described in Chapter 5. We revise these formal procedures to become the guidelines for creating the data set. For example, we revise some of the
formal procedures and the criteria in the evaluation scale for the metadata in order to use them in the guidelines for providing metadata for a data set.

Third, we include some of the practices gathered throughout our research in the guidelines. Such practices include recording the metadata in the data set file (as discussed in Chapter 4).

6.3 The guidelines for creating data sets

In this section, we present the guidelines for creating good-quality data sets. The guidelines aim to provide a complete data set documentation by providing a clear description of how researchers should define the contents of a data set. It also provides details on how to describe adequate metadata to allow the correct interpretation of data.

Our guidelines comprise two stages: (1) define the contents of a data set; (2) provide the metadata for the data set. Each stage of the guidelines is described in detail below:

(1) Define the contents of a data set.

A data set contains a collection of elements that must include value, label and metadata.

(a) Value is an element that must be recorded to describe a characteristic of an entity. The value must be either measurement value, record identifier or ancillary value.

i. Measurement value is a value that is obtained through the process of measurement that must be associated with a metric.

ii. Record identifier is a value in a record that must be used to represent an entity and it should be used to distinguish one entity from another.

iii. Ancillary value is a value that must be used to represent neither measurement value nor record identifier.

(b) A record must contain a list of values that is associated with a particular entity. The value for the record must be either measurement value, record identifier or ancillary value.

(c) Label is an element that must be used to represent a characteristic of an entity. The label must be either metric label, entity label or ancillary label.

i. Metric label is a label that must represent a metric that measures an attribute of an entity and must be associated with measurement value.

ii. Entity label is a label that must represent a property whose values distinguish between entities and must be associated with record identifier.

iii. Ancillary label is a label that must be neither entity label nor metric label and must be associated with ancillary value.

(d) Metadata is a set of elements that must be used to represent a description of the characteristics of an entity. The metadata must be either metric metadata, entity metadata, data type metadata or ancillary metadata.
6.4. HOW TO USE THE GUIDELINES

i. *Metric metadata* is a set of elements that must describe the metric used to measure an attribute of an entity and must be associated with *metric label*.

ii. *Entity metadata* is a set of elements that must describe the property whose values distinguish between entities and must be associated with *entity label*.

iii. *Data type metadata* is a set of elements that must describe the data type of a measurement value.

iv. *Ancillary metadata* is a set of elements that must be not either metric metadata, entity metadata or data type metadata and must be associated with *ancillary label*.

(2) Provide the metadata for the data set.

a) Metadata is a set of elements that must represent a description of the characteristics of an entity. The description in the metadata should:

i. state the essential meaning.
   The description should describe explicitly the characteristics of the entity.

ii. be precise and unambiguous.
   The meaning and interpretation of the characteristics of the entity should be apparent from the descriptions. The descriptions should be clear enough to allow for only one possible interpretation.

iii. be concise.
   The descriptions should be brief and comprehensive.

b) Metadata should be recorded in the same file as the measurement values; it is possible for metadata recorded in external locations to miss updates.

6.4 How to use the guidelines

These guidelines are aimed at software engineering researchers: both researchers who are the authors of data sets and researchers who intend to assess the quality of an existing data set (who are not the author of the data set). Researchers can use the guidelines in two ways: first, they can use our guidelines while creating data sets (e.g., during the creation of a new data set); second, they can use the guidelines after evaluating the quality of an existing data set (e.g., a data set from a software repository).

The two ways to use the guidelines both fit into the phases of our research data life cycle. As described in Chapter 3, the research data life cycle was introduced to describe the phases that data goes through in the research journey. It starts with the creation of data and goes through to the completion of research based on the data. The research data life cycle consists of five phases: *source, raw, refined, results* and *completion*. For the purposes of convenience, the research data life cycle is shown in Figure 6.2.

In the case of using the guidelines while creating a data set, the relevant phase is the *raw* phase. In this phase, researchers process the raw data using processing methods (e.g., scripts, software metrics tools). They can use the guidelines during these processing...
methods to transform the raw data (e.g., commit logs) into refined data (e.g., a data set). The use of the guidelines in this case is further discussed in section 6.4.1.

Another appropriate phase in which to use the guidelines is in the refined using analysis techniques to produce results data in their research. Before using the guidelines in this phase, researchers need to understand the quality of the refined data. Researchers who are the authors of the data set can use the guidelines immediately to improve the quality of the data set because they already know the quality of their own data set. However, researchers who are not the authors of the data set need to evaluate the quality of the data set before using the guidelines. Once they understand the quality of the data set, they can use the guidelines to improve the quality of the data set. We discuss this case in detail in section 6.4.2.

6.4.1 Using the guidelines while creating data sets

Researchers who are creating data sets using software metrics tools can use the guidelines during the data set creation process. First, they should examine the created data sets and identify which stages of the guidelines they can use to improve their quality. Second, they need to have access to a software metrics program to make changes to the data sets according to the guidelines.

Researchers who do have access to a software metrics program can make the changes to the program and reproduce a new version of the data set. For example, after reading the guidelines, the author of a data set realises that they need to add metadata for all the metric labels in the data set. To do this, they need to follow Stage 2 of the guidelines and make the changes to a software metrics program by adding the metadata. Finally, they re-create the data set using the software metrics tool. In the case of researchers who do not have access to a software metrics program, they can make the changes manually in
To illustrate an example for using the guidelines while creating data sets, we examined artificial data set 5 from Chapter 2. We found that the labels *Filename*, *Phase* and *nBug* do not have metadata. In this situation, according to the guidelines, we should apply Stage 2, which is ‘provide the metadata for the data set’. For this example, we add the metadata for the labels *Filename*, *Phase* and *nBug* because we created this data set manually.

Figure 6.3 illustrates a new version of artificial data set 5. The new elements are entity metadata (pink), ancillary metadata (grey) and metric metadata (light blue).

![Image](image.png)

Figure 6.3: An example of using the guidelines for artificial data set 5.

### 6.4.2 Using the guidelines after evaluating the quality of data set

Researchers who are not the author of a data set should use the guidelines after evaluating the quality of the data set. They should assess the structure and contents of the data set using our data quality framework. The data quality framework produces an assessment report that includes suggestions for improving the quality of the data set. Based on the suggestions, the researchers can make changes to the data set according to the guidelines.

In the following subsections, we present two examples of real data sets from Chapter 2 (PROMISE Boetticher and PROMISE Datatrieve) to illustrate how to use the guidelines after evaluating the quality of the data set.

#### 6.4.2.1 Example 1: PROMISE Boetticher

Part of the Boetticher data set [16] is shown in Figure 6.4; this consists of survey data presented in a typical tabular format. This example contains some issues related to the quality of the data set.

The quality issues described in Chapter 2 are summarised below:
6.4. HOW TO USE THE GUIDELINES

1. We do not know the meaning of the column headers because there is no metadata described in the data set. However, we found that the metadata is provided on the web page where the data set comes from. The metadata presents context information for the data set that includes the publication that uses the data set. We found that the publication does not clearly describe the necessary information required to interpret the meaning of the column headers.

2. We do not know what the actual entity being measured is because there is no entity identification (a value that represents the entity) in the data set.

In this chapter, we evaluate the PROMISE Boetticher data set using our data quality framework. The model of the data set is shown in Figure 6.5 (as described in Chapter 5) and we produce a full report on the results of the quality assessment in Figure 6.6 (note that Figure 7.6 shows only part of the full report). The full report for the PROMISE Boetticher data set is shown in Appendix H.

Figure 6.4: PROMISE Boetticher dataset.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Highest Degree</th>
<th>Undergrad Courses</th>
<th>Comp Sci Undergrad Courses</th>
<th>Comp Sci Grad Courses</th>
<th>Hardware Undergrad Courses</th>
<th>Hardware Grad Courses</th>
<th>MIS Undergrad Courses</th>
<th>MIS Grad Courses</th>
<th>Prof Mgmt Undergrad Courses</th>
<th>Prof Mgmt Grad Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Bachelor</td>
<td>3</td>
<td>18</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>Bachelor</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>Master</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 6.5: A model of the PROMISE Boetticher data set.

In the full report on the Boetticher data set, we examine the ‘Suggestions for improving the data set’. For each suggestion, we identify which stage of the guidelines we should apply to improve the data set. After identifying the applicable stage of the guidelines, we then make the according changes to the data set. Table 6.1 presents the suggestions, the stages of the guidelines and the changes made to the Boetticher data set.
6.4. HOW TO USE THE GUIDELINES

Figure 6.6: Full report on the PROMISE Boetticher data set.

Table 6.1: A summary of changes to the PROMISE Boetticher data set according to the guidelines

<table>
<thead>
<tr>
<th>Suggestions</th>
<th>Guidelines</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Record identifier</td>
<td>Stage 1</td>
<td>We add unique values as record identifiers.</td>
</tr>
<tr>
<td>2. Entity label</td>
<td>Stage 1</td>
<td>We add ‘UserID’ as an entity label.</td>
</tr>
<tr>
<td>3. Entity metadata</td>
<td>Stages 1 and 2</td>
<td>We add entity metadata for the entity label ‘UserID’.</td>
</tr>
<tr>
<td>4. Metric metadata</td>
<td>Stages 1 and 2</td>
<td>We contact the author for the actual metric metadata (for the purpose of example, we add artificial metric metadata based on assumptions).</td>
</tr>
</tbody>
</table>

We illustrate a new version of part of the Boetticher data set in Figure 6.7. The new elements in the Boetticher data set are record identifiers (light blue), entity label (light green), entity metadata (pink) and metric metadata (dark blue).
6.4. HOW TO USE THE GUIDELINES

6.4.2.2 Example 2: PROMISE Datatrieve

We represent part of the Datatrieve data set [16] in Figure 6.8, which consists of source code data presented in an ARFF format. As discussed in Chapter 2, the contents of this data set is organised differently to Example 1: PROMISE Boetticher and some of the metadata is located on the web page of the data set. The metadata on the web page is shown in Figure 6.9.

![Figure 6.7: A new version of part of the PROMISE Boetticher data set](image)

![Figure 6.8: The PROMISE Datatrieve dataset.](image)

The quality issues described in Chapter 2 are summarised below:

1. We found that some of the metadata was located on the web page of the data set. The metadata describes information about context, past usage and attribute information. Context information describes the date the data set was created, the author of the data set and the change history of the data set. Past usage information describes the publication that uses the Datatriev data set. The attribute information describes the meaning of the column headers in the data set.
6.4. HOW TO USE THE GUIDELINES

Figure 6.9: Information about the PROMISE Datatrieve data set.

2. Some of the metadata do not explicitly describe the meaning of the column headers. For example, the metadata for the LOC6_0 does not explicitly describe how to deal with blank lines and comment lines.

3. There is no entity identification (a value that represents the entity) in the data set.

As in Example 1, we evaluate the Datatrieve data set using our data quality framework. The model of the data set and the metadata are shown in Figures 6.10 and 6.11 (as described in Chapter 5). We produce a full report on the results of the quality assessment in Figure 6.12 (See Appendix H).

Figure 6.10: A model of the PROMISE Datatrieve data set.

The full report indicates that the Datatrieve data set has a data quality issue; namely, imprecise metadata for a metric. To solve this issue, we need to contact the author of the Datatrieve data set to get the full description of the metadata for the metric labels. In this case, we do not make any changes to the existing metric metadata.
6.4. HOW TO USE THE GUIDELINES

Based on the suggestions in the full report, we make changes to the Datatrieve data set according to the guidelines. We present the suggestions, the stages of the guidelines and the changes for the Datatrieve data set in Table 6.2.

The full report in Figure 6.12 indicates that the metric metadata is recorded on the webpage where we access the data set. As stated in the guidelines (Stage 2: ‘provide the metadata for the data set’), we should record the metadata in the same file as the measurement values. To apply this guideline, we add the metric metadata from the webpage to the data file.

We illustrate a new version of part of the Datatrieve data set in Figure 6.13. The new elements in the Datatrieve data set are record identifiers (light blue), entity label (light green), entity metadata (pink) and metric metadata (dark blue).
6.5 Conclusion

This chapter has described in detail the guidelines for creating good-quality data sets. The guidelines provide instructions on how to define the structure and contents of a data set. Furthermore, the guidelines provide details on how to describe metadata for the data set.

This chapter has presented two ways to use the guidelines: while creating the data sets and after evaluating the quality of the data set. We have shown the practical use of the guidelines by applying them to the examples of data sets from Chapter 2.
In the next chapter, we present a survey research method for the evaluation of part of the data quality framework. We describe the design, execution and analysis of the survey on the data quality framework. We also discuss the threats to validity and summarise our findings.
This chapter presents a survey to evaluate part of the framework for data quality assessment. The part of the framework presented consists of definitions of elements of the dataset metamodel and formal definitions of data quality issues. We begin by introducing the survey as our research method used for evaluation. Then we present the survey design, execution and results. This is followed by a discussion of the findings resulting from the survey. The chapter ends with some discussion of the threats to the survey's validity.

7.1 Introduction

Evaluation is an essential activity for software engineering research. Conducting a user study for the evaluation of a framework can provide useful feedback for improving the framework. User studies are important for convincing users of the usefulness and feasibility of a framework. The results from user studies can offer important insights into the real actions and preferences of users [91].

In this thesis, we conducted two user studies to evaluate the effectiveness and usability of part of the framework for data quality assessment. First, we conducted a survey as our preliminary evaluation to evaluate the effectiveness of definitions of dataset category elements and the usability of formal definitions of data quality issues. Second, we conducted an observational study to observe how software engineering researchers apply the new definitions of dataset category element and the formal definitions of data quality issues to the data sets (described in Chapter 8).

In this chapter, we used a cross sectional survey design, which is one of the most common types of survey designs in software engineering. This type of survey requires participants to give information at a particular point in time [92]. To design this type of survey, we decided to use an online survey for data collection as it appeared to be the most appropriate method for gathering the detailed information required to find answers
to our research problem over a fixed period of time.

In the following section, we present the design of the survey in section 7.2 and the procedures used to execute it in section 7.3. We discuss the results, potential threats and research validity and summarise our findings in section 7.4. The structure of the survey design, execution and results is based on guidelines for reporting by Jedlitschka [93] and Kitchenham et al. [92]. We followed these guidelines when documenting our survey to ensure that sufficient information was provided to help other researchers to conduct similar studies.

7.2 Survey Design

In this section, we present the design of the survey that was carried out to evaluate part of the framework for data quality assessment. We begin by describing the survey goals and research questions in section 7.2.1. We then discuss the target population of participants in section 7.2.2, the sampling methods in section 7.2.3, the survey material in section 7.2.4 and survey tasks in section 7.2.5.

7.2.1 Goal and survey research questions

We designed the survey goals according to a goal definition template by Wohlin et al. [94]. The goal template contains five important aspects that need to be properly defined to design and execute an online survey. The five criteria are as follows:

(a) Object of study:
   - Definitions of dataset category elements in the dataset metamodel.
   - Formal definitions of data quality issues during quality assessment.

(b) Purpose: The purpose is to evaluate part of the framework for data quality assessment.

(c) Quality focus: The main effect studied in the survey is the effectiveness of definitions of dataset category elements in the dataset metamodel and the effectiveness of formal definitions of data quality issues in the quality assessment.

(d) Perspective: The survey is analysed from the point of view of the researchers.

(e) Context: The survey is executed using researchers in software engineering or computer science as subjects and considering software engineering data sets as objects.

We constructed two main research questions and four secondary research questions for the survey. The survey research questions (SQs) are as follows:

1. SQ1: Is the dataset metamodel for the framework effective in describing the structure and content of the data sets?
   The effectiveness of the dataset metamodel for the framework depends on how
well participants understand the definitions of dataset category elements in the dataset metamodel and apply them to a range of data sets. It also includes participants’ overall acceptance of definitions of dataset category elements in the dataset metamodel, particularly in terms of ease of use.

a) Question: Can participants apply the definitions of dataset category elements from the dataset metamodel to a range of data sets?

b) Question: Do participants perceive the definitions of dataset category elements in the dataset metamodel as easy to use?

2. SQ2: Is the formal definitions for data quality issues in the quality assessment useful for determining quality issues across a range of data sets?
The usability of the framework's quality assessment depends on how well participants identify issues in quality across a range of data sets by applying formal definitions of data quality issues during the quality assessment process.

a) Question: Can participants identify quality issues in data sets using formal definitions of data quality issues?

b) Question: Do participants perceive the formal definitions of data quality issues in the quality assessment process as easy to use?

7.2.2 Participants

The target population for the survey was researchers, including academic, industrial and postgraduate researchers. We aimed to recruit researchers with backgrounds or experience in software engineering or computer science research. In particular, we were interested in researchers who had used or analysed data sets from data repositories.

One of the motivations for taking part in the survey was that participants would gain knowledge in understanding data sets. This would enable them to apply their newly acquired knowledge of data sets to their own data sets during future research. Participation in this survey was voluntary and participants were not required to sign consent forms because the survey was designed to be anonymous.

7.2.3 Sampling methods

We used convenience sampling and snowball sampling to invite potential participants because the target population of participants was specific and limited. Convenience sampling allowed us to invite participants who were available and willing to take part voluntarily. This sampling method was chosen to save time when finding potential participants.

We used snowball sampling to ask the participants who had received our invitation to participate in the survey to invite other potential participants that they believed would be willing to take part in the research. Other participants who were interested in participating in the survey were able to access information about the survey through
forwarded advertisement link. This sampling method helped us to expand the numbers of those participating in the survey.

### 7.2.4 Survey material

We designed a web-based survey to gather responses from participants to save time in distributing the survey. The web-based survey simplified the processes of manually entering data and of participants answering the survey. We developed the web-based survey using Qualtrics software.

The web-based survey consisted of 52 questions and was divided into two sections. (Please refer to Appendix J.1 for the web-based survey questions).

(a) Section 1: Demographics. In questions Q1(a)-Q1(f) we asked participants about their backgrounds in research and their experiences with data sets from data repositories. We also asked participants if they had encountered any quality issues with data sets and to comment on these quality issues. We used the responses collected from this section to classify the participants.

(b) Section 2: Dataset framework information. In questions Q2(a)-Q2(o), Q3(a)-Q3(o) and Q4(a)-Q4(o) we required researchers to apply definitions of dataset category elements to the data sets and to use formal definitions to identify quality issues in the data sets. We also asked an open question (Q5) to encourage participants to provide comments about our framework.

### 7.2.5 Tasks

The main task was for participants to answer all questions in the web-based survey, particularly the questions in Section 2: Dataset framework information. In Section 2, we provided participants with three different data sets for the same set of questions. We highlighted some elements of the data sets in bold face and labelled them using a set of answer options for each data set. In the questions we provided definitions for each dataset category element and asked participants to choose the correct answer options (as illustrated by the data sets). Some questions may have had more than one option to produce a correct answer.

In addition, we also asked participants to identify data quality issues in each given data set. We provided formal definitions for data quality issues in the answer options. As for the previous questions, the question about data quality issues also had more than one option to produce a correct answer.

### 7.3 Survey Execution

#### 7.3.1 Preparation

We applied to the UAHPEC (University of Auckland Human Participants Ethics Committee) for ethics approval to conduct the survey. We received approval from the ethics committee.
7.3. SURVEY EXECUTION

in October 2015 (Please refer to Appendix J.2 for our ethics application).

In our survey, we used three different methods for participant recruitment. First, we emailed invitations to participate to 125 researchers who we knew were involved in empirical software engineering research. The invitations contained an advertisement providing a web-based survey link for researchers to participate in the survey and contact information for the people conducting the study.

Second, we posted an advertisement to different mailing lists and groups chosen from Yahoo! and LinkedIn. We identified relevant groups, mainly in software engineering and designated by the titles software engineering research, software engineering professional, software quality engineering, software engineering process and productivity, data science, data analytics and big data software engineering. These groups were classified as both public and closed groups.

Third, we used a snowballing method to distribute the survey. We asked participants to forward the advertisement to other researchers they believed could be potential participants in the survey. Other researchers who were interested in participating in the survey were able to access the participant information sheet (PIS) and the survey through the forwarded link in the advertisement.

As mentioned earlier, the survey was designed to be anonymous and participants were not required to sign consent forms. However, we included a consent page as the first page of the web-based survey to inform participants about the survey. The consent page also informed participants about their rights to withdraw at any time before submitting the survey (please refer to Appendix I for the PIS and invitation email).

7.3.2 Procedure

Before publishing the survey we performed a pilot test by providing a link to the web-based survey to two potential participants via email. The pilot test was intended to identify any problems with the questions in the survey and to see how long it will take to complete the survey. We improved our survey based on early feedback from the participants.

We published the survey between October 2015 and April 2016. As mentioned earlier, we invited potential participants via email. We provided a link to the web survey in the advertisement in the email (see Figure 7.1 for a screen shot of part of the web-based survey).

The web-based survey started with the PIS and consent page. The main purpose of the PIS and consent page was to inform participants about the survey and ensure that they understood the conditions of taking part in the survey. As mentioned earlier, the survey was anonymous, so we did not ask the participants to sign their names or complete any information on the consent page. Participants could proceed to answer the survey if they chose to agree on the consent page.

The survey was divided into two sections:

1. Demographic information (six questions)
2. Dataset framework information

   a) Ant 1.7 data set (15 questions)
   b) KC2 data set (15 questions)
   c) An artificial data set (15 questions)
   d) An open-ended question (one question)

   The survey contained 52 questions in total and participants were not required to answer all of the questions. Some of the questions were linked based on specific conditions. We created a skip logic that allowed the survey to send participants to a future point in the survey if a specific condition was met.

   Participants progressed through the survey using the ‘Next’ and ‘Back’ buttons. When they reached the last page of the survey, they were asked to complete the survey by clicking the submit button. If participants did not click the submit button, they were considered to have withdrawn from participating in the research and their responses were not used in our analysis.
7.4 Results and Analysis

By the end of the survey period, we had received 49 responses in total. Out of the 49 responses, 16 completed only the first page of the survey —the participant information sheet —and 33 completed the first section —the demographic section. Of these 33 responses, only 14 clicked the submit button at the end of the survey.

Out of 14 participants, we found that one participant who purported to be a lecturer selected the ‘other’ category as the answer to the question that determine the category of participant (Q1(a)). This participant did not select the categories relating to research indicating that their position as a lecturer only involved teaching. Further, we observed this participant’s responses to other questions in the survey and found that s/he commented in the open ended question (Q5) of the data set framework section that the ‘survey must use layman terms that can be easily understood by the people who take the survey, then ... the surveyor must translate ... these to “so called” data set terms, and not the other way around’. We believed that this participant was not a researcher with a background in computer science or software engineering. Therefore, we removed this participant from the 14 participants because we considered that this participant was not in our target population. We accepted the 13 other participants as complete responses because they reached the final page of the survey and clicked the submit button to complete the survey.

Overall, the survey consisted of 52 questions. 11 out of 13 participants answered 23 out of 52 questions. As mentioned earlier, some questions were linked, based on specific conditions. We assumed that participants may have skipped some questions in the survey when these specific conditions were met.

Out of a total of 13 participants, ten participants did not answer the open-ended question (Q1(d)) in the demographic section or the open-ended question (Q5) in the data set framework section. Q1(d) asked participants to provide comments on quality issues in the data sets. Q5 asked participants to provide comments on the data quality framework.

The following sections present the results gathered from the 13 participants in the survey. We structured the results based on the categories of questions in the survey. We present the results beginning with the demographics section and followed by the results for the dataset framework section. We describe the comments from the participants about the data quality framework at the end of this section.

7.4.1 Demographics

The demographics section contained a total of six questions asking about the backgrounds of participants. The answers to these questions provided useful information for understanding whether the participants had experience with data sets, particularly data sets from data repositories.

Out of the 13 complete responses, eight participants were PhD students and five participants were academic researchers. Figure 7.2 illustrates the number of complete responses for each category of participant.
Three participants answered ‘yes’ to the question, ‘Have you used or analysed data sets from software repositories?’ and ten participants that answered ‘no’ to the same question. Three out of 13 participants answered the next question asking them to select the data repositories that contained the data sets they had used in research. This question allowed participants to select more than one data repository.

The three participants indicated that they used data sets from PROMISE (two responses), Eclipse Bug data (two responses), SIR (one response) and ‘other’ (one response). The ‘other’ category was specified as ‘Google Chrome’ and ‘Bug Zilla Bugdata’. Figure 7.3 illustrates the number of responses indicating the use of data sets in each software repository. Participants who indicated more than one data repository are counted for each repository they selected; therefore, the sum of the participant responses is greater than three. There were two such participants.
In question Q1(e), participants were asked about quality issues that they had encountered with data sets. This question allowed participants to select more than one quality issue. Ten participants reported that they encountered ‘missing data’, eight ‘inconsistent data’, eight ‘duplicate data’ and one ‘none’, indicating that s/he had encountered no quality issues.

Figure 7.4 illustrates the number of encounters with data quality issues in data sets. Participants that indicated more than one data quality issue are counted for each data quality issue they selected; therefore, the sum of the participant responses is greater than 13. There were nine such participants.

We asked participants to give their comments on quality issues in data sets in question 1(e). Three participants answered this question are academic researchers. Their comments are shown in Table 7.1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Precisely what the data measures is often not clear. The context of the work and recording can lead to similar work producing different data and the easily recorded data artefacts often do not answer the interesting questions. For example bug reports may not indicate specifically what and why something was changed. An ODC requires the contextual knowledge of the developer but is seldom included unless it is part of the fix protocol. The actual effort is seldom provided, we might get a calendar duration open, but this has little correlation to actual effort. Unless check-ins are highly atomic, the size of the fix and number modules touched can be ambiguous.</td>
</tr>
<tr>
<td>2.</td>
<td>Data entered by humans are prone to error and resulted in poor quality.</td>
</tr>
<tr>
<td>3.</td>
<td>Some of the data are not properly presented.</td>
</tr>
</tbody>
</table>

In the last question of the demographic section, participants were asked to select the programming languages that they knew from the following list: Java, C++, C#,
7.4. RESULTS AND ANALYSIS

Python, PHP, C and ‘other’. This question allowed participants to select more than one programming language.

We found more than half of total responses from participants had knowledge of Java (12 responses), followed by C++ (ten responses) and C (nine responses). Other responses to this question also included C# (six responses), Python (four responses), PHP (five responses) and ‘other’ (three responses). Figure 7.5 illustrates the number of responses for each programming language. Participants who indicated more than one programming language are counted for each programming language they selected; therefore, the sum of the participant responses is greater than 13. There were 12 such participants.

![Figure 7.5: Number of responses in each programming language.](image)

7.4.2 Dataset framework information

The second part of the survey was to answer 45 questions relating to part of our data quality framework. We presented a summary of our framework at the beginning of the second part of the survey. We then asked participants 15 questions per data set. Each data set was presented in a different format. The formats were comma separated values (CSV), attribute-relation file format (ARFF) and tabular. The following subsections describe the results for the second part of the survey.

7.4.2.1 Participant experiences analysing data sets

The first question for each data set related to how many times participants had analysed data sets in CSV, ARFF or tabular. The purpose of this question was to determine the experience of participants in analysing data sets. As described in the demographics section, two categories of participants answered our survey: academic researchers and PhD students. We analysed participant responses to this question by classifying responses into two categories of participants for each data set. The results are illustrated in Figure 7.6.

139
More than half of total participants reported that they had experience in analysing the CSV data set and the tabular data set. Out of 13 participants, 11 participants had analysed the CSV data set, seven participants had analysed the tabular data set and six participants had analysed the ARFF data set. It is possible that some of the participants in our survey were not familiar with the ARFF data set.

From Figure 7.6, we can see that nine participants had experience in analysing the CSV and tabular data sets more than 20 times — five out of 11 participants for CSV data sets and four out of seven participants for tabular data sets. In terms of categories of participants, we found that three participants were PhD students and two participants were academic researchers for the CSV data sets, whereas three participants were PhD students and one participant was an academic researcher for the tabular data sets. This indicates that participants with experience analysing data sets more than 20 times were from both categories of participants (PhD students and academic researchers), particularly in the CSV data set.

Regarding the number of times participants had analysed data sets, we found that more than half of total participants had experience in analysing the data sets from one to five times and more than 20 times. In terms of the participants who had experience in analysing data sets from one to five times, it can be seen in Figure 7.6 that three out of 11 participants used the CSV data set, six participants used the ARFF data set, and two out of seven participants used the tabular data set. As mentioned earlier, participants who had analysed data sets more than 20 times are five out of 11 participants for CSV data sets and four out of seven participants for tabular data sets. The percentage of total responses (39 responses) from participants who had analysed data sets from one to five times (28.2%) was more than the percentage of total responses (39 responses) from participants who had analysed data sets more than 20 times (23.1%). This shows that more than half of total participants with experience analysing data sets from one to five times in the survey.

Overall, the results suggest that both categories of participants had experience in analysing the three data sets, particularly in the CSV data sets and followed by the tabular data sets. It seems reasonable to assume that academic researchers do have experience...
and knowledge in research and some PhD students may have had experience analysing
data sets at least once during their doctoral research. It is speculated that there was a
relation between PhD students and academic researchers relative to their experience in
analysing data sets. Thus, we analyse participant responses in the following sections by
classifying them into PhD students and academic researchers to explore how background
influenced participant responses in the survey.

7.4.2.2 Participant responses: Applying the definitions of data set category
elements according to participant backgrounds

The next ten questions in each data set were related to the definitions of data set category
elements. Each data set consisted of five questions about the application of the definitions
of category elements to the data set and another five questions about the time taken to
identify category elements in the data set.

To answer SQ1 in Section 7.2 as to whether the definitions of dataset category ele-
ments in the dataset metamodel were effective in describing the structure and content
of data sets, we focused on participant responses to the questions (five questions for
each data set) that asked participants to apply definitions of dataset category elements to
each data set (Q2(b, d, f, h, j), Q3(b, d, f, h, j) and Q4(b, d, f, h, j)). We analysed correct
and wrong answers to identify the dataset category elements of three different data sets
grouped by participant background. We calculated the percentage of participants re-
sponses for correct and wrong answers for the three data sets. This analysis allowed us to
explore the relationship between participant background and the abilities of participants
to correctly identify dataset category elements in the data sets. We presented the results
from the dataset category elements in Figures 7.7 to 7.11.

Figure 7.7 shows the responses for record identifiers across the three data sets
(tabular, CSV and ARFF) according to participant background. Out of the 39 responses,
the percentage of responses containing correct answers was 36%: 13% (five of 39)
responses from the CSV data set, 13% (five of 39) responses from the tabular data set
and 10% (four of 39) responses from the ARFF data set.

In terms of participant background, we calculated the percentage of correct and
incorrect answers from the total responses of PhD students (24) and the total responses
of academic researchers (15) for the three data sets. The percentage of correct responses
from PhD students was 25% and the percentage of correct responses from academic
researchers was 53%. This shows that more than 50% of total responses from academic
researchers have responded with correct answers. It is likely that the academic researchers
who participated in our survey had more experience in research than the PhD students.

Figure 7.8 shows the responses for entity labels across the three data sets according to
participant background. Of the 39 responses who answered the questions, the percentage
of responses for correct answers was 36%. In particular, the percentage of correct answers
for the tabular data set was 15.4% (six of 39), the percentage for the CSV data set was
12.8% (five of 39) and the percentage for the ARFF data set was 7.7% (three of 39).
7.4. RESULTS AND ANALYSIS

In terms of participant background, 60% of total 15 responses from academic researchers and 20.8% of total 24 responses from PhD students had responded with correct answers for entity label. Similar to the results for record identifiers, more than 50% of total responses from academic researchers appear to have had successfully identified the entity label because they had more experience in research than the PhD students.

As can be seen in Figure 7.7 and Figure 7.8, 7 out of 8 responses from PhD students had responded with incorrect answer for record identifier and entity label in the ARFF data set. This indicates that most PhD students could not identify correctly the record identifier and entity label in the ARFF data set. It could be that they may not familiar with the format of ARFF data sets and have misunderstood the definition (for record identifier and entity label) because they had less experience in research.

Figure 7.9 shows the responses for metric labels in the three data sets according to participant background. Out of the 39 responses, the percentage of responses with correct answers was 56.4%. The percentage of correct answers for the tabular data set was 20.5% (eight of 39), the percentage for the CSV data set was 20.5% (eight of 39)
and the percentage for the ARFF data set was 15.4% (six of 39).

In terms of participants background, 80% of total 15 responses from academic researchers and 42% of total 24 responses from PhD students had responded with correct answers for metric labels. It is likely that most of the academic researchers were able to identify the metric labels correctly across the different formats of data sets because they may have been familiar with their use in research.

Figure 7.9: Number of responses for metric labels according to participant background.

Figure 7.9 shows that the ARFF data set (18%) received more incorrect responses than the other data sets. This could be explained by the fact that the structure of the ARFF data set is different from the structures of the other data sets and this may affect the abilities of participants — particularly PhD students with less experience in research — to identify the metric label correctly.

Figure 7.10 shows the responses for measurement values in the three data sets according to participant background. Out of the 39 responses, the percentage of correct responses was 74.4%. Both the ARFF and CSV data sets received 25.6% (10 of 39) of correct responses and the CSV data set received 23.1% (9 of 39) of correct responses.

In terms of participant background, 66.7% of total 24 responses from PhD students and 86.7% of total 15 responses from academic researchers had responded with correct answers for measurement values. This shows that more than 50% of total responses from participants were able to apply the definition for the measurement value correctly to the data sets. It may be that they had experience using measurement values in data sets.

From Figure 7.10, it can be seen that 33.3% responses of the total 24 responses from PhD students for identifying measurement value were incorrect: the ARFF and CSV data sets each received 12.5% incorrect responses and the tabular data set received 8.3% incorrect responses. This result seems reasonable because it is likely that the PhD students had less experience than the academic researchers and this could be the reason that they were sometimes unable to identify the measurement value correctly.

Figure 7.11 shows the responses for metric metadata in the three data sets according to participant background. Out of the 39 responses, the percentage of correct answers for metric metadata was 48.7%. In particular, the tabular data set and the ARFF data set each received 17.9% (7 of 39 each data set) correct responses, while the CSV data set received only 12.8% (five of 39) correct responses. One reason for this could be that...
the content of the CSV data set in the survey did not contain metric metadata; however, participants were able to answer this question correctly by selecting the option ‘none’.

Regarding the results of participant background, 73.3% of the total 15 responses from academic researchers and 33.3% of the total 24 responses from PhD students had responded with correct answers. It can be seen that four participants who answered correctly in both the tabular data set and the ARFF data set (26.7% for each dataset) and three participants who responded correctly in the CSV data set (20%) were academic researchers. This shows that most of the academic researchers identified the metric metadata correctly. It is speculated that the participants who were academic researchers might have had more experience analysing data sets and therefore had previously used the metric metadata in research.

Figure 7.10: Number of responses for measurement value according to participant background.

Figure 7.11: Number of responses for metric metadata according to participant background.
7.4.2.3 Participant responses: Providing correct answers for the definitions of dataset category elements considered according to participant background and experience analysing data sets

In this subsection, we focus on the correct answers to all questions related to the definitions of data set category elements. As mentioned in the previous section, each data set contained five questions that were related to the definitions of data set category elements. In total, there were 15 questions for the three data sets. To present the results, we counted the frequency of correct answers to the 15 questions from the 13 participants. We analysed the responses for correct answers based on participant categories to determine the relationship between background and the ability to answer all questions pertaining to the definitions of data set category elements in the data sets. The results are illustrated in Figure 7.12 (see Appendix K for detailed results).

In addition to determining the relevance of participant experience on the analysing of data sets, we also looked at the relationship between participant experience in analysing data sets and the total percentage of correct answers for the definitions of dataset category elements across the three data sets. The experience of each participant in analysing data sets is represented in Figure 7.13.

![Figure 7.12: Percentages of correct answers for definitions of elements in the data sets.](image)

We had expected that the PhD students would be able to answer these 15 questions as same with the academic researchers because the results depicted in Figure 7.6 show that the PhD students and the academic researchers had experience analysing data sets. However, the results in Figure 7.12 indicate that, for the most part, the academic researchers were able to answer the 15 questions more correctly than the PhD students. Of the 50% of the total participants who answered the 15 questions correctly across the three data sets, 27% were academic researchers and 23% were PhD students.

One participant who was a PhD student (P7) answered 100% of the questions correctly. This may have been because of the experience of P7 in analysing data sets as shown in Figure 7.13. It can be seen that P7 had previously used or analysed data sets
7.4. RESULTS AND ANALYSIS

Figure 7.13: Participant experience analysing data sets in data repositories.

From public data repositories—particularly data sets from PROMISE and Eclipse Bug Data. We also found that P7 had encountered two data quality issues in the data sets: duplicate data and missing data. It could be assumed that P7 was actively involved in research based on data sets, particularly data sets from public repositories.

In Figure 7.13, we can see that one participant who was an academic researcher (P1) did not have experience analysing data sets from public data repositories, but, nevertheless, as shown in Figure 7.12, s/he was able to answer the 15 questions 93.33% correctly. This contrasts with some of the academic researchers with experience analysing data sets (P4, P3 and P12) who were able to answer the 15 questions more than 70% correctly. It could be that P1 had a strong background of knowledge and experience in research that assisted him or her to understand and apply the definitions of data set category elements correctly in the survey.

It is apparent from Figure 7.12 that one PhD student (P10) answered all questions incorrectly across the three data sets. Although we found evidence (see Figure 7.13) that P10 had experience analysing CSV data sets and ARFF data sets between one and five times and had analysed tabular data sets more than 20 times, it could be assumed that these data sets were not from public data repositories. Figure 7.13 shows that P10 did not have experience analysing data sets from public data repositories (see Question 1(b)). We can assume that this participant did not form part of our target population for this survey. However, there were some participants who did not have experience analysing data sets from public data repositories and were able to answer some of the same questions correctly. This indicates that, although the participants did not have experience analysing data sets from public data repositories, they had general knowledge and experience in research that may have helped them to answer the questions in the survey.
Further, one academic researcher (P9) did not have experience analysing data sets from public data repositories; however, P9 was able to encounter a data quality issue: inconsistent data. P9 answered 13.33% correctly out of the 15 questions, as shown in Figure 7.12. This indicates that P9 may have some knowledge and experience in research.

### 7.4.2.4 Participant responses: Applying the definitions of data set category elements grouped by participant experience in analysing particular format of data sets

In this section, we focused on participant responses to five questions that asked participants to apply the definitions of dataset category elements to each data set —Q2 (b, d, f, h, j), Q3 (b, d, f, h, j) and Q4 (b, d, f, h, j). We compared these with responses to the questions, ‘Approximately how many times have you analysed CSV/ARFF/tabular data sets?’—Q2(a), Q3(a) and Q4(a), respectively. We considered correct participant responses alongside participant experience in analysing the three different formats of data sets. This analysis allowed us to explore the relationship between the way participant experience in a particular format of data set and participant ability to identify dataset category elements correctly. We present the distribution of correct answers applying the definitions of dataset category elements across the three data sets for each participant from Figures 7.14 to 7.16.

**Figure 7.14: Distribution of correct answers applying the definitions of dataset category elements to the CSV data set for each participant**

As shown in Figure 7.14, two out of 13 participants were able to identify the dataset category elements correctly in the CSV data set. One of the five participants who had analysed CSV data sets more than 20 times was an academic researcher and one of the two participants who had analysed CSV data sets between 11 and 20 times was a PhD student. Although only a few participants answered the five questions correctly, it is speculated that participant who had analysed CSV data sets more than 11 times may be able to apply the definitions of dataset category elements correctly to the CSV data set.
We expected that no participants would be able to answer the five questions correctly because most of participants had experience analysing ARFF data sets between one and five times (see Figure 7.15). However, we found that three participants had answered all five questions correctly for the ARFF data set. Two of the three participants who did not have experience analysing ARFF data sets were academic researchers and one participant who had analysed ARFF data sets between one and five times was a PhD student. Although the participants did not seem to have enough experience in analysing ARFF data sets, three participants were able to answer all the questions correctly. These participants may have had general knowledge and experience in research that allowed them to understand and apply the definitions of data set category elements correctly to the ARFF data set.

Figure 7.15: Distribution of correct answers applying the definitions of dataset category elements to the ARFF data set for each participant.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Q3(a) Approximately how many times have you analysed the ARFF data sets?</th>
<th>Q3(b) Record identifier</th>
<th>Q3(d) Entity label</th>
<th>Q3(f) Metric label</th>
<th>Q3(h) Measurement value</th>
<th>Q3(j) Metric metadata</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 (AR)</td>
<td>None</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P2 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>P3 (AR)</td>
<td>None</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>P4 (AR)</td>
<td>1-5 times</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>P5 (PS)</td>
<td>1-5 times</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>P6 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>P7 (PS)</td>
<td>1-5 times</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P8 (PS)</td>
<td>1-5 times</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P9 (AR)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P10 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>P11 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P12 (AR)</td>
<td>None</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P13 (PS)</td>
<td>1-5 times</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 7.16: Distribution of correct answers applying the definitions of dataset category elements to the tabular data set for each participant.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Q4(a) Approximately how many times have you analysed the tabular data sets?</th>
<th>Q4(b) Record identifier</th>
<th>Q4(d) Entity label</th>
<th>Q4(f) Metric label</th>
<th>Q4(h) Measurement value</th>
<th>Q4(j) Metric metadata</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 (AR)</td>
<td>None</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P2 (PS)</td>
<td>11-20 times</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>P3 (AR)</td>
<td>More than 20 times</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>P4 (AR)</td>
<td>1-5 times</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P5 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>P6 (PS)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P7 (PS)</td>
<td>None</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P8 (PS)</td>
<td>More than 20 times</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>P9 (AR)</td>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P10 (PS)</td>
<td>More than 20 times</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P11 (PS)</td>
<td>More than 20 times</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P12 (AR)</td>
<td>1-5 times</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>P13 (PS)</td>
<td>None</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: PS = PhD student, AR = Academic researcher
Five participants answered all five questions correctly in the tabular data set (see Figure 7.16). Out of these five participants, one participant who had analysed tabular data sets more than 20 times was a PhD student, two participants who had analysed tabular data sets between one and five times were academic researchers and two participants who did not have experience analysing tabular data sets were a PhD student and an academic researcher. This indicates that, although some participants did not have enough experience analysing tabular data sets, they were still able to answer the five questions correctly. Just as for the ARFF data set, it could be speculated that these participants were able to answer the five questions correctly because of their background knowledge and experience in research.

**7.4.2.5 Participant responses: Identifying data set category elements in the data sets within three time periods**

In Section 7.4.2.2, we mentioned that the participants were required to answer five questions to determine the time taken to identify the dataset category elements. These questions —Q2 (c, e, g, i, k), Q3 (c, e, g, i, k) and Q4 (c, e, g, i, k)— were designed to answer the secondary research question for SQ1: whether or not the participants perceived the definitions of dataset category elements as easy to use. We asked participants to answer these kinds of questions right after each question identifying the dataset category element in the survey. To answer the question related to the time taken to identify the dataset category element, participants were required to choose from three options: (a) Short (e.g., 1–10 seconds), (b) Medium (e.g. 11–30 seconds) and (c) Long (e.g., more than 30 seconds).

For this section, we analysed participant responses when identifying dataset category elements across three different data sets in terms of these three periods of time and in relation to participant backgrounds. We calculated the percentage of total responses for correct and incorrect answers across the three data sets for the three durations. The results of this analysis allowed us to explore the relationship between participant backgrounds and the ability of participants to identify dataset category elements in each data set within the three durations. We present the results for the three durations in Figures 7.17 to 7.21.

Figure 7.17 shows the responses identifying record identifiers within three periods of time across three data sets. Out of the 39 responses, the percentage of correct answers in a short duration for the three data sets was 25.6%. In particular, three academic researchers and two PhD students answered correctly for the tabular data set (12.8%), one academic researcher and two PhD students answered correctly for the CSV data set (7.7%) and one academic researcher and one PhD student answered correctly for the ARFF data set (5.1%). This indicates that participants were able to understand and apply the definition of the record identifier within a short duration, especially in the tabular data set. We speculated that these participants might have had some experience using data sets in the tabular format.

In terms of wrong answers given in a short period of time, the percentage for the
three data sets was 20.5%. In particular, 8% of total participants answered incorrectly for both the tabular data set and ARFF data set, and 5% of total participants answered incorrectly for the CSV data set. It could be that these participants answered the questions in a very short time because they had limited time to complete the survey.

We also found many incorrect responses in the three data sets (the percentage of incorrect responses was 38% of total 39 responses) that were made in longer time periods. Four PhD students and one academic researcher answered slowly but incorrectly for the tabular data set (13%), five PhD students and one academic researcher answered slowly but incorrectly for the ARFF data set (15%) and three PhD students and one academic researcher took their time with the CSV data set (10%), but answered wrongly. Some of the PhD students may have taken a longer time to answer the questions because they did not have experience using data sets in research.

![Figure 7.17: The number of responses identifying record identifiers in three periods of time.](image)

Figure 7.18 shows the responses identifying entity labels within three durations of time across the three data sets. There was a difference in participant background based on the formats of data sets when identifying the entity label in a short period of time. In a short period of time, out of the 39 responses, the percentage of correct answers for the three data sets (23%) was higher than the percentage of incorrect answers for the three data sets (21%).

In terms of participants background, four academic researchers and two PhD students obtained correct answers in the tabular data set, one academic researcher and one PhD student obtained correct answers in the ARFF data set and one academic researcher obtained correct answers in the CSV data set. This shows that most of academic researchers were able to quickly apply the definition to the entity label in the tabular data set.

Of the 24 total responses from PhD students, the percentage of responses provided wrong answers within a long period of time across the three data sets was 50%: 25% (six of 24) responses from the tabular data set, 12.5% (three of 24) responses from the ARFF data set and 12.5% (three of 24) responses from the CSV data set. Similar to the previous
7.4. RESULTS AND ANALYSIS

Results revealing slow and erroneous answers for record identifiers, these results show that many responses from the PhD students were incorrect and that these students took a longer time to identify entity labels. It could be that they were slower to understand the definition of the entity label because they had less knowledge of the research process.

![Figure 7.18: The number of responses identifying entity labels in three periods of time.](image)

Figure 7.18 presents the breakdown of responses with correct and incorrect answers identifying the metric label across three data sets given in three different time periods. Of the 39 responses, the percentage of responses for correct answers within a short duration of time across the three data sets (33%) was higher than the average percentage of responses for incorrect answers within a short duration of time across the three data sets (2.6%). Seven responses from academic researchers (the percentage of correct responses for the three data sets from the total responses of academic researchers was 47%) and six responses from PhD students (the percentage of correct responses for the three data sets from the total responses of PhD students was 25%) identified the metric label correctly across the three different data sets in less than ten seconds. It can be speculated that the academic researchers may have more experience than the PhD students and that this allowed them to determine the metric label within a shorter duration of time.

From the three data sets, we found the percentage of responses with wrong answers from PhD students was 58% for identifying the metric label. In particular, 12 responses were made over a long period of time (50%), one response was provided over a medium period of time (4.2%) and one response was given quickly (4.2%). There were six responses from PhD students who could not identify the metric label in the ARFF data set. This could be because the position of the metric label in the ARFF data set was different from those in the other data sets and this made it difficult for participants to identify the metric label correctly.

Figure 7.20 shows responses providing correct and incorrect answers identifying measurement values within three periods of time across the three data sets. It is apparent from Figure 7.20 that many responses from participants for the three data sets are correct.
answers provided within a short period of time — in particular, six responses from the CSV data set, five responses from the ARFF data set and eight responses from the tabular data set. Of the 39 responses, the percentage of correct responses for the three data sets within a short period of time was 48.7%. We found that these correct responses were made by both categories of participant (60% from total responses (9 of 15) of academic researchers and 42% from total responses (10 of 24) of PhD students). These results appear to be reasonable because the measurement value is presented as a string or a number and could be easily identified in the data sets.

We can see in Figure 7.20, many responses from PhD students could not identify the measurement value over a long duration of time. Of the 39 responses, the percentage of incorrect responses from PhD students over a long duration of time was 29.2% and the percentage of incorrect responses from PhD students within a short duration of time was 4.2%. This indicates that some PhD students could not identify the measurement value within a long period of time, possibly because they had less experience in research based on data sets. Conversely, some PhD students may have accidentally identified the wrong element for the measurement value because they attempted to answer the question in a very short time.

Figure 7.21 shows responses with correct and incorrect answers identifying metric metadata in three different periods of time across three data sets. Out of the 39 responses, the percentage of responses with correct answers (5.1%) given within a short period of time was less than the percentage of responses with incorrect answers (12.8%) provided within the same period of time. In particular, one PhD student had obtained correct answers within a short period of time in both the tabular and ARFF data sets. This PhD student may have been able to identify the metric metadata within a short period of time because the metric metadata was explicitly described in the tabular and ARFF data sets.

The results for incorrect answers indicate that many wrong responses from both categories of participants were completed within a long period of time than a short
7.4. RESULTS AND ANALYSIS

Figure 7.20: The number of responses identifying measurement values in three periods of time.

duration of time. In the short period, it can be seen that two wrong responses in the tabular data set, one wrong response in the ARFF data set and two wrong responses in the CSV data set. These participants may not have had enough experience dealing with data sets and may have attempted the questions quickly because they had limited time to answer the survey. In terms of the long period of time, we can see that four wrong responses in the tabular data set, five wrong responses in the ARFF data set and six wrong responses in the CSV data set. The percentage of wrong responses in a long duration of time for the three data sets from participants was 38.5%. It can be assumed that these participants had no experience using metric metadata in the data sets because many existing data sets in public data repositories do not contain metric metadata.

Figure 7.21: The number of responses identifying metric metadata in three periods of time.
7.4.2.6 Participant responses to identifying entities in the data sets

In the second part of the survey, we asked participants two questions related to the entities in each data set. The first question asked participants to identify the entities in the data sets and the second question asked participants to specify how they identified the entities in the data sets. We analysed responses from these two questions to investigate the relationship between participant backgrounds and their abilities to identify the entities correctly in the data sets. Participant responses to the first question identifying the entities in each data set are illustrated by Figure 7.22.

![Figure 7.22: Participant responses identifying the entities in data sets.](image)

The overall results, as shown in Figure 7.22, indicate that 30.8% of the total 39 responses from participants identified the entities correctly across the three data sets. Out of 39 responses, six responses from participants identified the entities correctly in the tabular data set, three responses from participants identified the entities correctly in the ARFF data set and three responses from participants identified the entities correctly in the CSV data set. In particular of tabular data set, four of five academic researchers and two of eight PhD students provided correct answers. This could be because some academic researchers might have had knowledge of and experience in determining the entities of data sets in tabular format.

In terms of incorrect answers, 33.3% of the total 39 responses from participants could not identify the entities across the three data sets. One of the five academic researchers and five of the eight PhD students misidentified the entities of the tabular data set, four PhD students misidentified the entities of the ARFF data set and two academic researchers and one PhD student provided wrong answers for the CSV.

In addition, 35.9% of the total 39 responses from participants provided a ‘don’t know’ answer for the three data sets. Figure 7.22 shows that most of PhD students provided a ‘don’t know’ answer at some stage, especially for the CSV data set (four of eight participants) and the ARFF data set (three of eight participants). Some PhD students may not have been able to determine entities clearly in the data sets because they did not know the correct way to represent entities in the data sets.
As described earlier in Section 7.4.1, more than half of total participants had knowledge of programming languages, particularly Java, C++ and C, which indicates that they were familiar with the common conventions for source codes. However, as demonstrated in Figure 7.22, few participants were able to identify entities correctly in the data sets although the entities were represented using the common conventions for source codes. We found that three out of 12 responses from participants familiar with Java correctly identified the entities in the CSV data set and four out of nine responses from participants familiar with C++ correctly identified the entities in the tabular data set. It can be speculated that knowledge of programming may not have been enough to help the participants in our survey to identify the entities correctly in the data sets.

As mentioned earlier, the second question related to the entities in the data sets was ‘How did you identify the entities in Figure ....?’. We provided four possible answers to this question: (a) ‘I know this data set’, (b) ‘it is stated in the data set’, (c) ‘I made a reasonable guess based on my experience’ and (d) ‘other’. This question aimed to determine the knowledge of participants when identifying the entities in the data sets. Participant responses about how they identified the entities in the data sets are shown in Figure 7.23.

Figure 7.23: Participant responses about how they identified the entities in the data sets.

Overall, the results indicate that 64.1% of the total participants answered the second question explaining the reasons for the first question; that is, identifying the entities in the data sets. As can be seen in Figure 7.23, 48.7% of responses from participants selected (c) ‘I made a reasonable guess based on my experience’, 7.7% selected (b) ‘it is stated in the data set’ and 7.7% selected (d) ‘other’. No participants selected (a) ‘I know this data set’.

As Figure 7.23 shows, more than half of total responses from academic researchers and PhD students selected (c) ‘I made a reasonable guess based on my experience’ as the answer to this question; these participants included four PhD students and four academic researchers in the tabular data set, four PhD students and two academic researchers in
the ARFF data set and three PhD students and two academic researchers in the CSV data set. More than half of total responses with answer (c) were from the tabular data set. It can be assumed that these participants were confident in determining entities in the tabular data set because they had experience in analysing tabular data sets.

We analysed the results in Figure 7.22 for correct answers and compared them with the results in Figure 7.23 where participants had chosen (c) ‘I made a reasonable guess based on my experience’. We found four academic researchers in the tabular data set, one PhD student and two academic researchers in the ARFF data set and two PhD students in the CSV data set who had responded with correct answers and selected the option (c) ‘I made a reasonable guess based on my experience’. The percentage of responses for correct answers and selected the option (c) ‘I made a reasonable guess based on my experience’ from the total responses of academic researchers (6 of 15) was 40% and the percentage of responses for correct answers and selected the option (c) ‘I made a reasonable guess based on my experience’ from the total responses of PhD students (3 of 24) was 13%. Therefore, more than half of total academic researchers successfully identified the entities in the data sets based on their own experience in research.

7.4.2.7 Participant responses identifying data quality issues

As discussed earlier, we provided three data sets, each with a set of the same questions in the second part of the survey. The three data sets presented in this survey contained data quality issues. We asked the participants to identify the quality issues for each data set using formal definitions for data quality issues that had been constructed by the quality assessment process in Chapter 6. This question aimed to answer SQ2 as to whether the participants were able to apply formal definitions to data quality issues in the data sets. We analysed the responses to this question to explore the relationship between participant background and experience, and the ability to identify the data quality issues correctly across the three data sets. Participant responses with correct, incorrect and ‘don’t know’ answers for each data set are illustrated in Figure 7.24.

![Figure 7.24: Participant responses providing formal definitions of data quality issues for each data set.](image-url)
The results shown in Figure 7.24 indicate that 46.2% of the total responses from participants answered correctly, 38.5% of the total responses from participants answered incorrectly and 15.4% answered by selecting ‘don’t know’. There were five of 13 participants (two PhD students and three academic researchers) who answered correctly for the CSV data set, six of 13 participants (three PhD students and three academic researchers) who answered correctly for the ARFF data set, and seven of 13 participants (five PhD students and two academic researchers) who answered correctly for the tabular data set. This indicates that more participants successfully identified the data quality issues in the tabular data set than in the CSV and ARFF data sets.

In terms of participant background, of the total 24 responses from PhD students, the percentage of correct responses for the three data sets was 41.7%. Out of the total 15 responses from academic researchers, the percentage of correct responses for the three data sets was 53.3%. In particular, the responses from PhD students for the tabular data set were more accurate than those for the CSV and ARFF data sets. This is consistent with the reported participant experience in analysing data sets shown in Figure 7.6, which indicated that four PhD students had experience in analysing tabular data sets. This suggests that participants’ experience in analysing data sets might help them to identify data quality issues correctly in data sets.

The results shown in Figure 7.24 indicating that some participants successfully identified the data quality issues accord with the results in Figure 7.22 indicating that some participants identified the entities correctly in the same data sets. This is because participants presumably knew the entities in a particular data set before identifying the data quality issues in that particular data set. For example, three PhD students identified the entities correctly in the CSV data set (see Figure 7.22) and two PhD students successfully identified the data quality issues in the same data set. In this example, the remaining PhD student identified the entities correctly in the CSV data set but responded with a wrong answer for the data quality issues in the same data set.

7.4.2.8 Participant responses identifying data quality issues in the data sets within three periods of time

In the second part of the survey in Q2(o), Q3(o) and Q4(o), we asked participants, ‘How long does it take to identify the quality issues in Figure ..... ?’. This question aimed to determine the time taken to identify quality issues in the data sets. We provided three possible answers to these questions: (a) ‘short’ (e.g., 1-10 seconds), (b) ‘medium’ (e.g., 11-30 seconds) and (c) ‘long’ (e.g., more than 30 seconds). We analysed the participant responses for the three periods of time when identifying the data quality issues in three different data sets in relation to each participant's background. This analysis allowed us to explore the relationship between participant backgrounds and the ability of participants to identify the data quality issues correctly within the three periods of time. Figure 7.25 presents the results of participant responses identifying data quality issues in the three periods of time.

The results in Figure 7.25 show that 85% of the total responses from participants
7.4. RESULTS AND ANALYSIS

Figure 7.25: Participant responses identifying data quality issues in three periods of time.

answered the questions determining the time taken to identify the quality issues in the data sets. 46% of the total responses from participants responded with correct answers and 39% of the total responses from participants responded with incorrect answers across the three durations of time. 15% of participants did not answer the questions about duration because they had answered ‘none’ or ‘don’t know’ to the previous question identifying the quality issues in the particular data set.

Comparing the responses for correct answers given in a short period of time, it can be seen that the tabular data set received the highest number of correct responses, followed by the ARFF data set and then by the CSV data set. For the 13 participants in the tabular data set, two correct responses were from academic researchers and three correct responses were from PhD students, whereas, in the ARFF data set, two correct responses were from PhD students and one correct response was from an academic researcher. For the CSV data set, two correct responses were from academic researchers and one correct response was from a PhD student.

In terms of the participants background, out of the 24 total PhD students, the percentage of correct responses for the three data sets in a short period of time was 25%. Of the 15 academic researchers, the percentage of correct responses for the three data sets in a short period of time was 33.3%. This indicates that more than half of total academic researchers answered the questions to identify the quality issues quickly.

In terms of responses with wrong answers, the tabular data set and the CSV data set received more incorrect responses than the ARFF data set. Five PhD students responded with wrong answers within a long period of time in the CSV data set. In addition, the tabular data set received one response from an academic researcher and one response from a PhD student within long periods of time and the ARFF data set received one response from an academic researcher within a long time period. This indicates that more than half of total PhD students spent more than 30 seconds identifying the quality issues, particularly in the CSV data set. It can be speculated that some of the PhD students in this survey may have had less experience in determining data quality issues in certain
7.4.3 Open-ended question

In the last part of the survey, we provided participants with an open-ended question and with space for them to write comments about the data quality framework. Three of 13 participants answered the open-ended question. The issues that were raised by the three participants are depicted in Table 7.2. The first two answers indicated that the participants had issues understanding the definitions of data set category elements. The third answer indicated that the participants had issues understanding the representations of the data sets. We discuss these issues in section 7.5.5.

Table 7.2: Comments and feedback about the open-ended question.

<table>
<thead>
<tr>
<th>No.</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>It took some time to understand the meaning of some terms used.</td>
</tr>
<tr>
<td>2</td>
<td>I usually use statistical languages such as R and SAS to analysis CSV and EXCEL sheets data. It was a bit challenging for me to clearly understand your definition of identifier, I assumed you mean the primary key of the data.</td>
</tr>
<tr>
<td>3</td>
<td>The representation of data is clearer and easy to read for the third one and is so complicated on the first one. It is easier to find missing information on the third representation compare to those two earlier.</td>
</tr>
</tbody>
</table>

7.5 Discussion of the findings of our survey

The overall findings discussed here are based on the results of a user study conducted between October 2015 and April 2016. The user study was conducted with the aim of evaluating the effectiveness and usability of part of the framework for data quality assessment (as described in Section 7.2). We used an online survey as our user study method. In the survey, we had included questions that evaluated whether the participants understood the definitions of dataset category elements and were able to apply them to a range of data sets. We also included questions that evaluated whether the participants were able to use formal definitions of data quality issues to identify quality issues across the same range of data sets. We received 49 responses in total; however, we considered only 13 complete responses. Although the sample size was small, it complied with the general rule regarding sample sizes in usability evaluation [95].

The findings from the survey are indicative due to the small number of participants who responded to the survey. The survey results suggest mild support to a conclusion that is participants with background knowledge and experience in research —particularly in analysing data sets—were able to apply our definitions of dataset category elements to a range of data sets and to apply the formal definitions of data quality issues to the same range of data sets. This conclusion was reached by analysing the participant responses in the survey based on participant background; that is, whether they were academic researchers or PhD students. We also analysed participant responses based on
the participants’ experience in analysing or using data sets in their research. Table 7.3 and table 7.4 present a summary of the analyses and results as discussed in the previous subsections.

Table 7.3: Summary of the analyses and results of the application of definitions of dataset category elements to the data sets.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Summary of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Considered participant responses in terms of their experience in analysing data sets and according to participant background (see Section 7.4.2.1).</td>
<td>Both categories of participants (academic researchers and PhD students) had experience in analysing data sets, particularly in the CSV data sets.</td>
</tr>
<tr>
<td>Analysed the participants’ correct and incorrect answers according to participant background when participants applied the definitions of data set category elements to the data sets (see Section 7.4.2.2).</td>
<td>The percentage of correct responses was higher than the percentage of incorrect responses for metric label and measurement value. The percentage of incorrect responses was higher than the percentage of correct responses for record identifier, entity label and metric metadata.</td>
</tr>
<tr>
<td>Analysed participant responses for the percentage of correct answers in relation to participant background and experience in analysing data sets when participants applied the definitions of data set category elements to the data sets (see Section 7.4.2.3).</td>
<td>Four of the five academic researchers obtained a higher percentage of correct answers than six of the eight PhD students, although four of the five academic researchers did not have any experience in analysing data sets from data repositories.</td>
</tr>
<tr>
<td>Analysed the distribution of correct answers in applying the definitions of data set category elements when participants had experience in analysing a particular format of data set. (see Section 7.4.2.4).</td>
<td>Three of the participants had experience in analysing tabular data sets, two had experience in analysing CSV data sets and one participant had experience in analysing ARFF data sets; these participants answered all of the questions correctly in the particular data set in which they had experience.</td>
</tr>
<tr>
<td>Analysed participant responses in terms of the time taken to apply the definitions of dataset category elements to the data sets in relation to participant background. (see Section 7.4.2.5).</td>
<td>Both categories of participants (academic researchers and PhD students) quickly identified the four dataset category elements (record identifier, entity label, metric label and measurement value).</td>
</tr>
</tbody>
</table>

In the following subsections we summarise the aggregation of findings in terms of the definitions of the dataset category elements and the formal definitions for data quality issues.

7.5.1 Analysis of the application of the definitions of dataset category elements to the data sets

The first survey research question in this user study sought to determine the effectiveness of the definitions of the dataset category elements in the dataset metamodel. The survey results for the application of these definitions to the data sets revealed that most
### Table 7.4: Summary of the analyses and results of the application of the formal definitions of data quality issues to the data sets.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Summary of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysed participant responses in which the entities in the data sets were identified according to participant background (see Section 7.4.2.6).</td>
<td>The percentage of correct responses from total responses of academic researchers identifying the entities based on their own research experience was 40% and the percentage of correct responses from total responses of PhD students identifying the same entities was 13%.</td>
</tr>
<tr>
<td>Analysed participant responses identifying data quality issues in the data sets according to participant background (see Section 7.4.2.7).</td>
<td>The percentage of correct responses from total responses of academic researchers was 53.3% and the percentage of correct responses from total responses of PhD students was 41.7%.</td>
</tr>
<tr>
<td>Analysed participant responses in terms of the time taken to identify the data quality issues in the data sets according to participant background (see Section 7.4.2.8).</td>
<td>The percentage of correct responses within a short period from total responses of academic researchers was 33.3% and the percentage of correct responses within a short period from total responses of PhD students was 25%.</td>
</tr>
</tbody>
</table>

Participants were able to understand and apply the definitions of the dataset category elements by identifying these elements correctly across the three different data sets. This can be seen in Figures 7.7 to 7.11, which indicate that, 36% of the total responses from participants provided correct answers for the record identifier and entity label, 56.4% of the total responses from participants provided correct answers for the metric label, 74.4% of the total responses from participants provided correct answers for the measurement value and 48.7% of the total responses from participants provided correct answers for the metric metadata. More than 50% of total responses from participants thus provided correct answers for the measurement value and metric label. This result could be explained by the fact that the most common elements of data sets are measurement values and these elements are associated with metric labels. Some participants who have had experience in research may have used measurement values and metric labels frequently in data sets.

In contrast, we noticed that few participants were able to provide correct answers for record identifiers and entity labels across the three different data sets. The few participants who answered these correctly were academic researchers who may have had more experience in research than the PhD students. This indicates that participant background and experience in research play a role in the ability to understand and apply the definitions of dataset category elements to the data sets.

The analysis of the participants’ correct answers according to participant background also indicates that more academic researchers (27%) obtained correct answers than PhD students (23%). As can be seen in Figure 7.12, four of the five academic researchers who
responded to the survey answered 70% of the questions correctly; however, only one of these four academic researchers (P4) had experience analysing data sets from public data repositories, indicating that more than half of total academic researchers used their general knowledge and prior experience in research to apply the definitions of dataset category elements to the data sets.

As mentioned in the results (see Section 7.4.2.3), P7 — a PhD student — answered 100% of the questions correctly. We found that P7 had no experience analysing tabular data sets, but did have previous experience analysing CSV data sets (between 11 and 20 times) and ARFF data sets (between one and five times). This indicates that participant experience analysing data sets does help to apply the definitions of dataset category elements correctly to the data sets.

Comparing the analysis of participant responses based on their experience analysing data sets in Figures 7.14 to 7.16, we can see that participant experience with particular formats of data sets also has provided guidance to help the participants to apply the definitions of dataset category elements correctly to the data sets. This is evident from the results shown in Figure 7.14, where an academic researcher (who had analysed CSV data sets more than 20 times) and a PhD student (who had analysed CSV data sets between 11 and 20 times) responded correctly to all of the questions in the CSV data set.

However, we can also see from Figures 7.14 to 7.16 that four participants who did not have experience analysing data sets were able to answer all of the questions correctly for particular data set formats, namely, the ARFF data set and the tabular data set. Of these four participants, three were academic researchers and one was a PhD student. An explanation for this is that the academic researchers have prior experience in research that may help them to apply the definitions of the dataset category elements correctly to the data sets. The participants' general knowledge and prior experience in research may therefore also help them on the correct application of the definitions of dataset category elements to the data sets.

### 7.5.2 Analysis of the time spent identifying the dataset category elements

With respect to the first survey research question in this study, we found four dataset category elements that could easily be identified by participants across the three different data sets. The four dataset category elements are measurement values, metric labels, entity labels and record identifiers. The results indicate that more academic researchers responded with correct answers within short periods of time than PhD students for the four dataset category elements. This could be explained by the supposition that the academic researchers may have more general experience in research than the PhD students.

Comparing the analysis of participant responses identifying dataset category elements in three different time frames in Figures 7.17 to 7.21, we can see that the measurement value was likely to be easily identified within a short period of time because it was a common element in the data sets. It seems possible that the metric label was also able
to be easily identified within a short period of time because it was associated with the measurement value.

The results indicate that the number of responses from participants identifying the dataset category elements in a short period of time for measurement values, metric labels, entity labels and record identifiers in the tabular data set were higher than those in the CSV and the ARFF data sets. It could be that some participants were familiar with the structure of the tabular data set format. This made it easier for them to apply the definitions of dataset category elements correctly to the data sets.

In terms of responses for metric metadata, only two responses for correct answers were provided in a short period of time for the tabular and the ARFF data sets. This result may be explained by the fact that the academic researchers and PhD students were not familiar with the common conventions for metric metadata in data sets. It could be speculated that this happened because many data sets from public data repositories often do not contain metric metadata.

### 7.5.3 Analysis of the application of formal definitions of data quality issues to the data sets

The second survey research question was to determine how well participants were able to identify quality issues across the three data sets. As mentioned earlier, the results of analysing participant responses identifying data quality issues indicated that 46% of the total participants identified data quality issues correctly across the three data sets. We found that more than half of total participants responded with correct answers in the tabular data set. This could be because some participants in our survey were familiar with the formats of tabular data sets.

In terms of participant background, we found that the number of responses with correct answers from PhD students was higher than that of academic researchers for the tabular data set. This is consistent with the results shown in Figure 7.6 highlighting three responses from PhD students who had experience in analysing data sets more than 20 times. It was evident that these PhD students were familiar with the structure of the tabular format—an advantage that allowed them to identify the data quality issues correctly. Further, the tabular data set was arranged as the third data set in the structure of the survey. We believe that the participant experience through the structure of data sets while answering the survey does also help them to identify the data quality issues correctly.

In addition, we found that more academic researchers responded with correct answers than PhD students in the CSV data set. This result is likely to be related to the responses from academic researchers to Question 1(d) (see Figure 7.13), indicating that four out the five academic researchers had encountered some data quality issues in the data sets. We speculated that the academic researchers' experience with data quality issues in data sets allowed them to apply the formal definitions correctly to the data sets.
7.5.4 Analysis of time spent identifying data quality issues

With respect to the second survey research question in this study that is to assess the time spent identifying data quality issues in data sets, we looked at how long participants spent answering the relevant questions. The results (see Figure 7.25) show that the percentage for correct answers given in a short period of time from total responses of PhD students was 25% and the percentage for correct answers given in a short period of time from total responses of academic researchers was 33.3%. This result could be related to the responses from participants to Question 1(d), which indicated that eight PhD students and four of the five academic researchers had encountered some data quality issues in the data sets.

In terms of participant experience with a particular format of data set, the results for participants who responded within a short period of time for the tabular data set and the CSV data set (see Figure 7.25) are consistent with the results presented in Figure 7.6 indicating that most participants had experience in analysing the tabular data set and the CSV data set more than 20 times. As can be seen in Figure 7.6, more than half of total participants who had previously analysed both data sets (tabular and CSV) were PhD students, which implies that they were familiar with the tabular and the CSV data sets. However, the results in Figure 7.25 indicate that fewer PhD students responded with correct answers in a short period of time in the CSV data set. This suggests that, in some cases, the PhD students may have had difficulty identifying quality issues in a short period of time even though they were familiar with the formats of the data sets. It may be that the PhD students made mistakes because they attempted to answer the questions in a very short time.

Additionally, we found five responses from PhD students who responded with wrong answers within a long period of time for the CSV data set. It seems that these PhD students took more time to understand the formal definitions for data quality issues before choosing their answers. It is difficult to explain this result; however, it may be related to the participants’ abilities to identify the entities in the data sets. As discussed earlier, in the results for identifying the entities in Figure 7.22, we found that some PhD students could not identify the entities correctly in the data sets. This implies that there is a relation between participant ability to identify the entities in the data set and participant error in answers provided in a longer period of time. It can therefore be assumed that participants should identify the entities correctly to identify the data quality issues correctly and in a reasonable amount of time for each data set.

7.5.5 Analysis of the open-ended questions about the framework

We received three responses from participants to the open-ended questions about the framework. In general, the comments and feedback indicated that our definitions for dataset category elements were not completely clear or easy to understand. This result is likely attributable to the participants' background knowledge and experience in analysing the data sets. The participants who answered this question were an academic researcher and two PhD students. These participants reported that they did not have experience
7.6 IMPROVEMENT IN THE DATA QUALITY FRAMEWORK

analysing data sets from data repositories. This implies that presumably they were not familiar with the structure of data sets nor with the common conventions of data sets from data repositories.

Further, we found that the academic researcher and one of the PhD students did not have any experience in analysing data sets in CSV, ARFF or tabular formats. However, another PhD student had experience in analysing CSV data sets more than 20 times and in ARFF between one and five times. It seems plausible that this PhD student may have analysed other kinds of data sets such as data sets from databases.

Another possible explanation for this is that the participants might have had different perspectives for understanding the structure of data sets because they had already used other kinds of data sets. For example, one of the participants commented, ‘I usually use statistical languages such as R and SAS to analyse CSV and EXCEL sheets data. It was a bit challenging for me to clearly understand your definition of identifier. I assumed you mean the primary key of the data’. This comment could be interpreted as indicative of a good understanding on the part of the participant of the structure of data sets from databases because the participant was able to relate the definition of record identifier to the primary key used in databases. This suggests that some of the definitions of data set category elements can be generalised in certain situations.

7.6 Improvement in the data quality framework

Some of the results of our user study have indicated that there is a need to improve the definitions of dataset category elements in the dataset metamodel. The most obvious result is obtained from the participant comments responding to the open-ended question about the framework, which indicated that the definitions of dataset category elements were not easy and clear to understand (see Section 7.5.5). Although this result is likely to be related to participant backgrounds and levels of experience in research, we consider the participants’ comments useful as feedback for improving the definitions of dataset category elements.

Another result emerges from the five questions for applying the definitions of data set category elements to the three different data sets. The result showed that few participants were able to provide correct answers for record identifiers and entity labels across the three data sets. This result is also related to participant backgrounds and levels of experience in research; however, we are concerned that it could reveal a potential issue with the definitions for record identifiers and entity labels. This concern helps strengthen the need to improve the definitions of dataset category elements in the dataset metamodel.

We improved the definitions for dataset category elements in the dataset metamodel after the survey conducted in evaluation. In particular, we revised the definitions to capture a more precise meaning and to make them clearer and easier to understand. The new improved definitions for dataset category elements are described in Chapter 5. We conducted an observational study to evaluate the new definitions for dataset category
elements. (The original definitions for dataset category elements that were used in the survey can be found in Appendix J)

7.7 Threats to validity

In this section, we describe threats to the validity of the survey and the possible mitigation of these threats. We consider four types of validity as discussed by Wohlin et al. [94]: (a) conclusion validity, (b) internal validity, (c) construct validity and (d) external validity.

7.7.1 Conclusion validity

Conclusion validity is ‘concerned with the relationship between the treatment and the outcome’ [94]. The threats to this validity are related to issues such as low statistical power, violated assumptions of statistical tests, reliability of measures, reliability of treatment implementation and random heterogeneity of subjects. These issues could affect the ability to draw conclusions about the statistical relationship between outcome and treatment [94].

In our survey, we did not investigate the relationship between the definitions of dataset category elements and the data sets. Moreover, we did not apply any treatment to the survey results that might affect the conclusions made about our survey research questions. However, there is a possible threat regarding the random heterogeneity of participants in our survey. According to Wohlin et al. [94], if the subjects are very heterogeneous, the individual differences could have a larger effect on the outcome than the treatments.

In the survey, we focused on two types of participants, academic researchers and PhD students. The amount of research experience and knowledge that our participants had could affect how they answered the questions in the survey. For example, academic researchers have vast knowledge and experience in general research, while PhD students may have limited knowledge and experience in general research because they are still carrying out their very specific doctoral projects.

We collected data on the participants’ experience analysing data sets from public data repositories to observe the relationship between the participants’ experience analysing data sets and their ability to identify the dataset category elements correctly in the data sets. In the analysis of the results, we did not find that the difference in experience analysing data sets had a significant influence on the effectiveness of our definitions of dataset category elements. This was because we found that some participants (three of the five academic researchers), who did not have any experience analysing data sets, had successfully identified the dataset category elements. However, we believe these participants might have applied the definitions of dataset category elements based on their knowledge and experience in general research.
7.7.2 Internal validity

Internal validity is 'concerned with whether there is a causal relationship between the treatment and the outcome of an experiment' [94]. The threats to this validity are related to issues such as experimental procedures, treatments, the selection of the participants, each of which may affect the validity of the conclusions [94].

The treatment in our survey is the definitions of the dataset category elements and the formal definitions of data quality issues. There was no group free from treatment in our survey and we are therefore unable to say whether treatment has an effect or not. However, the participants who completed the survey probably answered the questions based on the given definitions of the dataset category elements and the formal definitions of data quality issues. Therefore, we could say that the definitions of the dataset category elements and the formal definitions of data quality issues probably have some effect.

7.7.3 Construct validity

Construct validity is ‘concerned with how well the treatment and the outcome reflect the concept or theory behind an experiment’ [94]. The threats to this validity are related to the design of the experiment and the behaviour of the subjects and experimenters [94]. We thus need to evaluate whether the questions used in the survey measured the effectiveness of the definitions for the dataset category elements in a data set and the usability of the formal definitions for data quality issues.

The main threat to construct validity in our survey was the survey design. In the survey, the dataset framework section contained three sets of questions about data sets and one open-ended question. The three sets of questions about data sets used three different examples of data sets. Two examples of data sets were real data sets from a public repository and one example was an artificial data set. We created the artificial data set to illustrate the format and structure of a real data set. However, we believe other examples of data sets from public data repositories could be used to better represent the dataset category elements than the existing examples of data sets in the survey. While the quality of the examples of data sets presented in the survey might be a potential threat to the validity of our survey, it was not obvious that these examples had an effect on our conclusion.

The length of the survey could also be a threat to the construct validity of our survey. Some questions might have been left unanswered in our survey because the participant needed to answer 52 questions to complete it. Our evaluation of the effectiveness of the definitions of dataset category elements and the usability of formal definitions for data quality issues could be affected by incomplete answers. To mitigate this threat, we allowed users to skip related questions if a specific condition was met in order to minimise the number of questions in the survey. This was to minimise the number of questions in the survey. As described in Section 7.3.2, we conducted a pilot test to see how long it took to complete the survey and to identify any problems with the questions. This pilot test helped us to improve the questions in the survey.
Another threat to the construct validity is related to how the participants answer the questions related to the identification of the dataset category elements. We intended to use these questions to observe whether or not the participants understood the definitions of the dataset category elements that correspond to the concepts in the dataset metamodel and were able to identify the correct dataset category elements in the data sets. However, it is possible that the participants applied their own understanding of dataset concepts to identify the correct dataset category elements. For example, academic researchers might not have read the given definitions of the dataset category elements in order to identify them correctly because they were likely to have had extensive knowledge and research experience with data sets.

If the academic researchers identified the dataset category elements incorrectly for this reason, this could affect our analysis of the effectiveness of the definitions of dataset category elements. On the other hand, if the academic researchers identified the dataset category elements correctly for this reason, this indicates that the definitions of the dataset category elements in the metamodel were consistent with their understanding of dataset concepts. Moreover, the academic researchers might have been able to recognise the correct dataset category elements in the data sets because the constructed definitions had captured what they recognised as the elements in a data set. This shows that the definitions of dataset category elements are consistent with whether the participants identified the elements in the data sets that correspond to the definitions or the participants identified the elements in the data sets by applying their understanding about the concepts of data sets. Therefore, we believe that the constructed definitions are appropriate for identifying the dataset category elements because the concepts in the dataset metamodel are consistent with the way that the participants with research experience think. This conclusion was supported by the findings of the survey that indicated that the participants with research experience were able to correctly identify the dataset category elements in the data sets.

Another possible threat to the construct validity would be if the definitions of the dataset category elements were inconsistent with the concepts in the metamodel. The participants might have answered the questions incorrectly, which, in turn, would affect the conclusions of our survey. For example, if the given definitions for record identifier was inconsistent with the entity in the metamodel, the participants might have answered the questions incorrectly. We think that this threat is unlikely to be relevant because we constructed the definitions of the dataset category elements carefully to be as precise as possible and ensure that the participants could clearly understand the definitions of the dataset category elements.

### 7.7.4 External validity

External validity is related to generalising the survey results. A threat to external validity is a condition that limits our ability to generalise the survey results [94]. For example, a threat is a clarification of how we might be wrong in making a generalization [96].

In our survey, the main threat to external validity was the way the participants were
7.8 Conclusion

In this chapter, we have presented the design, the execution and the results of the survey based on the data gathered, as well as the analysis based on the findings of these results. We have also described the threats to the validity of our survey in this chapter. The survey was conducted towards the end stage of our research to evaluate the effectiveness of the definitions of the dataset category elements and the usability of the formal definitions of data quality issues.

The results from the survey did not provide strong support for our conclusion because of the small number of participants who responded. However, the findings did suggest that participants with relevant background knowledge and experience in research, particularly in analysing data sets, were able to apply the definitions of the dataset category elements and the formal definitions of data quality issues to the data sets successfully. The participants' ability to identify the dataset category elements in the three data sets by applying the definitions of these elements shows that 50% of the participants were able to correctly identify some of these. The participants' responses show that the dataset category elements identified correctly most often were measurement values, followed by...
In addition, the results for identifying data quality issues in the three data sets by applying the formal definitions of these show that 46% of the participants were able to identify these correctly. The responses from the participants show that the data quality issues were easily identified in the tabular data set. The analysis indicated that the participants’ ability to identify the dataset category elements and data quality issues were likely to have relations with their demography, particularly their background, general knowledge and experience in analysing data sets.

Despite the small number of participants who responded, the responses from participants help us to discover some problems with the definitions of dataset category elements. We improved some definitions of the dataset category elements to capture a more precise meaning and to make them clearer to understand.

In the next chapter, we present an observational study for the evaluation of part of the framework. This observational study evaluate the effectiveness of the new definitions of dataset category elements and the usability of the formal definitions of data quality issues. We present our design, execution and results of the observational study in the next chapter. We also discuss threats to the observational study’s validity and summarise the findings.
This chapter presents an observational study to evaluate the data quality framework. We begin by introducing the observational study, then we present the study design, execution and results. This is followed by a discussion of the findings resulting from the study. The chapter ends with some discussion of the threats to the study’s validity.

8.1 Introduction

In Chapter 7, we reported on a survey we conducted as our preliminary evaluation to determine the effectiveness of dataset category element definitions and the usability of formal definitions of data quality issues. The survey results are indicative because of the small number of participants who responded. However, the findings do suggest that participants with relevant background knowledge and experience in research, particularly in analysing datasets, were able to apply the definitions of dataset category elements and the formal definitions of data quality issues to the datasets correctly. As the findings also indicate there were some problems with the definitions of dataset category elements, we improved these after conducting an analysis of the survey. Therefore, in this chapter, we report on an observational study we conducted to test the new definitions of dataset category elements and the formal definitions of data quality issues, aiming to observe how researchers apply the definitions of dataset category elements and the formal definitions of data quality issues.

The goal of the observational study was same as that of the previous survey; that is, to evaluate the effectiveness and usability of part of the framework for data quality assessment. However, the research questions of the observational study were different to those of the survey, because we used different methods. We used observation, a questionnaire and think aloud as our observational study methods to provide further insights into the framework through participant thought processes while applying the
part of the framework. We provided examples of applying the part of the data quality framework for guidance.

In this chapter, we present the design of the observational study in Section 8.2 and the procedures used to execute it in Section 8.3. Our analysis of the responses collected is presented in Section 8.4. Finally, in Section 8.5, we discuss the potential threats to the study’s validity and summarise the findings in Section 8.6. The structure of this chapter is based on guidelines for reporting experiments in software engineering by Jedlitschka [93], as in Chapter 7.

8.2 Design of the study

In this section, we present the design of the observational study carried out to evaluate the effectiveness and usability of part of the framework for data quality assessment. We begin by describing the observational study goals and research questions in Section 8.2.1. We then discuss the target population of participants in Section 8.2.2, the sampling methods used in Section 8.2.3, the study materials in Section 8.2.4, the tasks required of the participants in Section 8.2.5 and the study hypotheses in Section 8.2.6.

8.2.1 Goal and study research questions

The observational study was designed to determine the effectiveness of the application of definitions of dataset category elements and the effectiveness of the application of formal definitions for data quality issues to a range of datasets.

We constructed three observation research questions (ORQs) for the observational study:

- **ORQ1:** Can most software engineering researchers correctly apply the new definitions of dataset category elements?
  
  We constructed this question to test whether most SERs are able to apply the new definitions of dataset category elements correctly to a range of datasets. We planned to answer this question by comparing the correct and incorrect responses made by SERs. A majority of correct responses would indicate that most SERs interpreted the definitions of dataset category elements correctly.

- **ORQ2:** Can most software engineering researchers correctly identify the quality issues in data sets using the formal definitions of data quality issues?
  
  With this question, we tested whether most SERs identify quality issues correctly by applying the formal definitions of data quality issues. We planned to compare the correct and incorrect responses made by SERs. A majority of correct responses would indicate that most SERs interpreted the formal definitions of data quality issues correctly.

- **ORQ3:** Can most software engineering researchers correctly identify and evaluate the metadata in the data sets?
8.2. DESIGN OF THE STUDY

We constructed this question to compare the correct and incorrect responses made by SERs to identify and evaluate the metadata of datasets. A majority of correct responses would indicate most SERs identified and evaluated the metadata in the datasets correctly.

8.2.2 Participants

The target population for the observational study, as in the survey study, was researchers, including academic and postgraduate researchers. In addition, we aimed to recruit postgraduate students in software engineering and computer science. Postgraduate students studying in these areas are familiar with datasets because they are still carrying out very specific research projects that require them to analyse datasets. Postgraduate students who participated in the observational study would gain knowledge in understanding datasets, and this new knowledge could be applied to their own datasets in future research.

8.2.3 Sampling

As in our survey, convenience sampling was used for the observational study. We considered any researcher who was willing to participate as a potential participant. Any postgraduate student majoring in software-engineering or computer-science-related degrees willing to participate was also considered a potential participant.

8.2.4 Materials

The observational study involved the use of a task list with observation and a questionnaire.

Task List  The task list contained a set of tasks that participants had to complete and the observation contained a set of observation questions to be answered by the researcher (who conducted the study). We asked participants to perform four different tasks relating to the data quality framework (please refer to Appendix L for the task list). These tasks were constructed to allow the participant to explore the data quality framework and to apply the new definitions of dataset category elements and the formal definitions of data quality issues to the datasets.

The tasks were:

1. Explore the data quality framework.
   In this task, the participant was given a set of data quality framework documents that contained definitions of dataset category elements, formal definitions of data quality issues, an example of applying the definitions of dataset elements and an example of identifying the data quality issues. We required the participant to read and explore the data quality framework documents and we encouraged them to communicate with the researcher to gain understanding about the data quality framework.
2. Identify the dataset category elements.
   In this task, the participant was given two datasets: Dataset A and Dataset B. Dataset A (Comma Separated Value format) was a real dataset from public data repositories and Dataset B (tabular format) was an artificial dataset. We required the participant to identify the dataset category elements in the two datasets by applying the new definitions of dataset category elements.

3. Identify data quality issues.
   We required participants to apply the formal definitions of data quality issues to identify quality issues in the two datasets, A and B.

4. Evaluate metadata of data set.
   Participants were asked to evaluate the metric metadata and entity metadata in the two datasets, A and B.

**Observation**  We applied a combination of two methods for the observation: (1) unobtrusive observation, and (2) obtrusive observation. With unobtrusive observation, we observed the participant’s use of the data quality framework while performing the given tasks. This method was performed in the early stage of the observational study. We aimed to observe two aspects:

1. how participants applied the definitions of dataset category elements and the definitions for data quality issues
2. whether participants managed to complete the tasks on identifying dataset category elements and identifying data quality issues in the datasets.

In the obtrusive observation, we asked each participant what they thought about the framework after they had completed the given tasks.

With both observation methods, no personal information about the participants was collected, as participation in this study was treated anonymously. To collect the observation data, we observed the participant’s activity while undertaking each task, according to the following observation questions (OQs):

1. Task 1: Explore the data quality framework
   a) OQ1: Does participant look at the page containing the definitions of dataset category elements, the definitions of data quality issues and the examples of applying part of the data quality framework? (Yes/No)

2. Task 2: Identify the dataset category elements
   a) OQ2a: Does the participant communicate with the researcher? (Yes/No)
8.2. DESIGN OF THE STUDY

b) OQ2b: Does the participant look at the page containing the definitions of dataset category elements while performing the task? (Yes/No)

c) Does the participant complete the given task? (Yes/No)

3. Task 3: Identify data quality issues

a) OQ3a: Does the participant communicate with the researcher? (Yes/No)

b) OQ3b: Does the participant look at the page containing the formal definitions of data quality issues while performing the task? (Yes/No)

c) OQ3c: Does the participant complete the given task? (Yes/No)

4. Task 4: Evaluate metadata of dataset

a) OQ4a: Does the participant communicate with the researcher? (Yes/No)

b) OQ4b: Does the participant look at the page containing the definitions of dataset category elements while performing the task? (Yes/No)

c) OQ4c: Does the participant complete the given task? (Yes/No)

Questionnaire The questionnaire consisted of five multiple-choice questions (Q1- Q5) and two open-ended questions (Q6 and Q7). The five multiple-choice questions were about the participants’ background in research and their experience with datasets from data repositories. In Q6, we asked participants if they had any comments on quality issues in datasets. In Q7, participants could comment on ways to improve the data quality framework. (Please refer to Appendix L for the questionnaire.)

8.2.5 Tasks

The observational study required participants to read the set of data quality framework documents, complete a set of tasks and answer a questionnaire. From the pilot test of the observational study, we estimated that the observational study would take approximately 30 minutes. The participants’ main task was to complete the tasks on the task list and answer the questionnaire.

8.2.6 Observational study hypotheses

In this subsection, we describe the hypotheses to be tested in the observational study.
8.2.6.1 Hypotheses

We formulated the following observational study hypotheses (Hs). Each hypothesis corresponds to each of the research questions, ORQ1, ORQ2 and ORQ3.

- H1: The majority of SERs will correctly apply the new definitions of dataset category elements in datasets.
- H2: The majority of SERs will correctly identify one or more quality issues in datasets using the formal definitions of data quality issues.
- H3: The majority of SERs will correctly identify and evaluate one metadata or more in datasets.

8.3 Execution

8.3.1 Preparation

We applied to the University of Auckland Human Participants Ethics Committee for ethics approval to conduct the observational study and received approval on 1 December 2017.

In our observational study, we used two different methods for participant recruitment. First, we advertised our observational study using posters within Universiti Teknologi MARA (UiTM), Shah Alam campus, Malaysia. The posters were posted on the notice board in the Department of Computer Science. Second, we emailed an invitation to participate to potential researchers who we knew were involved in software engineering research.

Both posters and email invitations contained an advertisement providing information on the observational study and our contact details. When receiving emails expressing their interest in participating, we replied with available times for the observational study and attached a participant information sheet (PIS). The PIS contained information about the procedure of the observational study, data storage, the anonymity of responses, and contact details. When we scheduled the time with potential participants, we sent details of the location where the in-person observational study would be held. (Please refer to Appendix M for the PIS and invitation email.)

8.3.2 Procedure

The observational study was undertaken from December 2017 to February 2018. The study was conducted on a one-to-one basis in meeting rooms within the Department of Computer Science. As previously indicated, the study involved participants completing tasks on a list, observation of participants during completing the tasks, and a participant questionnaire.

As noted in Section 8.3.1, before the start of the observational study, each participant was given a PIS. They were also provided with a consent form and an assurance letter from the Head of Department of Computer Science. The consent form ensured that participants understood the conditions for taking part in the observational study. Participants were
asked to sign and return the consent forms if they agreed to participate. The assurance letter assured participants that their grades or employment with the UiTM would not be affected. Once each participant had signed and returned the consent form, they were given the task list and questionnaire.

Before participants started performing the observational study, they were asked to follow the instructions written on the task list and the questionnaire. Participants were also encouraged to ask any questions during the observational study session. During the study, we observed how the participants performed the tasks on the list and answered the questionnaire. Participants were asked to return the task list and the questionnaire when they finished.

**8.4 Analysis**

By the end of the recruitment period, we had recruited 25 participants—20 postgraduate students and five academic researchers. All participants took part in the one-to-one study.

As discussed, the observational study consisted of two parts: (1) task list and observation, and (2) background information. The task list contained four tasks: (1) Explore the data quality framework, (2) Identify the dataset category elements, (3) Identify the quality issues, and (4) Evaluate metadata of datasets. As described earlier, we constructed the four tasks to allow the participant to explore the data quality framework and to apply the definitions of dataset category elements and the formal definitions of data quality issues to the datasets. The questionnaire contained seven questions related to participants’ background in research and analysing datasets. All participants managed to complete the two parts.

In the following sections, we present analysis of the data collected from the observational study. The analysis is organised based on the structure of the observational study.

**8.4.1 Task list and Observation**

In part one of the study, we observed how participants applied the definitions of datasets category elements and used the formal definitions of data quality issues when carrying out the four tasks that we structured in the study. As described earlier, the aspects that we observed were:

1. how participants applied the definitions of dataset category elements and the definitions for data quality issues

2. whether participants managed to complete the tasks for identifying dataset category elements and identifying the data quality issues in the datasets.

As detailed in Section 8.2.4, we used a combination of unobtrusive and obtrusive observation methods. First, we performed unobtrusive observation while the participants performed their tasks on the data quality framework. After they had completed the tasks,
8.4. ANALYSIS

we then asked participants to express what they thought about the data quality framework via a think-aloud approach. The observation was performed with the researcher completing a set of structured questions, OQ1- OQ4c, while watching participants complete the tasks. Table 8.1 shows the results of the observation. We discuss these results in the following subsections.

8.4.1.1 Task 1: Explore the data quality framework.

In Task 1, participants were required to read the definitions of dataset category elements and formal definitions of data quality issues, and to read and explore the two examples of applying the data quality framework: (1) an example of applying the definitions of dataset category elements, and (2) an example of identifying the data quality issues.

Observation results: The results for OQ1, shown in Table 8.1, indicate that all participants looked at the related data quality framework documents (the definitions of dataset category elements, the formal definitions of data quality issues and the examples of applying part of the data quality framework). With the think-aloud approach, most of the participants communicated with the researcher to gain an understanding of the definitions of dataset category elements and the formal definitions of data quality issues. For example, some of participants asked one or two questions related to the definitions of dataset category elements, and after the researcher explained the examples of applying the definitions of dataset category elements, all participants managed to understand the definitions of dataset category elements.

8.4.1.2 Task 2: Identify the dataset category elements.

In Task 2, the participants were required to identify the dataset category elements by applying the definitions of dataset category elements in the given datasets (Dataset A and
Dataset B). The participants needed to list the selected elements that met the definitions of the dataset category elements in the space given on the task-list sheet.

**Observation results:** All participants managed to perform the second task. The results for OQ2a, shown in Table 8.1, indicate that nine participants communicated with the researcher, through the think-aloud approach, while performing this task. We received some useful feedback; two of the comments were:

- “It would be easier to identify the dataset category elements after the definitions of dataset elements were well understood.”
- “The dataset definitions provide me with a useful knowledge and reference for understanding data sets”.

We noticed that all participants read the Dataset Definitions sheet and the Example 1 sheet while identifying the dataset category elements in the given datasets. This seems reasonable for first-time use because of participants' unfamiliarity with the terminology for dataset elements. Overall, all participants managed to complete the second task by identifying the dataset category elements using the dataset definitions.

**Task 2 results:** Table 8.2 shows the number of participants who gave the correct answer and the number who gave the incorrect answer when applying the definitions of dataset category elements to the datasets. Most participants were able to apply the definitions of the dataset category elements by identifying these elements correctly across the two different datasets. This can be seen in Table 8.2, which indicates that all participants provided correct answers for measurement value, and there were 45 correct responses for metric label. This result can be explained by the most common elements of datasets being measurement values and these elements being associated with metric labels. Some participants with experience in research may have used measurement values and metric labels frequently in datasets.

We noticed that few participants provided incorrect answers for entity metadata across the two different datasets. In particular, seven responses for Dataset A and eight responses for Dataset B. All responses are from postgraduate students. These participants probably had no experience in using entity metadata in the datasets, because many existing datasets in data repositories do not contain metadata.

We also noticed that seven participants — six postgraduate students and one academic researcher — provided incorrect answers for record identifier in Dataset A. They may not have been familiar with a record identifier and misunderstood the definition for it.

**8.4.1.3 Task 3: Identify the data quality issues.**

For Task 3, the participants were required to apply the formal definitions of data quality issues to the given datasets. In particular, we asked the participants to execute the formal procedure to identify the quality issues (as described in Chapter 5, Section 5.3.2) to
Table 8.2: Numbers of participants who correctly and incorrectly applied the definitions of dataset category elements

<table>
<thead>
<tr>
<th>Dataset category element</th>
<th>Dataset</th>
<th>Participant Responses</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement value</td>
<td>Dataset A</td>
<td>25</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>25</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Metric label</td>
<td>Dataset A</td>
<td>24</td>
<td>1</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>25</td>
<td>0</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Record identifier</td>
<td>Dataset A</td>
<td>18</td>
<td>7</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>22</td>
<td>3</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Entity label</td>
<td>Dataset A</td>
<td>23</td>
<td>2</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>22</td>
<td>3</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Metric metadata</td>
<td>Dataset A</td>
<td>22</td>
<td>3</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>24</td>
<td>1</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Entity metadata</td>
<td>Dataset A</td>
<td>18</td>
<td>7</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>17</td>
<td>8</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Dataset A</td>
<td>130</td>
<td>20</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dataset B</td>
<td>135</td>
<td>15</td>
<td>150</td>
<td></td>
</tr>
</tbody>
</table>

The given datasets. The participants were required to determine the specific dataset category elements and use them to make decisions when identifying the quality issues. For example, participants needed to determine the record identifiers, measurement values and metric labels in the dataset to make decisions when identifying duplicate data.

Observation results: All participants managed to perform the third task. Table 8.1 shows that all participants read the Formal Definitions sheet that contained the definitions of data quality issues and the Example 2 sheet that contain the example of identifying the data quality issues while identifying the data quality issues in the given datasets. However, some participants appeared to forget some of the definitions of dataset category elements (e.g. record). Thus, they read the Dataset Definitions sheet to identify some of the elements in the given datasets. Overall, all participants managed to complete the third task. For this task, six of the 25 participants communicated with the researcher through the think-aloud approach. We received some useful comments; two were:

- “It is better to have automatic detection for common quality issues such as duplicate data and missing data.”
- “It would be much easier if the quality issues were explicitly highlighted in the given datasets”.

Task 3 results: Table 8.3 shows the numbers of participants who correctly and incorrectly applied the formal definitions of data quality to the given datasets.

Most participants were able to identify one or more data quality issues using the formal definitions of data quality issues in the two datasets. This can be seen in Table 8.3, which indicates all participants provided correct answers for identifying missing...
8.4. ANALYSIS

Table 8.3: Numbers of participants who correctly and incorrectly applied the formal definitions of data quality issues

<table>
<thead>
<tr>
<th>Data quality issues</th>
<th>Dataset</th>
<th>Participant Responses</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicate data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>24</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Dataset B</td>
<td>22</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Inconsistent data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>22</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Dataset B</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>22</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Dataset B</td>
<td>25</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Incomplete metadata for a metric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>12</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>Dataset B</td>
<td>23</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Incomplete metadata for an entity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>10</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Dataset B</td>
<td>13</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset A</td>
<td>80</td>
<td>35</td>
<td>115</td>
</tr>
<tr>
<td>Dataset B</td>
<td>88</td>
<td>27</td>
<td>115</td>
</tr>
</tbody>
</table>

data in Dataset B. This result could be because the most common quality issue in datasets is missing data. Participants who had experience in analysing datasets may have encountered missing data in datasets frequently. Ten participants —nine students and one academic researcher —did not correctly identify inconsistent data in Dataset B. These participants may not have encountered inconsistent data in datasets before.

Thirteen postgraduate student participants provided incorrect answers for identifying the incomplete metadata for a metric in Dataset A. This seems reasonable because the postgraduate students probably had less experience in analysing datasets, so this could be why they were sometimes unable to identify the quality issue correctly.

8.4.1.4 Task 4: Evaluate the metadata in the data sets.

Finally, the participants were required to find the metadata in the given datasets then assess whether it explicitly described the meaning of the property of the entity.

Observation results: All of the participants managed to perform the fourth task. Table 8.1 shows that all participants read the 'Dataset Definitions' sheet to identify the entity metadata and metric metadata in the given datasets. The participants managed to complete the fourth task by evaluating the metadata in the datasets. Of the 25 participants, seven communicated with the researcher through the think-aloud approach. We received some feedback from participants; two comments were:

• “It is much easier to evaluate the metadata if you actually know the meaning of the metric label”.

• “It takes times to understand the given metadata in the data set”.
8.4. ANALYSIS

**Task 4 results:** Table 8.4 shows the numbers of participants who correctly and incorrectly identified and evaluated the metadata in the datasets.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Dataset</th>
<th>Participant Responses</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Availability of metric metadata</td>
<td></td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Completeness of metric metadata</td>
<td></td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>Availability of entity metadata</td>
<td></td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Completeness of entity metadata</td>
<td></td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>A</td>
<td>92</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>67</td>
<td>33</td>
</tr>
</tbody>
</table>

Most participants were able to identify and evaluate the metric metadata in the two datasets. This can be seen in Table 8.4, which indicates that all 25 participants provided correct responses for the availability of metric metadata in Dataset B, while 24 participants provided correct responses for the same in Dataset A. In addition, 21 participants provided correct responses for the completeness of metric metadata in Dataset A, and 22 participants provided correct responses for the same in Dataset B. This indicates that most of the participants were able to identify and evaluate the metric metadata correctly in both datasets, which may be because they have used metric metadata previously in research.

Some participants provided incorrect answers for evaluation of entity metadata in Dataset B—Table 8.4 indicates that 15 participants provided incorrect responses for the availability of entity metadata and 15 participants provided incorrect responses for the completeness of entity metadata. All the incorrect responses were from postgraduate students. They may have misunderstood the definition (for ‘entity metadata’) because they had less experience in research.

### 8.4.2 Demographics

In Part 2 of the study, we asked participants to answer seven questions related to their background information. The aim of this part of the study was to find out whether the participant had experience in analysing datasets and had encountered data quality issues with datasets.

Five academic researchers and 20 postgraduate students participated in the observational study. Of the five academic researchers, two had more than 10 years of experience and three had five to 10 years of experience doing research in computer science or software engineering. All postgraduate students had less than five years of experience doing
research in computer science or software engineering. Figure 8.1 shows participants’ experience in research.

![Figure 8.1: Participants’ experience in research](image)

Four of the five academic researchers had experienced analysing datasets in their research. Two had used University of California Irvine (UCI) Machine Learning datasets, one had used International Software Benchmarking Standards Group (ISBSG) datasets and one had used Kaggle datasets. Of the 20 postgraduate students, 12 had experienced analysing datasets in their research. Of these 12 postgraduate students, two had used datasets from data.gov.my repository; one had used datasets from the Kaggle repository; one had used datasets from sinar.gov.my; one had used datasets from the GPS repository; and seven indicated they had used other datasets by selecting option ‘Other’, without reporting the name of the repository they had used.

In the questionnaire, participants were also asked about dataset quality issues that they had encountered. This question allowed participants to select more than one quality issue. Although some participants reported that they did not have experience in analysing datasets, all participants reported encountering data quality issues with datasets —possibly, these participants encountered the data quality issues with datasets not in their research, but in their work or assignments. Sixteen participants reported encountering ‘missing data’, 10 ‘inconsistent data’, 12 ‘duplicate data’ and six ‘incorrect data’. Participants who indicated more than one quality issue were counted for each quality issue they selected; therefore, the sum of the participant responses is greater than 25. Figure 8.2 shows the number of encounters participants indicated they had with data quality issues in datasets.

We also asked participants to comment on quality issues in datasets in Question 6. Of the 11 participants who answered this question, four were academic researchers and seven were postgraduate students. Their comments are shown in Table 8.5.

Further, we asked participants about how to improve the data quality framework in Question 7. The four participants answered this question were academic researchers. Their comments are shown in Table 8.6.
8.4. ANALYSIS

Figure 8.2: Number of participant encounters with data quality issues

Table 8.5: (Q6) Comments on quality issues in datasets

<table>
<thead>
<tr>
<th>No.</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>'Experiment using datasets that have missing data might cause reliability issues.'</td>
</tr>
<tr>
<td>2.</td>
<td>'Quality of data sets determines the reliability of the results research.'</td>
</tr>
<tr>
<td>3.</td>
<td>'There is no metadata provided with the datasets.'</td>
</tr>
<tr>
<td>4.</td>
<td>'Need automation to check for the quality of data.'</td>
</tr>
<tr>
<td>5.</td>
<td>'Some of the data are not consistent.'</td>
</tr>
<tr>
<td>6.</td>
<td>'Some data in CSV file are missing, we have to search on Google for the correct information to fill up the empty space.'</td>
</tr>
<tr>
<td>7.</td>
<td>'Not complete metadata, no reference (date, time, source).'</td>
</tr>
<tr>
<td>8.</td>
<td>'Highlighted the null values to solve missing data.'</td>
</tr>
<tr>
<td>9.</td>
<td>'Metadata is not provided in the datasets.'</td>
</tr>
<tr>
<td>10.</td>
<td>'Specifically need to identify quality of data.'</td>
</tr>
<tr>
<td>11.</td>
<td>'Found missing data, duplicate data, inconsistent data and incorrect data.'</td>
</tr>
</tbody>
</table>

Table 8.6: (Q7) Comments on the data quality framework.

<table>
<thead>
<tr>
<th>No.</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>'It is good for automating the process of this framework.'</td>
</tr>
<tr>
<td>2.</td>
<td>'Automate the manual process of handling data sets.'</td>
</tr>
<tr>
<td>3.</td>
<td>'Data creator should concern on producing better quality data.'</td>
</tr>
<tr>
<td>4.</td>
<td>'The creator of dataset must ensure the quality of data produced.'</td>
</tr>
</tbody>
</table>

8.4.3 Analysis of participants’ application of the new definitions of dataset category elements according to participants’ background

To answer ORQ1, outlined in Section 8.2 —that is, whether most SERs can correctly apply the new definitions of dataset category elements —we focused on participant responses to the questions in Task 1. We analysed participants’ correct responses grouped by participant background (academic researcher and postgraduate student). This analysis allowed us to explore which group provided more correct answers in the observational
study. The results of this analysis are presented in Figure 8.3 for Dataset A and Figure 8.4 for Dataset B.

Figure 8.3 shows that most participants correctly applied the new definitions of dataset category elements with Dataset A. In particular, all academic researchers correctly applied the new definitions for measurement value, entity label, entity metadata and metric metadata, and all postgraduate students correctly applied the new definitions for measurement value and metric label. Eighty per cent of academic researchers’ applications of the new definitions for metric label and record identifier were correct. This shows most academic researchers responded with correct answers for most of the dataset category elements in Dataset A.

Figure 8.3 also shows that 70% of postgraduate students responded with correct answers for record identifier, 90% for entity label, 85% for metric metadata and 65% for entity metadata. They incorrectly applied the definitions for elements related to entity (record identifier, entity label and entity metadata). This could be because they did not read the definitions of dataset category elements carefully.

Figure 8.4 shows that most participants correctly applied the new definitions of dataset category elements with Dataset B. All academic researchers correctly applied the new definitions for all dataset category elements, while all postgraduate students correctly applied the new definitions for measurement value and metric label. This shows more academic researchers responded with correct answers than did postgraduate students.

In Figure 8.4, we can see that 85% of postgraduate students responded correctly for both record identifier and entity label, 95% for metric metadata and 60% for entity metadata. As with Dataset A, some postgraduate students responded incorrectly for elements related to entity. This might be because they were not familiar with the dataset elements or misunderstood some of the definitions of dataset category elements.

Figures 8.3 and 8.4 show that most of the academic researchers correctly applied
the new definitions of dataset category elements to datasets. As described in Section 8.4.2, all academic researchers who participated in this study had more than five years’ experience in research. This suggests that the academic researchers who participated in this observational study had a strong background of knowledge and experience in research that assisted them to apply correctly the new definitions of dataset category elements.

8.4.4 Analysis of participants' application of the new definitions of dataset category elements according to observation data

We analysed participants’ correct responses grouped by the observation question OQ2a (‘Does participant communicate with the researcher? (Yes/No)’). This analysis allowed us to explore which group ((1) participants who communicated with the researcher, or (2) participants who did not communicate with the researcher) provided the most correct answers in the observational study. As shown in Table 8.1, nine participants communicated with the researcher (‘Yes’) and 16 participants did not (‘No’). We calculated the percentage of total correct responses for six dataset category elements for the two groups, (1) participants who communicated with the researcher, and (2) participants who did not communicate with the researcher. The results of this analysis are presented in Figure 8.5 for Dataset A and Figure 8.6 for Dataset B.

Figure 8.5 shows that 92.59% of the correct responses were made by participants who communicated with the researcher while performing Task 2 with Dataset A, while 82.29% were made by participants who did not communicate with the researcher. This reveals that participants who communicated with the researcher responded correctly more often than participants who did not. It seems that the participants who communicated with the researcher gained a better understanding of the definitions of dataset category elements than those who did not communicate with the researcher.

Figure 8.6 shows that 90.74% of correct responses were made by participants who communicated with the researcher while performing Task 2 with Dataset B, while 89.58%
8.4. ANALYSIS

Figure 8.5: Percentage of responses for six dataset category elements in Dataset A

Figure 8.6: Percentage of responses for six dataset category elements in Dataset B

were made by participants who did not communicate with the researcher. This highlights a small difference in correct responses between participants who communicated with the researcher and participants who did not. Some of the participants who did not communicate with the researcher but responded correctly may have read carefully the definitions of dataset category elements and the examples of applying the definitions.

8.4.5 Analysis of participants’ application of the new definitions of dataset category elements according to experience in analysing datasets

We analysed participants’ correct responses grouped by participants’ experience in analysing datasets. This analysis allowed us to explore which group ((1) participants with experience in analysing datasets, or (2) participants without experience in analysing datasets) gave the most correct answers in the observational study. From the demographic data, we found 16 participants had experience in analysing datasets and nine participants did not.

We calculated the percentage of correct and incorrect responses for six dataset
category elements for the two groups, (1) participants with experience in analysing datasets, and (2) participants without experience in analysing datasets. The results of this analysis are presented in Figure 8.7 for Dataset A and Figure 8.8 for Dataset B.

Figure 8.7 shows that 86.46% of correct responses were made by participants with experience in analysing datasets, while 87.04% were made by participants without experience in analysing datasets. This shows participants without experience in analysing datasets responded correctly more than participants with experience in analysing datasets. Although there is a small difference between the two groups, we speculate that participants who did not have experience in analysing datasets responded correctly because they understood the definitions of dataset category elements well.

Figure 8.7: Percentage of responses for six dataset category elements in Dataset A

Figure 8.8 shows that 90.63% of correct responses were provided by participants with experience in analysing datasets, while 88.89% of correct responses were made by participants without experience in analysing datasets. This shows participants with experience in analysing datasets responded more correctly than participants without such experience. This result seems reasonable because it is probable that participants with experience in analysing datasets responded correctly.

Figure 8.8: Percentage of responses for six dataset category elements in Dataset B
8.4.6 Analysis of participants’ application of the formal definitions of quality issues according to participants’ background

To answer ORQ2, described in Section 8.2—that is, whether most SERs can correctly apply the formal definitions of data quality issues—we focused on participant responses to the questions in Task 3. We analysed participants’ correct responses grouped by participant background (academic researcher and postgraduate student). This analysis allowed us to explore which group provided the most correct answers in the observational study. We present the results of this analysis in Figure 8.9 for Dataset A and Figure 8.10 for Dataset B.

Figure 8.9 shows that most academic researchers correctly applied the formal definitions of data quality issues in Dataset A. In particular, all academic researchers correctly applied the formal definitions for duplicate data, missing data and incomplete metadata for a metric. Eighty per cent of all responses from academic researchers were correct for inconsistent data and incomplete metadata for an entity, with researchers correctly applying the formal definitions for these.

In Figure 8.9, we can see that 95% of postgraduate students’ responses were correct for duplicate data, 85% for inconsistent data, 85% for missing data, 35% for incomplete metadata for a metric and 25% for incomplete metadata for an entity. They responded correctly less than 50% of the time for metadata-related quality issues (incomplete metadata for a metric). This could be because some of the postgraduate students may not have been familiar with metadata-related quality issues, so responded incorrectly.

Figure 8.9: Percentage of correct responses for five data quality issues in Dataset A

Figure 8.10 shows that most academic researchers correctly applied the formal definitions of data quality issues in Dataset B. In particular, all academic researchers correctly applied the formal definitions for duplicate data, inconsistent data, missing data and incomplete metadata for a metric, while 80% of academic researchers correctly applied the formal definitions for incomplete metadata for an entity. This shows most academic researchers responded correctly for most of the data quality issues in Dataset B. In Figure 8.10, we can see that 100% of postgraduate students’ responses were correct
for missing data, 90% for incomplete metadata for a metric, 85% for duplicate data, 50% for inconsistent data and 45% for incomplete metadata for an entity. As with Dataset A, some of the postgraduate students may have misunderstood the formal definitions for incomplete metadata for a metric, duplicate data, inconsistent data and incomplete metadata for an entity because they might not have been familiar with these quality issues.

Figures 8.9 and 8.10 show that most of the academic researchers correctly applied the formal definitions of data quality issues to the datasets. As described earlier in section 8.4.2, all academic researchers who participated in the study had more than five years’ experience in research, indicating they had a strong background of knowledge and experience in research that assisted them to apply the formal definitions of data quality issues in the datasets correctly.

8.4.7 Analysis of participants’ application of the formal definitions of data quality issues according to observation data

We analysed participants’ correct responses grouped by the observation question OQ3a (‘Does participant communicate with the researchers? (Yes/No)’). This analysis allowed us to explore which group ((1) participants who communicated with the researcher, or (2) participants who did not communicate with the researcher) most often correctly applied the formal definitions of quality issues. Table 8.1 shows that six participants communicated with the researcher (‘Yes’) and 19 participants did not communicate with the researcher (‘No’). We calculated the percentage of correct responses for five data quality issues across the two groups, (1) participants who communicated with the researcher, and (2) participants who did not communicate with the researcher. The results of this analysis are presented in Figure 8.11 for Dataset A and Figure 8.12 for Dataset B.

Figure 8.11 shows that 83.33% of correct responses were made by participants
who communicated with the researcher while performing Task 3 in Dataset A, while 68.42% of the correct responses were made by participants who did not communicate with the researcher. This shows participants who communicated with the researcher responded more correctly than participants who did not. It seems that participants who communicated with the researcher gained a better understanding of the formal definitions of data quality issues than those who did not communicate with the researcher, allowing them to respond correctly in Dataset A.

Figure 8.11: Percentage of responses for five quality issues in Dataset A

Figure 8.12 shows that 73.33% of correct responses were made by participants who communicated with the researcher while performing Task 3 in Dataset B, and 80.00% of correct responses were made by participants who did not communicate with the researcher. Thus, participants who did not communicate with the researcher responded more correctly than participants who did communicate with the researcher. As described in Section 8.4.2, most participants had encountered data quality issues in datasets. It could be that participants who did not communicate with the researcher responded correctly because they were familiar with data quality issues in datasets.

Figure 8.12: Percentage of responses for five quality issues in Dataset B
8.4.8 Analysis of participants’ application of the formal definitions of data quality issues according to experience in analysing datasets

We analysed participants’ correct responses grouped by their experience in analysing datasets. This analysis allowed us to explore which group ((1) participants who had experience in analysing datasets, or (2) participants who did not have experience in analysing datasets) provided the most correct answers in the observational study. As detailed earlier, the demographic data revealed 16 participants had experience in analysing datasets and nine participants did not.

We calculated the percentage of correct and incorrect responses for five data quality issues across the two groups, (1) participants with experience in analysing datasets, and (2) participants without experience in analysing datasets. The results for this analysis are presented in Figure 8.13 for Dataset A and Figure 8.14 for Dataset B.

![Figure 8.13: Percentage of responses for five quality issues in Dataset A](image)

Figure 8.13 shows that 70.00% of correct responses were made by participants who had experience in analysing datasets, while 75.56% of correct responses were made by participants without experience in analysing datasets. This shows the participants without experience in analysing datasets responded more correctly than those participants with experience. It could be that the participants without experience in analysing datasets responded correctly because they understood the formal definitions of data quality issues well. This suggests that our formal definitions of data quality issues were able to help participants to identify the quality issues in datasets correctly.

Figure 8.14 shows that 86.25% of correct responses were made by participants with experience in analysing datasets, while 64.44% of correct responses were made by participants without experience in analysing datasets. Thus, participants who had experience in analysing datasets responded more correctly than participants who did not. This result seems reasonable because participants with experience in analysing datasets probably responded correctly because they were familiar with datasets.
8.4.9 Analysis of participants’ evaluation of the metadata in datasets according to participants’ background

To answer ORQ3, as detailed in Section 8.2—that is, whether most SERs can correctly identify and evaluate the metadata in datasets—we focused on participant responses to the questions in Task 4. We analysed participants’ correct responses grouped by participant background (academic researcher and postgraduate student). This analysis allowed us to explore which group provided the most correct answers in the evaluation of the metadata in datasets. The results for this analysis are presented in Figure 8.15 for Dataset A and Figure 8.16 for Dataset B.

Figure 8.15 shows that most participants correctly identified and evaluated the metadata criteria in Dataset A. In terms of participants’ background, all academic researchers correctly identified and evaluated all the metadata criteria, while 95% of postgraduate students correctly identified the entity metadata and metric metadata. We can also see that 90% and 80% of postgraduate students correctly evaluated the completeness of entity metadata and the completeness of metric metadata, respectively. Some of the postgraduate students who responded incorrectly may have done so because they did not read the definitions of entity metadata and metric metadata carefully.

Figure 8.16 shows that all academic researchers correctly identified and evaluated all the metadata criteria in Dataset B. For postgraduate students, all correctly identified the availability of metric metadata, 85% correctly evaluated the completeness of metric metadata and 25% correctly identified the entity metadata and evaluated its completeness in Dataset B. We can see in Figure 8.16 that more than 50% of the postgraduate students incorrectly evaluated the entity metadata. Possibly, some of the postgraduate students were not familiar with metadata in datasets or misunderstood the definitions of entity metadata and metric metadata.
8.4.10 Analysis of participants’ evaluation of the metadata in datasets according to observation data

We analysed participants’ correct responses grouped by the observation question OQ4a (‘Does participant communicate with the researchers? (Yes/No)’). This analysis allowed us to explore which group ((1) participants who communicated with the researcher, or (2) participants who did not communicate with the researcher) most often correctly evaluated the metadata in datasets. Table 8.1 states seven participants communicated with the researcher (‘Yes’) and 18 participants did not communicate with the researcher (‘No’) for OQ4a. We calculated the percentage of correct responses for five data quality issues for both groups, (1) participants who communicated with the researcher, and (2) participants who did not communicate with the researcher. The results of this analysis are presented in Figure 8.11 for Dataset A and Figure 8.12 for Dataset B.

Figure 8.17 shows that 92.86% of correct responses were made by participants who communicated with the researcher while performing Task 4 in Dataset A, and 91.67% of correct responses came from participants who did not communicate with the researcher.
There is a small difference in the percentage of correct responses between the two groups. This shows that most participants responded correctly even if they did not communicate with the researcher. It seems the definitions of dataset category elements and the examples of identifying the dataset category elements assisted most participants to identify and evaluate the metadata in the dataset correctly.

Figure 8.17: Percentage of correct responses for metadata evaluation in Dataset A

Figure 8.18 shows that 75.00% of correct responses were made by participants who communicated with the researcher while performing Task 4 in Dataset B, and 63.89% of correct responses were made by participants who did not communicate with the researcher. Thus, participants who communicated with the researcher responded more correctly than participants who did not. This indicates that participants who communicated with the researcher gained a better understanding of the definitions of entity metadata and metric metadata that allowed them to respond correctly in Dataset B.

Figure 8.18: Percentage of correct responses for metadata evaluation in Dataset B
8.4.11 Analysis of participants’ evaluation of the metadata in datasets according to experience in analysing datasets

We analysed participants’ correct responses grouped by participant experience in analysing datasets. This analysis allowed us to explore which group ((1) participants with experience in analysing datasets, or (2) participants without experience in analysing datasets) provided the most correct answers in the observational study. From the demographic data, we found that 16 participants had experience in analysing datasets and nine participants did not.

We calculated the percentage of correct and incorrect responses for four metadata criteria for both groups, (1) participants with experience in analysing datasets, and (2) participants without experience in analysing datasets. The results of this analysis are presented in Figure 8.19 for Dataset A and Figure 8.20 for Dataset B.

Figure 8.19 shows that 92.19% of correct responses with Dataset A were made by participants with experience in analysing datasets, and 91.67% of correct responses were provided by participants without experience in analysing datasets. We can see that there is a small difference in the percentage of correct responses between the two groups. This indicates that most participants responded correctly even those without experience in analysing datasets.

Figure 8.19: Percentage of responses for metadata evaluation in Dataset A

Figure 8.20 shows that 67.19% of correct responses with Dataset B were provided by participants with experience in analysing datasets, while 66.67% of correct responses came from participants without experience in analysing datasets. As with Dataset A, there is a small difference in the percentage of correct responses between the two groups. Again, this indicates that most participants responded correctly even those without experience in analysing datasets. It seems that our definitions of dataset category elements and the example of identifying the dataset category elements assisted most participants to identify and evaluate the metadata correctly.
8.4.12 Correctness of application of the new definitions of dataset category elements

We used confidence intervals to estimate the percentage of prospective participants in the entire SER population that would correctly apply the new definitions of dataset category element to the datasets. We applied the adjusted Wald method to provide coverage for the 95% confidence interval that contains the observed proportion on average about 95% of the time. Figure 8.21 shows the 95% confidence interval graph for the six dataset category elements in Dataset A.

From Figure 8.21, we can see the 95% confidence interval for metric label is 79% to 99%, that for record identifier is 52% to 86%, that for entity label is 74% to 99%, that for metric metadata is 69% to 97%, that for entity metadata is 52% to 86% and that for measurement value is 88% to 99%. Overall, the lowest 95% confidence interval percentage is for entity metadata and record identifier, 52% to 86%, and the highest
for is measurement value, 88% to 99%. The confidence intervals for the application of new definitions of dataset category elements suggest that most SER in the population will get the correct answers for measurement value, metric label, entity label and metric metadata, and for record identifier and entity metadata, it is quite likely that most SER will get the correct answers.

![Figure 8.22: The 95% confidence intervals for the correct responses by participants for Dataset B](image)

Figure 8.22 shows the 95% confidence interval graph for the six dataset category elements in Dataset B. The confidence interval for metric label is 88% to 99%, that for record identifier is 69% to 97%, that for entity label is 69% to 97%, that for metric metadata is 79% to 99%, that for entity metadata is 48% to 83% and that for measurement value is 88% to 99%. The lower boundaries of the confidence intervals for metric label, record identifier, entity label, metric metadata and measurement value in Dataset B are higher than 50%. This suggests that we are 95% confident that most SER will get the correct answers for these five dataset category elements using the new definitions of dataset category elements in Dataset B.

We notice that the lower boundary of the confidence interval for entity metadata (48%) is less than 50%. This gives us low confidence that more than 50% of prospective participants in the entire SER population will get the correct answers for entity metadata. Possibly, some participants would misunderstand the new definition of entity metadata because they did not read the definition carefully. Overall, the confidence intervals for Dataset A and Dataset B suggest that most participants in the entire SER population will likely to identify correctly most of dataset category elements using the new definitions of dataset category elements.

8.4.13 Correctness of application of the formal definitions of data quality issues

We used confidence intervals to estimate the percentage of prospective participants in the entire SER population likely to identify correctly one or more quality issues in datasets. We applied the adjusted Wald method to provide coverage for the 95% confidence interval
that contains the observed proportion on average about 95% of the time. Figure 8.23 shows the 95% confidence interval graph for the five data quality issues in Dataset A.

![Figure 8.23: The 95% confidence intervals for the correct responses by participants for Dataset A](image)

From Figure 8.23, we can see the confidence interval for duplicate data is 79% to 99%, that for inconsistent data is 69% to 97%, that for missing data is 69% to 97%, that for incomplete metadata for a metric is 30% to 67% and that for incomplete metadata for an entity is 23% to 59%. The lower boundaries of the confidence intervals for duplicate data, inconsistent data and missing data are higher than 50%. This suggests that most participants in the entire SER population will get the correct answers for the three quality issues (duplicate data, inconsistent data and missing data). In turn, this gives us confidence to support our second hypothesis, H2.

We notice that the lower boundaries of the confidence intervals for incomplete metadata for a metric and incomplete metadata for an entity are lower than 50%. This gives us low confidence that more than 50% of prospective participants in the entire SER population will get the correct answers for incomplete metadata for a metric and incomplete metadata for an entity. It could be that some participants in the SER population would misunderstand the formal definitions for incomplete metadata for a metric and incomplete metadata for an entity because they are not familiar with these quality issues.

Figure 8.24 shows the 95% confidence interval graph for the five data quality issues in Dataset B. The confidence interval for duplicate data is 69% to 97%, that for inconsistent data is 41% to 77%, that for missing data is 88% to 99%, that for incomplete metadata for a metric is 74% to 99% and that for incomplete metadata for an entity is 34% to 70%. The lower boundaries of the confidence intervals for duplicate data, missing data and incomplete metadata for a metric are higher than 50%. This suggests that most participants in the entire SER population will get the correct answers for the three quality issues (duplicate data, missing data and incomplete metadata for a metric) in Dataset B. This gives us confidence to support H2.
8.4. ANALYSIS

The lower boundaries of the confidence intervals for inconsistent data and incomplete metadata for an entity are less than 50%. This gives us low confidence that more than 50% of prospective participants in the entire SER population will likely to identify correctly the inconsistent data and incomplete metadata for an entity using the formal definitions of data quality issues in Dataset B. Possibly some participants would not read the formal definitions for inconsistent data and incomplete metadata for an entity carefully. Overall, the confidence intervals suggests that most participants in the entire SER population will likely to get correct answers for one or more quality issues in the datasets using the formal definitions of data quality issues.

8.4.14 Correctness of identifying and evaluating the metadata in datasets

We used confidence intervals to estimate the percentage of prospective participants in the entire SER population likely to correctly identify and evaluate the metadata in datasets. We applied the adjusted Wald method to provide the coverage for the 95% confidence interval that contains the observed proportion on average about 95% of the time. Figure 8.25 shows the 95% confidence interval graph for the four criteria of metadata evaluation in Dataset A.

From the Figure 8.25, we can see the confidence interval of correct responses for the availability of metric metadata is between 79% and 99%, that for completeness of metric metadata is between 74% and 99%, that for availability of entity metadata is between 79% and 99%, and that for completeness of entity metadata is between 65% and 94%. Overall, the lowest percentage for the lower boundaries of the confidence intervals is 65% and the highest percentage for the upper boundaries of the confidence intervals is 99%. This suggests that most participants in the entire SER population will get correct answers for identifying and evaluating the entity and metric metadata in Dataset A, which supports H2.

From Figure 8.26, we can see the confidence interval of correct responses for availabil-
8.4. ANALYSIS

Figure 8.25: The 95% confidence intervals for the correct evaluation of metadata for Dataset A

Figure 8.26: The 95% confidence intervals for the correct evaluation of metadata for Dataset B

ity of entity metadata is between 23% and 59%, that for completeness of entity metadata is between 23% and 59%, that for availability of metric metadata is between 88% and 99%, and that for completeness of metric metadata is between 69% and 97%.

The lower boundaries of the confidence intervals for entity metadata evaluation are less than 50%, which gives us low confidence to support our third hypothesis, H3. The confidence intervals for metric metadata evaluation are above 50%. This suggests that most participants in the entire SER population will get correct answers for identifying and evaluating the metric metadata in Dataset B, which supports H3. Overall, the confidence intervals suggest that most participants in the entire SER population will likely to identify and evaluate correctly entity and metric metadata in Dataset A and metric metadata in Dataset B.
8.4.15 Discussion of the findings of our observational study

Findings discussed in this section are based on the results of an observational study conducted between December 2017 and February 2018. We conducted the observational study to determine the effectiveness of the application of the new definitions of dataset category elements and the effectiveness of the application of the formal definitions of data quality issues to two datasets, A and B. We presented an analysis of the frequency of data in the observational study to support H1, H2 and H3, and we used confidence intervals to estimate the prospective proportion of the entire SER population likely to apply successfully the new definitions of dataset category elements and the formal definitions of data quality issues.

The findings from the observational study give us high confidence to support our hypotheses. First, we find that the majority of participants applied the new definitions of dataset category elements in the two datasets correctly (H1). Second, we find that the majority of participants identified missing data and duplicate data using the formal definitions of data quality issues in the two datasets correctly (H2). Finally, we find that the majority of participants correctly identified and evaluated metric metadata in the two datasets (H3).

To strengthen the findings, we analysed participant responses based on the participants' background and their experience in analysing or using datasets in their research. We also analysed participant responses based on the observation data. Tables 8.7 to 8.9 present a summary of the analyses and results discussed in the previous subsections.

Table 8.7: Summary of the analyses and results of the application of definitions of dataset category elements to the datasets

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Summary of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants' responses to Task 2</td>
<td>Most participants applied the new definitions of dataset category elements in the two datasets correctly</td>
</tr>
<tr>
<td>Participants' application of the new definitions of dataset category elements according to participants' background</td>
<td>Academic researchers responded more correctly than postgraduate students</td>
</tr>
<tr>
<td>Participants' application of the new definitions of dataset category elements according to observation data</td>
<td>Participants who communicated with the researcher more often responded correctly than participants who did not communicate with the researcher</td>
</tr>
<tr>
<td>Participants' application of the new definitions of dataset category elements according to experience in analysing datasets</td>
<td>Most participants who responded correctly with the two datasets, but particularly with Dataset B, had experience in analysing datasets</td>
</tr>
<tr>
<td>Correctness of application of the new definitions of dataset category elements</td>
<td>The confidence intervals (for Dataset A and Dataset B) suggest that most participants in the entire SER population will likely to identify correctly most of dataset category elements using the new definitions of dataset category elements</td>
</tr>
</tbody>
</table>
Table 8.8: Summary of the analyses and results of the application of formal definitions of data quality issues to the datasets

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Summary of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants’ responses to Task 3</td>
<td>Most participants correctly identified missing data and duplicate data using the formal definitions of data quality issues in the two datasets</td>
</tr>
<tr>
<td>Participants’ application of the formal definitions of data quality issues according to participant’s background</td>
<td>Academic researchers responded more correctly than postgraduate students</td>
</tr>
<tr>
<td>Participants’ application of the formal definitions of data quality issues according to observation data</td>
<td>Most participants who communicated with the researcher responded correctly for Dataset A and most participants who did not communicate with the researcher have responded correctly for Dataset B</td>
</tr>
<tr>
<td>Participants’ application of the formal definitions of data quality issues according to experience in analysing datasets</td>
<td>For Dataset A, most participants who responded correctly did not have experience in analysing datasets, while for Dataset B, most participants who responded correctly had experience in analysing datasets</td>
</tr>
<tr>
<td>Correctness of application of the formal definitions of data quality issues</td>
<td>The confidence intervals for duplicate data and missing data in both datasets are higher than 50%, giving us high confidence that the majority of prospective participants in the SER population would be likely to identify correctly one or more quality issues using the formal definitions of data quality issues</td>
</tr>
</tbody>
</table>

In summary, we found that the results of the observational study provide more evidence than the results of the survey for the evaluation of part of the data quality framework, not only due to the number of participants who responded but also because of the design of the study. For example, we provided some general information about measurements in datasets in the online survey to guide the participants to answer the survey questions, whereas in the observational study, we provided examples of applying the definitions of dataset category elements and the formal definitions of data quality issues to help them perform the given tasks.

We also found that the results of the observational study are better than the survey results. This could be because we allowed participants to ask the researcher questions, via the think-aloud approach, while performing the tasks in the study. We found that participants who communicated with the researcher made fewer errors. However, the observation data in Table 8.1 show that some participants who did not communicate with the researcher also successfully applied the definitions of dataset category elements and the formal definitions of data quality issues. This suggests that our framework helped most of the participants to respond correctly in the observational study.
Table 8.9: Summary of the analyses and results of the evaluation of the metadata in datasets

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Summary of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants’ responses to Task 4</td>
<td>Most participants correctly identified and evaluated one metric metadata in the two datasets</td>
</tr>
<tr>
<td>Participants’ evaluation of the metadata in datasets according to participants’ background</td>
<td>Academic researchers responded more correctly than postgraduate students</td>
</tr>
<tr>
<td>Participants’ evaluation of the metadata in datasets according to observation data</td>
<td>Participants who communicated with the researcher responded more correctly in the evaluation of metadata than participants who did not communicate with the researcher</td>
</tr>
<tr>
<td>Participants’ evaluation of the metadata in datasets according to experience in analysing datasets</td>
<td>Participants with experience in analysing datasets responded more correctly than participants without experience in analysing datasets</td>
</tr>
<tr>
<td>Correctness of identifying and evaluating the metadata in datasets</td>
<td>The confidence intervals for metric metadata in both datasets are higher than 50%, giving us high confidence that the majority of prospective participants in the SER population would be likely to identify and evaluate correctly the metric metadata in datasets</td>
</tr>
</tbody>
</table>

8.5 Threats to validity

In this section, we describe threats to the validity of the observational study and the possible mitigation of these threats.

8.5.1 Conclusion validity

Conclusion validity concerns the conclusions made about the program’s outcome relationship [94]. In our observational study, we did not investigate the relationship between the definitions of dataset category elements and the two datasets. In addition, we did not apply any treatment to the study results. However, we need to be aware of any threats that could affect the conclusions made about the study research questions.

Similar to the threat to the conclusion validity of our survey, the possible threat in the observational study is the random heterogeneity of participants. We focused on two types of participants, academic researchers and postgraduate students. Their amount of research experience, whether in general research or analysing datasets, may have affected how they answered the questions in the observational study. However, from the analysis of the results, we did not find that the difference in experience in analysing datasets had an influence on how effectively participants applied the definitions of dataset category elements. This was clear in that some participants, who had less than five years’ experience in research and did not have experience in analysing datasets, successfully
8.5. THREATS TO VALIDITY

identified the dataset category elements using the new definitions of dataset category elements and identified the quality issues using the formal definitions of data quality issues.

8.5.2 Internal validity

A possible threat to internal validity is related to participants’ reactions during the observational study. Since the study involved participants’ performing the tasks on the task list using the definitions of dataset category elements and formal definitions of data quality issues, our participants could have felt frustrated if they could not find answers. They may also have felt tired if the tasks took too long to perform. This might have led participants to answer randomly or not complete a task or tasks. To reduce the likelihood of participants feeling negative about our observational study, we tried to minimise the length of the task list and to include examples that could help the participants to perform the tasks. However, our analysis of the results indicates all participants managed to complete all the tasks on the list, so it is unlikely that they felt frustrated or tired.

8.5.3 Construct validity

Construct validity threats are related to the design of the experiment and the behaviour of the subjects and experimenters [95]. The first threat to construct validity in our observational study was the study design. We used the task list and questionnaire to collect data from participants. The wording of the questions in the observational study could be a threat to construct validity, as if the questions were ambiguous, the participant could not answer them correctly. In the fourth task in the study, we found the procedures to evaluate the metadata of datasets might have been ambiguous to our participants. We intended to use this task to observe how participants assessed the given metadata in the datasets. However, we believe the wording of these procedures did not make the steps to evaluate the metadata of datasets clear. Some participants might have found it difficult to perform this task. However, the analysis of results indicates that all participants managed to complete the task.

The difference in participants’ years of experience in research or in analysing datasets may have affected the outcome of our observational study, which, in turn, would have affected the conclusion of our study. For example, academic researchers may have performed better than postgraduate students in the study because they were more likely to have had more experience in research. However, the analysis of results did not show any significant difference in how well the participants performed the tasks on the task list with respect to their experience in research or in analysing datasets.

8.5.4 External validity

Similar to our survey, the main threat to external validity in our observational study is the way the participants were recruited. We used convenience sampling for our target
population of participants. We could not generalise our survey to the entire population involved in our research, because our participants were limited to computer science researchers or SERs due to the specific content of the study. Further, our participants were not representative of the whole population of researchers in computer science or software engineering because our participants were postgraduate students and a small number of academic researchers located in one geographical location, UiTM, Malaysia. However, the analysis of results from the study indicates most participants were able to apply the definitions of the dataset category elements and the formal definitions of data quality issues to the datasets successfully. The results also give us high confidence to support our hypotheses.

8.6 Conclusions

In this chapter, we have presented the design, execution and results of our observational study, as well as analyses based on the findings of these results. We have also described the threats to the validity of our observational study. The observational study was conducted to determine the effectiveness of the application of the new definitions of the dataset category elements and the effectiveness of the application of the formal definitions of data quality issues. The observational study results show that most participants successfully applied the definitions of dataset category elements and the formal definitions of data quality issues to the datasets.

The study findings give us confidence to support our hypotheses. First, the analysis of results for applying the new definitions of dataset category elements supports H1, which states the majority of SERs would correctly apply the new definitions of dataset category elements in datasets, and the results show most participants applied the new definitions of dataset category elements in the two datasets correctly.

Second, the analysis of results for applying the formal definitions of data quality issues supports H2, that the majority of SERs would identify correctly one or more quality issues using the formal definitions of data quality issues. The results show that most participants identified correctly missing and duplicate data using the formal definitions of data quality issues. Lastly, the analysis of results for evaluating the metadata in datasets supports H3, which states the majority of SERs would correctly identify and evaluate one metadata or more in datasets. The results show that most participants identified and evaluated metric metadata in the two datasets.

In the next chapter, we discuss the research questions addressed by this thesis. We also discuss relevant points concerning the systematic mapping study, the dataset metamodel and the data quality framework. Additionally, we discuss the limitations of our research and recommended improvements to the survey and the observational study.
In this thesis, we have explained in detail the importance of understanding the quality of data sets by reviewing data set quality problems in the literature and by observing real and artificial data sets. The review of quality problems in the literature found that many definitions of data quality issues are unclear because of the inconsistent terminology used in these definitions. The results from the observation of real data sets indicated that few data sets contain metadata to describe what the data mean, which can lead to misinterpretation. Further, the results of the observation revealed that some aspects of data are commonly present across the varying structures of the data sets. This motivated us to design a metamodel as a standard way to describe the structure of a data set. We modelled real data sets from public data repositories using standard terminology in the metamodel. The standard terminology allowed us to apply a common interpretation to understanding the content of data sets. We then developed a framework to evaluate the quality of the data sets. The framework consists of two parts: modelling the data set (a process for modelling the data set) and quality assessment (a process for assessing quality). We applied the framework to the real data sets from public data repositories in order to understand the quality of the data sets. Finally, we evaluated the usability and the effectiveness of part of our framework by conducting a survey and an observational study. This chapter presents interesting observations and interpretations from our research findings. We also describe the implications and limitations of our research and suggest improvements to our evaluation approach. We begin the chapter with a discussion of the research questions that we answered in this thesis.

9.1 Revisiting the research questions

In this thesis, our main research question was: ‘How can we help researchers to understand the quality of data sets in software engineering?’ As described in Chapter 1, we constructed three subordinate research questions to guide our efforts to answer the main
9.1. Revisiting the Research Questions

We answered the three research questions as follows:

1. **To what extent does the existing research take into consideration the quality of data sets in software engineering?**

   We reviewed the relevant research on data quality in software engineering literature in Chapter 3. The evidence from our review revealed that the existing research paid little attention to the quality of data sets in software engineering. It also highlighted a research gap: many studies used inconsistent terminology to report the problems with data quality, such as missing data and incomplete data. We further reviewed the literature on data quality issues in software engineering data sets. This review identified another research gap: many different definitions were used to describe data quality issues and these definitions used a variety of terminology that could be interpreted in many different ways. After the review, it was concluded that there was no standard definition of data quality issues in software engineering.

2. **What are the content, format and structure of data sets, particularly from public data repositories?**

   We observed real data sets from public data repositories and discussed the outcomes of our observations in Chapter 2. We found that most of the real data sets contained only values and labels, with few containing metadata and other information. In the case of some data sets, the metadata was stored in external locations, which could pose a challenge with regard to interpretation. This suggests that data sets from public data repositories that only contain values and labels are likely to have problems that prevent the correct interpretation of the data.

3. **How do we ensure that data sets from public data repositories are of acceptable quality for research?**

   We developed a data quality framework to support researchers in understanding the quality of data sets. The framework consists of two parts: modelling the data set (a process for modelling the data set) and quality assessment (a process for assessing quality). The dataset modelling process aims to model data sets using a formal procedure based on the dataset metamodel. The dataset metamodel provides a standard terminology to describe the structures and concepts in a data set and the relationships between them. This standard terminology allows researchers to apply a common interpretation to understanding the content of data sets. The quality assessment process provides tasks that determine whether a data set contains sufficient information to facilitate the correct interpretation of the data. It also allows researchers to identify quality issues in the data sets. We modelled real data sets from public repositories using the dataset metamodel (Chapter 4) and assessed the quality of data sets using the quality assessment process (Chapter 5).

We have answered the three subordinate research questions by collecting evidence about the quality of data sets from a literature review, the observation of data sets and a data quality framework. In the following section, we describe and interpret the findings.
collected from the evaluation of the dataset metamodel and the quality assessment, as well as the findings from the survey.

### 9.2 Evaluating the use of the dataset metamodel

We evaluated the use of our dataset metamodel using two approaches. First, we applied the dataset metamodel using the modelling data set process on 92 data sets from the public data repositories (Chapter 4). Second, we asked participants to apply our definition of *dataset category* elements in the dataset metamodel through an online survey (Chapter 7) and an observational study (Chapter 8).

As discussed in Chapter 4, the first evaluation approach revealed that many real data sets contained values (e.g., measurement values) and labels (e.g., metric labels) due to the structure of data sets that typically have measurements and column headers. This is in line with the second evaluation approach that revealed that participants from the survey were more easily able to determine the measurement values and metric labels in the given dataset examples. Both approaches showed a consistency in terms of the identification of the dataset category elements in the data sets. It may be the case that participants from the survey were already familiar with the positions of these dataset category elements in the structure of data sets. In the following paragraph, we discuss three important observations from the evaluation of the dataset metamodel.

The first observation is about the structure of real data sets in public data repositories. Before we designed our metamodel, we observed the structures and formats of real data sets from public data repositories in Chapter 2. We found that many real data sets typically have measurements with multiple columns and column headers are used to distinguish between these columns. We noticed that some elements in the data sets, such as the column headers, always appear in the same position in the structure of the data sets and were therefore easy to identify.

However, we found that a few data sets contained elements other than measurements and column headers, such as metadata. Metadata provide additional information to convey the meaning of the measurements in the data sets. In Chapter 4, we constructed a precise definition of metadata and included it as the key element in our metamodel to support the correct interpretation of data.

In the online survey and the observational study, we found that not many participants were able to identify the metadata in the data sets because they may not have been familiar with the conventions for metadata in the structure of data sets. This was also supported by our observation in Chapter 2 that metadata did not appear consistently in the same position in the real data sets. It is therefore challenging for researchers to correctly identify certain elements, such as metadata, in the data sets.

The second observation is about the common practices in creating data sets for a data repository environment. In Chapter 2, we observed that researchers often recorded basic information using only column headers and measurements in the data sets. While some researchers also recorded metadata to convey the meaning of the measurements, these
metadata were often recorded in external locations, such as web pages and publications.

We discussed the implications of metadata that are recorded in external locations in Chapter 4. In particular, we anticipated a potential for quality issues to arise because researchers may not be aware of the existence of the metadata. In addition, we also described how our metamodel was able to capture the metadata that were located in an external location, particularly on the web page of a data set. This is because we consider the web page to be the first location where the data set is accessed. However, we still believe the ideal way to create a data set is to record all of the elements in a single file. This will almost certainly avoid any potential risk of quality issues in the data set.

The third observation is about the extent of the metadata in the data set that may influence the interpretation of data. As described in Chapter 4, we modelled every data set with assumptions that we made based on the metadata. We found that the process of modelling data sets with insufficient metadata can result in the misinterpretation of data because, without sufficient metadata, we do not know what is actually being measured. It is thus possible to construct an incorrect model of a data set because we do not know the intended meaning of the data.

In the following section, we discuss the conclusions drawn from the application of our dataset metamodel.

9.2.1 Application of dataset metamodel

The results of modelling the data sets (as described in Chapter 4) demonstrate the usefulness of our dataset metamodel for describing the structure of data sets. The results show that we can model real data sets from public data repositories successfully. These data sets originate from two public software engineering data repositories: PROMISE and Qualitas Corpus.

In the PROMISE repository, we applied our dataset modelling procedures to 91 of the 96 selected data sets (as at May 2014) from five categories (defect prediction, effort estimation, text mining, model-based software engineering and general) to discover the various structures of data. In addition, we modelled a data set (Qualitas Weka 3.7.5) from the Qualitas Corpus to observe the differences in the structure of data between the PROMISE and Qualitas data sets.

While we did not apply the dataset metamodel to all of the data sets in the two public data repositories, we actually covered at least all of the common formats and structures of data sets that are often used in research. In Chapter 2, we also considered more uncommon data structures, which we expected could appear in real data sets, and illustrated these uncommon data structures in the examples of artificial data sets.

In this thesis, we focus on model software engineering data sets, as there is currently no standard method to support the interpretation of data. As mentioned earlier, we modelled some data sets from two public software engineering data repositories with our dataset metamodel. While it is possible that not all types of software engineering data sets exist in these two data repositories, at least some of them can be modelled successfully with our dataset metamodel.
We believe that our dataset metamodel could be useful for modelling data sets, not only those related to software engineering, but also other kinds of data sets that have the same structural elements as software engineering data sets, such as label and value. This is because our metamodel is designed based on a common and standard structure of a data set (e.g., containing value, label and metadata). In particular, we introduced metadata as our key element in the dataset metamodel because we wanted to solve quality issues related to the interpretation of data. These quality issues are critical because, if the data set is interpreted incorrectly, we could reach incorrect conclusions in our research. We also considered common quality issues such as duplicate data, missing data and incorrect data, as these quality issues commonly appear in real data sets.

9.3 Evaluating the use of the quality assessment process

We evaluated the use of the quality assessment process in a data quality framework using two approaches. First, we applied the quality assessment process to real data sets from public data repositories (Chapter 5). Second, we asked participants to identify data quality issues in three examples of data sets through an online survey (Chapter 7) and an observational study (Chapter 8). We used real data sets from public data repositories for the examples of data sets in the online survey and the observational study.

The results of applying the quality assessment process (as described in Chapter 5) suggested that some data quality issues could not be identified due to the absence of some dataset category elements in the data sets. These results are likely to be related to the results from the survey, which indicated that 54% of the total participants had responded either incorrectly or answered 'don’t know' when asked to identify data quality issues in the dataset examples. This result could be explained by the fact that the participants were unable to identify some of the dataset category elements required to identify quality issues in the data sets. In the following part of this section, we reflect on the importance of metadata and the procedural element of the framework and how these relate to other forms of quality assessment.

One of the main concerns that motivated this thesis was the lack of clarity about metadata. In the process of conducting this research, we observed a particular example, in Chapter 1, that illustrated an issue with the misinterpretation of data due to insufficient metadata. In Chapter 2, we found that very few real data sets from public data repositories contained metadata. In Chapter 3, we pointed out that there were very few discussions related to metadata and the existing quality assessment techniques; moreover, the quality model produced in other research did not include the treatment of metadata. The framework that we proposed includes metadata as the key element in the dataset metamodel (as described in Chapter 4). The application of the framework has identified the importance of these metadata in Chapter 5 and the process of evaluating the framework reveals this particular concern about metadata.

In this chapter, we confirm that metadata play a key role in the framework. There are particular kinds of quality issues that we can identify as unknowable, that is, whether or
not the data set has quality issues, unless the data set contains metadata. The metadata play a critical role in establishing not only if the data set has quality issues, but also in determining the data set has quality issues.

In terms of the procedures of the framework, our quality assessment procedure is different to other kinds of quality assessment in the way the assessment results are presented. In Chapter 3, we found that a few existing techniques and data quality models presented the quality assessment results by providing the identified quality issues in the data sets. This may allow researchers to decide whether the data set is of good quality or not.

Our assessment results, on the other hand, do not indicate whether the data set is of good quality or not; however, the results provide information that indicates whether or not we have enough information in the data sets to perform the assessment. Furthermore, our quality assessment procedure allows researchers to not only identify the quality issues, but also the information (the dataset category elements) that is not present in the data sets. It is useful for researchers to know whether the data sets contain sufficient information to support a correct interpretation in data analysis conducted during research.

In the following section, we discuss the research findings from the application of our quality assessment process.

9.3.1 Application of quality assessment process

The results of applying the quality assessment process (as described in Chapter 5) illustrate the usefulness of this process in evaluating dataset quality. The results show that we can perform the quality assessment steps successfully on real data sets from public data repositories. Although some of the data sets do not contain the dataset category elements that are essential in identify the quality issues (step 2), our procedures are able to indicate that these data sets might contain unidentified data quality issues or might be at risk of the misinterpretation of data.

We use the evaluation scale for metadata to assist researchers with the assessment of the extent of the metadata, in terms of completeness and availability. Although the results show that few data sets actually contained adequate metadata, these results also inform researchers (particularly authors of data sets) to what extent they need to improve the quality of the metadata, in terms of completeness and availability. We believe that when a data set is being created, it is important to provide sufficient information about the metadata, particularly for the metrics and the entities. This is to ensure that the data set has sufficient metadata to support the correct interpretation of data for analysis in research.

The results illustrated that data sets that contain more dataset category elements tend to have fewer identified and unidentified quality issues and a lower risk of misinterpretation than data sets that do not contain the dataset category elements that are essential in the identification of quality issues. This suggests that researchers (particularly the authors of data sets) should provide all of the dataset category elements that are
essential in the evaluation of the quality of data sets, thereby reducing the potential risk of quality issues and the misinterpretation of data.

We believe that the quality assessment process is useful not only for identifying quality issues and assessing the extent of metadata in data sets, but the assessment process also assists researchers in determining the information that is present in the data set to support the correct interpretation of data for analysis in research. Furthermore, the results of the quality assessment may allow researchers to choose which particular data sets to use for analysis in their research, according to the information contained in those data sets.

9.4 Evaluating part of the data quality framework

We conducted an online survey and an observational study to evaluate the effectiveness and usability of part of the framework. The part of the framework in question consists of the definitions of the dataset category elements and the formal definitions of data quality issues. In the following parts of this section, we discuss some interesting observations from the analysis of survey and the analysis of observational study.

9.4.1 Observations from the analysis of survey

First, the participants’ background and experience with data sets might have affect the ability of participants to apply the definitions of the dataset category elements to the data sets. When asked about their experience analysing data sets, the results of survey indicated that both categories of participants (academic researchers and PhD students) had experience in analysing data sets, particularly in CSV data sets. However, the participants’ responses in applying the definitions of the dataset category elements in the data sets indicated that the academic researchers responded correctly more often than PhD students. These results are likely to be related to knowledge and experience in general research. It could be argued that the PhD students in our survey did have experience in analysing data sets, but they may not have been concerned with measurements in data sets, particularly metrics and entities. Although we described some general information about measurements in data sets in the survey, they might have overlooked the information and made mistakes when applying the definitions of the dataset category elements in the data sets.

In addition, the participants’ ability to apply the definitions of the dataset category elements correctly in the data sets may be increase after applying them to the three different data sets in the survey. In particular, it was noted that many participants responded correctly to the questions about the tabular data set because the tabular data set was the third data set in the survey. We believe that some participants could have established an understanding of the definitions of the dataset category elements in the dataset metamodel after answering the questions for the CSV and ARFF data sets and therefore, it is possible that they could have answered all of the questions easily for the tabular data set.
However, a few participants commented that the definitions of the dataset category elements in the dataset metamodel were not easy to understand. They mentioned that they spent a certain amount of time understanding the definitions of the dataset category elements before applying the definitions to the data sets. We speculated that this was likely to be related to the background knowledge of the participants and their familiarity with the terms that we used in the dataset metamodel. It is possible that the participants had experience using the same terms in their research area where they may have had a different meaning to the terms’ meanings in our metamodel. For example, as mentioned in Chapter 4, the term ‘record’ can generally have a different meaning in different circumstances, such as in databases, programming languages and data sets.

Second, the participants’ background and experience with data sets might affect the ability of the participants to identify the quality issues in data sets. The results of the survey indicated that the participants with less experience analysing data sets were not able to identify the quality issues correctly. It is possible that these participants did use data sets for analysis in research, but they were not concerned with the quality of the data sets. In this case, it would have been difficult for them to identify the quality issues as a result of having no experience in the assessment of the quality of data sets.

Another possible explanation for this finding might be that some participants interpreted our formal definitions of data quality issues differently to what was originally intended. We suspected that they may have had experience in analysing different kinds of data sets other than software engineering data sets, such as data sets from databases. Such experience might have influenced the participants’ interpretation of our formal definitions of data quality issues and affected their assessment of data quality issues in the data sets. This suggests a need to improve the assessment approach for the identification of the quality issues in data sets, and illustrates a need to improve the evaluation of the quality of software engineering data sets in the future.

One last interesting observation from the survey: participants may have been put off by the amount of time required to complete the survey. We found that 49 participants expressed an interest in the survey by clicking the survey link and reading the participant information sheet on the first page of the survey. The participant information sheet contained detailed information about the survey, including the amount of time required to complete the survey, which was approximately 30 to 40 minutes. We suspect that some participants were not interested in answering the next section of the survey after reading the estimated time for the completion of the survey.

In the survey, out of the 49 participants responses, we considered 13 complete responses. A possible alternative explanation to the low number of respondents is that researchers feel threatened by the data quality framework. It is possible that researchers may feel insecure if the framework were to reveal that their data sets contain poor quality data. Quality issues with their data sets could affect the credibility of existing research results.

After the analysis of survey was conducted, we improved the definitions of the dataset category elements to capture a more precise meaning and to make them clearer to
understand. We then conducted an observational study to test the new definitions of dataset category elements and the formal definitions of data quality issues.

9.4.2 Observations from the analysis of observational study

First, we did not find the difference in experience analysing data sets had a significant influence on the effectiveness of our definitions of dataset category elements and the formal definitions of data quality issues. This was because we found that some participants, who have less than five years experience in research and did not have experience analysing data sets, had successfully identified the dataset category elements using the new definitions of dataset category elements and identified the quality issues in the data sets using the formal definitions of data quality issues. We suspected that they have gained a better understanding about the data quality framework through the examples of applying the definitions of dataset category elements and the formal definitions of data quality issues that we had provided in the first task as their guideline to answer the tasks list. Moreover, we also allowed participants to ask questions via the think aloud approach in the observational study. The results show that many participants did asked questions while performing the tasks.

Second, participants managed to perform the given tasks with reviewing the Dataset definitions sheet. The results show that all participants reviewed the Dataset definitions sheet in each task. This could be because they are unfamiliar with the terminology for dataset elements. We also suspected that they seems appeared to forget some of the definitions of dataset category elements (e.g. record) although they have read them in the first task. This could be some participants may have had experience in analysing different kinds of data sets other than software engineering data sets, such as data sets from machine learning repository. Such experience might have influenced the participants’ interpretation for the definitions of dataset category elements.

9.5 Implications

The research described in this thesis attempts to provide a standard way to understand the quality of data sets. As such, it explores several research implications, both regarding the framework itself and also with regard to other software engineering fields. This section elaborates on some interesting points from the research implications.

9.5.1 The state of quality for data sets in software engineering

With respect to the first research question in this thesis, it was found that very little research considered data quality in software engineering. In order to verify this finding again, we performed two simple searches for literature papers in the SCOPUS online database. The first search used the same search strings that we used previously in our systematic mapping study. We found 276 papers in the first search. The second search used the same search strings, but we excluded the phrases that related to data quality. We found 34,795 papers in the second search. By doing this, we concluded that only 0.007%
of papers paid attention to the quality of data sets in the research. This is consistent with our findings in the systematic mapping study that concluded that the research community is still not paying sufficient attention to the quality of data sets.

A possible explanation for this lack of interest in data quality could be that researchers are more interested in developing the best way to analyse the data than focusing on the underlying data with which they operate. We noticed that some researchers had developed robust data analysis techniques without actually considering the quality of the data sets that they used. We thus inferred that these researchers may not realise the impact of poor quality data on the conclusions of their research.

Nevertheless, some researchers have included a pre-processing step in the data analysis techniques to eliminate quality issues before analysing the data. We suspect that this step is useful for some common quality issues, such as missing and duplicate data, because these could be identified by analysing the structure of the data sets. However, we think that the pre-processing step might not be appropriate for quality issues that relate to the interpretation of data because this requires the analysis of metadata.

During our observation of data sets from the PROMISE repository, we found that very few of the data sets contained adequate metadata about the entities and metrics for the measurements. This indicates that researchers could easily misinterpret the intended meaning of the measurements in the data sets. It also raises an issue regarding the usefulness of these data sets for the research community: we believe that the data sets from the PROMISE repository are of limited use in research as a result of insufficient metadata.

9.5.2 Why should researchers care about data quality in software engineering?

Researchers should care about data quality issues in software engineering because they have to rely on data sets, particularly from public data repositories. We suspect that the data sets are not always of good quality because the researchers who are sharing the data sets do not always make it explicit what is contained in the data sets, possibly because they were fully dependent on software metrics tools that produced the data sets that they used in analysis. Therefore, the researchers may not be knowledgeable about the detailed methodology that generates measurements for the data sets and may be unfamiliar with the metrics that they used for analysis. For these reasons, we assume that some researchers are actually unaware of the essential things that should be reported in data sets.

Another possible reason for the low quality of data sets could be that public data repositories, such as PROMISE, do not have any standard guidelines for the publication of data sets. If researchers intend to share their data sets, they might need to explore the structure and content of existing data sets first and there is thus a strong possibility that they replicate the structure of the existing data sets in the repositories in their own data sets. This has been proven by the results from the modelling of data sets in Chapter 216.
9.5. IMPLICATIONS

4, which indicated that 87 of 92 data sets contained only labels, values and ancillary elements for the physical structure in the dataset metamodel.

The existing structure of data sets from data repositories could also be related to the software metrics tools that were used to produce the data sets. If researchers used the same software metrics tools, it is almost certain that they published data sets with the same structure. We suspect that many data sets from the PROMISE repository were produced by the same software metrics tools, although there is no detailed information about the software metrics tools in the data sets.

We noticed that most of the software metrics tools do not use a standard methodology to generate measurements, even though they are referring to the same metrics. We speculate that this is because there is no standard for measurement in software engineering, unlike the common SI metrics, and raise the question: ‘Why is there no standard for measurement in software engineering?’

A possible answer is that software engineering is a complex process that involves various kinds of methodologies, which are difficult to define in a valid and consistent metric to be practised in all domains. In addition, many existing software metrics have been defined in research and some of them do not capture whatever it is that they are supposed to measure. We speculate that some metrics that were defined for a particular software are unique (specifically constructed for that particular software system); therefore, we think it would be very useful if researchers could provide adequate metadata for the metrics that they have defined for that particular software.

9.5.3 How researchers use data sets in software engineering?

Researchers use data sets in several different ways. Some researchers create data sets from software systems for private data analysis. Some create and share the data sets with other researchers in the same research areas. Other researchers create the data sets and make the data sets available for sharing in public data repositories.

Our data quality framework is capable of evaluating the quality of the data sets that relate to software engineering, regardless of how researchers use the data sets in their research. This is because we developed our framework based on the observation of real data sets from data repositories and took into consideration almost every kind of data set that existed in the data repositories. However, we suspect that there are some aspects of data that are not being considered in the framework, especially for the data sets that are created for private use.

Researchers who create the data sets for private use usually design their own data structure. They may include some aspects of data that are important only for them to interpret the data sets and sometimes these metadata are not recorded anywhere, perhaps because these researchers prefer that the information relating to the data sets be confidential and not public. It is possible that some aspects of these kinds of data sets do not exist in the dataset metamodel; however, we do believe that researchers can model any elements in the data sets using our metamodel because it uses standard terminology that can be generalised in all cases.
Some researchers who create data sets and share these with other researchers do provide additional information about the data sets, such as metadata. We suspect that some metadata are not sufficient to support the interpretation of data because they might be defined in a particular way for a group of researchers. In this case, researchers can use our framework to model these data sets, particularly the metadata, to determine whether they contain a precise description for the metric or the entity. Furthermore, they can evaluate the quality of the metadata using our quality assessment process in the framework.

Researchers who create data sets and make these available for sharing in public data repositories intend to make their data sets available for wider use in the research community. To create new data sets, researchers should provide data and sufficient metadata to facilitate the correct interpretation of the data. For this particular purpose, researchers can use our guidelines for creating data sets that provide the best practices for the creation of good quality data sets.

In order to verify the created data sets before making them available in the data repositories, researchers can use our framework to evaluate the quality of the data sets. The framework will identify what information might be missing and also suggest ways to improve the quality of the data sets. Moreover, researchers who want to use the existing data sets from data repositories can also use our framework to determine whether or not the data sets are usable for their research purpose. In this case, they need to model the data sets and assess their quality. The framework will allow researchers to better understand the quality of the data sets and choose the appropriate data sets for their research.

9.5.4 Interpreting threats to validity and data quality

Poor data quality in data sets can be a threat to the validity of research results. This is evident in the number of studies in our systematic mapping study that reported that several common data quality issues affected their empirical results. These common data quality issues include duplicate data, missing data, inconsistent data and incorrect data. In the observation of real data sets, we found another kind of data quality issue that relates to the interpretation of data. This quality issue occurs due to insufficient metadata and may also affect the research results.

There are two typical threats to the validity of data identified in this research: conclusion validity and construct validity. We assume that the quality issue that relates to the interpretation of data can be classified as a threat to construct validity because it is concerned with the misinterpretation of data. For example, the metric LOC is associated with measurement values and there are no metadata that describe the meaning of LOC. In this example, we might interpret LOC as physical lines of code or physical lines of code excluding comments and non-blank lines. This clearly indicates that there is a potential for misinterpretation that will affect the research results.

We classified the common data quality issues such as missing data, duplicate data and incorrect data as threats to conclusion validity because these quality issues can occur
even though the metrics or entities for the measurement values were well defined in the metadata. These quality issues were likely to cause bias in the conclusions due to changes in the distribution of the analysed data.

9.6 Limitations

9.6.1 Limited to software engineering data sets

The data quality framework developed in this research can be applied to most software engineering data sets because we observed many different data sets from software engineering repositories to determine the different structures and formats that exist. As mentioned in Chapter 2, we focused on software engineering data sets because there is no standard for measurements in these and this makes them more likely to have data interpretation problems. However, many other kinds of data sets exist that do not have standards for measurements either, such as social science data sets. Further research efforts are required to understand these kinds of data sets in order to apply our data quality framework to them.

9.6.2 Manual modelling of data sets

The process for modelling data sets in the data quality framework was carried out manually. The automation of this process was not possible due to time constraints and it was not in the scope of this research. Additional effort would be required to build an automation process to model the elements of data sets using the dataset metamodel.

9.6.3 Sample size of survey

Another limitation in our survey was that we could not generalise the results to the entire population of computer science and software engineering researchers due to the small sample size. However, the findings from the survey helped us to gain insights into the effectiveness of the definitions of some elements of data sets, such as measurement values and metric labels. It also gave us an indication that these dataset category elements were easily identified in the data sets because they represented the common structure of data sets that only contain values and labels. Furthermore, we introduced other elements, such as record identifiers and metric metadata, which are important to facilitate the interpretation of data in data sets.

9.6.4 Minimal guideline in the observational study

In the observational study, we provided participants with a set of documents that consist of definitions of dataset category elements, formal definitions of data quality issues and two examples of applying the data quality framework. The two examples contain minimal explanation on how to apply the definitions of dataset category elements and how to apply the formal definitions of data quality issues. Although we encouraged participants to ask questions during the observational study, only few participants asked questions.
However, we can resolve this limitation by providing more detailed guidelines for the examples in order to assist participants in the study.

### 9.6.5 Incomplete evaluation of framework

The complete evaluation of the framework has not been carried out in a formal way. Although we demonstrated examples for each process in the framework in Chapter 4 and Chapter 5, we did not have enough opportunities to evaluate the complete framework during the course of this research. However, we conducted an evaluation of the usability and effectiveness of part of the framework using the online survey and the observational study. The results of the observational study gave us an indication that our framework was useful can be applied effectively and is usable for researchers.

### 9.7 Improvements on survey and observational study

In this section, we discuss how the survey and observational study could be improved to allow researchers to reuse our survey and observational questions in the evaluation of dataset quality.

There are three improvements that might be needed for anyone who wished to reuse the survey questions. First, the demographic questions could include more detailed questions about the participants’ experience in analysing different kinds of data sets (the existing questions focused on data sets from software engineering). We found that it was difficult to categorise the participants into particular groups of experience from the total responses. This suggests that we need to ask participants what kinds of data sets (e.g., data sets for databases, data sets for social sciences) they might have used in their research.

Second, we need to highlight the importance and benefit of answering all of the questions in the survey. We discovered that some participants answered only the demographic questions and so these were considered incomplete responses. We believe that the results could have been more significant if we had a larger sample size for the completed responses.

Lastly, another improvement on the survey is related to the brief description of the dataset metamodel in section 2. We suspect that this information was not enough to attract the participants’ attention and encourage them to answer the questions related to the framework. It would be useful if we could demonstrate an example identifying the elements and the quality issues in the data set.

For administering the observational study, we suggest two improvements to reproduce the observational study. Firstly, a complete guideline for the examples of applying the data quality framework should be provided to participants before commencing the observational study. We believe that participants will refer to the examples when performing the tasks given in the study.

Second, the observational study was limited to test two common structures of data sets that widely used in research. We suggest to add another session with participants
who have more than ten years experience to test data sets that have complex structures.

9.8 Conclusion

In this chapter, we have discussed the answers to our research questions and the observations from the evaluation of the dataset metamodel and the quality assessment process. We have also discussed some interesting points from the evaluation approach for part of the data quality framework. The interesting observations were related to the real data sets in data repositories and the participants’ background and experience in analysing data sets. We have also discussed some points of view for our research perspectives, the limitations of our research and possible improvements for the survey and the observational study.

In the next chapter, we summarise the thesis and the contributions of our research. We also suggest areas for future research and make the final concluding remarks.
This chapter concludes this thesis by presenting a summary of the work carried out in supporting the thesis proposition. It discusses the overall results and the contributions made by this research. This chapter also suggests some future work to extend the research.

### 10.1 Research Summary

Researchers working with data sets in empirical software engineering need to understand the quality of the data that form the basis of the results. To facilitate this type of research, the meaning of the data must be able to be interpreted correctly. Data sets contain measurements that are associated with metrics and entities; however, in some data sets, it is not always clear which entities have been measured and exactly which metrics have been used. This means that measurements could be misinterpreted. In particular, some data sets from public data repositories contain data that may not be able to be interpreted correctly due to insufficient metadata in the data sets. This is a critical problem for research based on data sets, which may lead to the misinterpretation of data and incorrect conclusions. To ensure the credibility of the results obtained from the use of data sets, the quality of the data must first be examined.

One goal of our research was to help researchers understand the quality of data sets by developing a framework for quality evaluation. The framework incorporated a dataset metamodel to allow a common interpretation of the description of the structure of data sets, as well as an assessment process to evaluate quality. The research also succeeded in filing gaps that we found in software engineering research, such as the need for standard terminology to describe the structure of a data set and the need for formal definitions of data quality issues in data sets.

The research was carried out in six steps. First, we reviewed the literature on data quality in empirical software engineering. Second, we described information about data sets. Third, we designed a dataset metamodel that described the structure and contents
of data sets. Fourth, we developed a framework for data quality assessment. Fifth, we constructed guidelines for the creation of good quality data sets. Finally, we evaluated part of our framework by using an online survey and an observational study. A summary of the research and the key findings at each step is provided in the following sections.

10.1.1 A review of data quality research in software engineering

The aim of the literature review, presented in Chapter 3, was to understand the importance of data quality and to investigate to what extent the research community had addressed this topic. The review consisted of three parts: a review of the systematic literature reviews (SLR) on data quality research, a systematic mapping study on data quality and a review of the data quality issues in software engineering data sets.

The previous SLRs [11–13] investigated the current state of research conducted on data quality in empirical software engineering. The reviews revealed a clear research gap: the research community is not paying sufficient attention to the quality of data sets in empirical research. We conducted a systematic mapping study to update the status of the reviews and focused on the extent to which the research considered the impact of data quality issues on the empirical results. The mapping study revealed the same gap as the reviews and found an additional research gap: many studies used unclear and inconsistent terminology to report problems with data quality in data sets.

Drawing on this second research gap, we reviewed the literature on the definition of data quality issues, examples of data quality issues and how to deal with them. The review of the definition of data quality issues highlighted an important issue: many different definition were being used to describe data quality issues. This issue was problematic for the research community when it came to assessing the quality of data sets because there were no standard definition of data quality issues. Different terminology was being used to define data quality issues and this terminology could be interpreted in many different ways.

The review of the examples of data quality issues indicated that a number of studies considered quality issues such as missing data, duplicate data or inconsistent data; however, these studies did not pay much attention to one quality issue in particular: insufficient metadata. This quality issue relates to the interpretation of data in data sets, for which insufficient metadata may lead to the misinterpretation of data and incorrect conclusions.

Our review of the approaches to dealing with poor quality data sets indicated that there was a lack of research looking at the role of metadata in the assessment of the quality of data sets. This issue was deemed to be critical for the research community because the research results based on data sets depend on the metadata to provide the meaning of the data. The research results would be more meaningful if the metadata were documented in a standard way to allow researchers to apply a common interpretation of the data.

As a first step in solving the data quality issues, we proposed the development of a standard terminology to describe the structure of data sets. We believe that creating a
common understanding of the terminology and concepts in data sets is an essential step in the construction of formal definitions of data quality issues. For the definitions to be useful for research practice, the terminology needs to be clear and consistent and the relationships between the data set concepts need to be explicitly represented.

10.1.2 Examining the content of data sets

The second step of our research, described in Chapter 2, consisted of a description of the information about data sets that aimed to identify the various components of the data. The information about data sets can be divided into four parts: general information about data sets, examples of data sets, software engineering data sets and software metrics tools. The general information about data sets explained the typical content and structure of a data set.

We presented the information about the variety of formats and structures of data sets by using examples of real and artificial data sets. Using the examples, we described how to model some aspects of the data sets and pointed out some potential quality issues that we found. In addition, we provided detailed information about software engineering data sets and the software metrics tools that produced these.

The observations of the data set examples highlighted a number of findings. We found that two of five real data sets contained only values and labels, while three of five data sets contained metadata and other information. Some of these data sets stored the metadata in external locations, thereby making the interpretation of the data difficult.

We also found that data sets use a variety of formats and structures to represent their content; however, there are some aspects of data that often appear in the same way across the different structures of data sets. This indicated that these aspects of data could be easily identified across the different structures of data sets, motivating us to develop a metamodel with standard terminology to describe the structure of data sets.

10.1.3 Designing a dataset metamodel

In the third step of our research, described in Chapter 4, we proposed a dataset metamodel as a standard way to describe the structure of a data set. It classifies the data set elements into three levels: physical structure, dataset category and dataset concepts. Further, the metamodel describes the relationship between each level.

We introduced a process for data set modelling by applying the dataset metamodel using a formal procedure. The process will help researchers to understand the contents of data sets and to identify any missing information. This process can also indicate any potential risks of misinterpretation of the data due to the limitations of metadata.

We applied the dataset modelling process to the existing data sets from public data repositories. The results from the modelling process showed that many real data sets (43 out of 92) contain only values, labels and ancillary for physical structure elements. We suspected that these data sets could be misinterpreted due to insufficient metadata. Further, the results of the modelling process also showed that the data sets could be improved by indicating which dataset category elements were missing.
10.1.4 Developing a data quality assessment framework

In the fourth step of our research, we developed a data quality assessment framework to determine whether a data set contains adequate metadata to facilitate the correct interpretation of data for analysis in empirical research. The framework described in Chapter 5 consisted of two processes: dataset modelling and quality assessment. In this step, we focused on the quality assessment process, which includes four steps: assessing the dataset model, identifying the data quality issues, evaluating the metadata and preparing the assessment report.

We constructed formal procedures for every task in the quality assessment process. We introduced formal definitions of data quality issues to identify these, as well as an evaluation scale to evaluate the quality of the metadata in data sets. We demonstrated the purpose of each task in the quality assessment process by using examples from real data sets. We presented the results of the quality assessment process in two types of assessment report: a summary report and a full report.

We applied the quality assessment process to the existing data sets from public data repositories. The results from step two in the quality assessment process showed that many real data sets that contain only values, labels and ancillary elements have metadata-related quality issues: incomplete metadata for a metric (69 out of 92) and incomplete metadata for an entity (79 out of 92). The results of step three in the quality assessment process indicated that 70 out of 92 data sets were assigned 0 stars for rating metric metadata and entity metadata because of the absence of metadata. We suspected that these data sets carried risks of misinterpretation.

10.1.5 Constructing guidelines for creating good-quality data sets

In the fifth step of our research, we constructed guidelines for the creation of good quality data sets to help researchers improve the quality of their data sets. These guidelines were described in Chapter 6. The guidelines were prepared for researchers who are the creators of data sets and who intend to share the data sets with the research community. They were also intended for researchers who want to know the quality of real data sets from public data repositories.

The guidelines were constructed based on the essential terminology from the dataset metamodel and the procedures from the data quality assessment framework. In the guidelines, we provided instructions on how to define the structure and contents of a data set. We also provided details on how to describe adequate metadata for the data set and presented two ways to apply the guidelines to the data sets. The practical use of the guidelines was demonstrated by applying them to the examples of real data sets from Chapter 2.

10.1.6 Performing a user evaluation of a data quality framework

In the final step of our research, we presented an evaluation of part of the data quality framework by conducting an online survey (described in Chapter 7) and an observational
study (described in Chapter 8).

We conducted a survey as our preliminary evaluation to evaluate the effectiveness of definitions of dataset category elements and the usability of formal definitions of data quality issues. Although the number of participants who responded was small, the survey results indicate that participants with relevant background knowledge and experience in research, particularly in analysing data sets, were able to apply the definitions of the dataset category elements and the formal definitions of data quality issues to the data sets successfully. The survey results also indicate that there is a need to improve the definitions of dataset category elements. After the analysis of survey was conducted, we improved the definitions of the dataset category elements to capture a more precise meaning and to make them clearer to understand.

We then conducted an observational study to test the new definitions of dataset category elements and the formal definitions of data quality issues. In the observational study, we aim to observe how researchers apply the definitions of dataset category elements and the formal definitions of data quality issues. The goal of the observational study was same as that of the previous survey; that is, to evaluate the effectiveness and usability of part of the framework for data quality assessment. However, the research questions of the observational study were different to those of the survey, because we used different methods. We used observation, a questionnaire and think aloud as our observational study methods to provide further insights into the framework through participant thought processes while applying the part of the framework. From the observational study analysis, we conclude that most participants successfully applied the definitions of dataset category elements and the formal definitions of data quality issues to the datasets. In addition, the analysis reveals that most participants successfully identified and evaluated metadata in the datasets.

10.2 Research Contributions

Our research has demonstrated that a data quality assessment framework designed for software engineering data sets succeeds in identifying whether there is adequate metadata to facilitate the interpretation of data when evaluating dataset quality. We have made four major contributions at the doctoral level, which are summarised as follows:

1. We conducted a systematic mapping study on data quality research and developed a research data life cycle. The systematic mapping study identified the use of unclear and inconsistent terminology in reporting problems with dataset quality issues. It also identified the fact that there are no standard definitions of data quality issues in the data sets. The research data life cycle describes the stages a data set may go through and can be used to identify when data quality issues may be introduced.

2. We designed a dataset metamodel to describe the structure of and concepts in a data set, as well as the relationship between each concept. We defined each concept
using standard terminology to allow for a common interpretation by researchers and a better understanding of the contents of data sets.

3. We developed a data quality assessment framework to help researchers understand the quality of data sets. The framework provides researchers with a systematic approach to modelling the data sets and evaluating the quality of these.

4. We provided formal guidelines for the creation of complete and good quality data sets as there are no existing standard guidelines. The guidelines provide a clear insight into and a structured approach to the development of a data set from scratch, based on the essential elements and their relationships, using the dataset metamodel.

10.3 Future Work

During the course of this research, several possible directions for future investigation have been identified. Some of these directions include overcoming the current limitations of this study and unifying the terminology of data sets into an ontology.

10.3.1 Evaluating other types of data sets

As discussed, our approach to understanding data sets from public data repositories has provided insights and has allowed us to develop a framework for data quality assessment. This framework has been successfully applied to the assessment of software engineering data sets. It is therefore believed that the application of this approach could be beneficial to other types of data sets. Further research should be aimed at extending the metamodel’s capability to model other types of data sets and refining the quality assessment process to be able to assess these.

One of the types of data set that could potentially be modelled by our metamodel are data sets for machine learning research (data sets from the UCI Machine Learning Repository). These data sets consist of values (in a data file), attribute and metadata (in an external file). Such data sets would be compatible with the dataset metamodel that is built on the common structure of data sets. The extension of the metamodel’s capability would enhance the framework’s interoperability and portability across different types of data sets.

10.3.2 Automated modelling of the data sets

Although this thesis has demonstrated the process for modelling a data set, this process was carried out manually in our research. The application of the dataset metamodel to the existing data sets in the PROMISE repository shows that it is possible to model data sets using our dataset metamodel. However, this involves developing a program that is able to transform the existing modelling processes into an automated modelling process that uses the physical structure of the data set to generate the dataset model. The modelling needs to capture the definition of each element in the dataset metamodel.
in the automated program. By successfully automating this modelling process, it is hoped that dataset models could be successfully generated.

10.3.3 To unify dataset terminology into an ontology

Metamodels and ontologies appear to be used for similar purposes, except that an ontology focuses on specific domain knowledge and metamodels focus on conceptual definitions. In our research, our metamodel focused on software engineering data sets. Thus, it would be interesting to unify the dataset terminology into an ontology for data sets in general. The idea behind this ontology would be to provide a core terminology that focuses on the metadata used to evaluate data set quality. Having this metadata available in the data sets would enable researchers to understand data set quality more efficiently and consistently.

10.3.4 To improve and refine the quality assessment process in the data quality framework

At present, our quality assessment process consists of four steps that are carried out manually. Each step requires researchers to execute a set of formal procedures. In the future, we could potentially refine the steps to provide more simplified and efficient procedures for researchers to follow. Furthermore, we could also automate some of the steps in the quality assessment that do not require any subjective evaluation (for example, identifying data quality issues, such as duplicate data and inconsistent data). The refining of the procedures and the automation should help to improve the usability of the data quality framework in the evaluation of data set quality.

10.4 Final Remarks

This research project developed from the need to evaluate the quality of data sets that form the basis of empirical results. If the quality is poor, the results of the analysis will be questionable. Research has indicated that quality issues exist in data sets, particularly data sets from public data repositories. These quality issues are concerned with the accuracy of data, such as missing data, duplicate data, inconsistent data and incorrect data.

Apart from the quality issues, data sets also have an issue with the misinterpretation of data. This issue arises due to insufficient metadata in the data sets that can lead to the misinterpretation of the data and affect the analysis results. We thus saw a need to help researchers to correctly interpret the data sets and to identify what potential problems a data set could have. The focus of our research was as follows: to develop a framework for data quality assessment that determines whether a data set has sufficient information to support the correct interpretation for analysis in empirical research. The framework incorporates a dataset metamodel to allow for a common interpretation in describing the structure of the data set and a quality assessment process to evaluate the data set quality.
This thesis provides the results of the development of a framework for data quality assessment and presents a practical way to understand the quality of data sets. It also provides guidelines for the creation of good quality data sets that can be used to improve data set quality in the future.
A.1 Summary of search results from online databases
### Summary of Search Results from Databases

<table>
<thead>
<tr>
<th>Database name</th>
<th>Search strategy</th>
<th>Date searched</th>
<th>Years covered</th>
<th>No. of search results</th>
<th>No. of relevant articles</th>
<th>Duplicates found</th>
<th>Fields searched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scopus</td>
<td>TITLE-ABS-KEY((software OR system OR application OR program) AND (method OR technique OR approach) AND (metric OR measure) AND (&quot;data quality&quot; OR &quot;noisy data&quot; OR &quot;noisy data set&quot; OR &quot;noisy dataset&quot; OR &quot;noise data set&quot; OR &quot;noise dataset&quot; OR &quot;noise data&quot; OR &quot;inconsistent pieces&quot; OR &quot;inconsistent value&quot; OR &quot;inconsistent data&quot; OR &quot;erroneous data&quot; OR &quot;erroneous value&quot; OR &quot;clean data&quot; OR &quot;data cleaning&quot; OR &quot;empirical data&quot;) AND PUBYEAR &gt; 2007 AND SUBJAREA(COMP) AND ( LIMIT-TO(LANGUAGE,&quot;English&quot; ) ) AND ( EXCLUDE(SUBJAREA,&quot;MATH&quot; ) OR EXCLUDE(SUBJAREA,&quot;SOCI&quot; ) OR EXCLUDE(SUBJAREA,&quot;DECI&quot; ) OR EXCLUDE(SUBJAREA,&quot;MEDI&quot; ) OR EXCLUDE(SUBJAREA,&quot;PHYS&quot; ) OR EXCLUDE(SUBJAREA,&quot;BIOC&quot; ) OR EXCLUDE(SUBJAREA,&quot;MATE&quot; ) OR EXCLUDE(SUBJAREA,&quot;BUSI&quot; ) OR EXCLUDE(SUBJAREA,&quot;EART&quot; ) OR EXCLUDE(SUBJAREA,&quot;ENVI&quot; ) OR EXCLUDE(SUBJAREA,&quot;HEAL&quot; ) OR EXCLUDE(SUBJAREA,&quot;CHEM&quot; ) OR EXCLUDE(SUBJAREA,&quot;NEUR&quot; ) OR EXCLUDE(SUBJAREA,&quot;CENG&quot; ) OR EXCLUDE(SUBJAREA,&quot;ECON&quot; ) OR EXCLUDE(SUBJAREA,&quot;ENER&quot; ))</td>
<td>3/12/12</td>
<td>2008 to 2012</td>
<td>139</td>
<td>7</td>
<td>1</td>
<td>Title and abstract</td>
</tr>
<tr>
<td>Database</td>
<td>Query</td>
<td>Date</td>
<td>Year Range</td>
<td>Results</td>
<td>Categories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>--------</td>
<td>------------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science Direct</td>
<td>pub-date &gt; 2006 and TITLE-ABSTR-KEY(software OR system OR application OR program) AND TITLE-ABSTR-KEY(method OR technique OR approach) AND TITLE-ABSTR-KEY(metric OR measure) AND TITLE-ABSTR-KEY(&quot;data quality&quot; OR &quot;noisy data&quot; OR &quot;noisy data set&quot; OR &quot;noisy dataset&quot; OR &quot;noise data set&quot; OR &quot;noise dataset&quot; OR &quot;noise data&quot; OR &quot;inconsistent pieces&quot; OR &quot;inconsistent value&quot; OR &quot;inconsistent data&quot; OR &quot;erroneous data&quot; OR &quot;erroneous value&quot; OR &quot;clean data&quot; OR &quot;data cleaning&quot; OR &quot;empirical data&quot;) [All Sources(Computer Science)]</td>
<td>5/12/12</td>
<td>2008 to 2012</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>Title and abstract</td>
</tr>
<tr>
<td>SpringerLink</td>
<td>(software OR program OR system OR application) AND (&quot;technique&quot; OR &quot;method&quot; OR &quot;approach&quot;) AND (&quot;metric&quot; OR &quot;measure&quot; OR &quot;measurement&quot;) AND (&quot;data quality&quot; OR &quot;noisy data&quot; OR &quot;noisy dataset&quot; OR &quot;noisy data set&quot; OR &quot;noise data set&quot; OR &quot;noise dataset&quot; OR &quot;noise&quot; OR &quot;inconsistent pieces&quot; OR &quot;inconsistent value&quot; OR &quot;inconsistent data&quot; OR &quot;erroneous data&quot; OR &quot;erroneous value&quot; OR &quot;clean the data&quot; OR &quot;data cleaning&quot; OR &quot;empirical data&quot; OR &quot;not very consistent&quot;)</td>
<td>10/12/12</td>
<td>2008 to 2012</td>
<td>102</td>
<td>3</td>
<td>1</td>
<td>All fields</td>
</tr>
</tbody>
</table>
A.2 List of included studies in systematic mapping
List of studies in systematic mapping review (2008-2012).


List of studies in systematic mapping review (2013-2016).


A.3 Data extracted from S7
<table>
<thead>
<tr>
<th>Study ID</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Dealing with noise in defect prediction.</td>
</tr>
<tr>
<td>Author</td>
<td>Kim, S., H. Zhang, et al.</td>
</tr>
<tr>
<td>Year of Publication</td>
<td>2011</td>
</tr>
<tr>
<td>Study Setting</td>
<td>Industry</td>
</tr>
<tr>
<td>Type of paper</td>
<td>Propose a solution</td>
</tr>
<tr>
<td>Reference type</td>
<td>Conference</td>
</tr>
<tr>
<td>Research Problem</td>
<td>Data collected from mining software repositories contain noises because current defect collection practices are based on optional bug fix keywords or bug report links in change logs.</td>
</tr>
<tr>
<td>Objective / Aim</td>
<td>Propose approaches to deal with the noise in defect data</td>
</tr>
<tr>
<td>Contribution</td>
<td>1. Noise resistance measuring technique</td>
</tr>
<tr>
<td></td>
<td>2. Empirical study of measuring noise resistance</td>
</tr>
<tr>
<td></td>
<td>3. Noise detection technique - A new method called CLNI for identifying noisy instances in defect data</td>
</tr>
<tr>
<td>Domain research</td>
<td>Defect prediction</td>
</tr>
<tr>
<td>Data Quality Problem</td>
<td>Noisy</td>
</tr>
<tr>
<td>What definition of data quality problem was used?</td>
<td>Noisy data are found in the extracted defect information from change logs and bug reports.</td>
</tr>
<tr>
<td>Examples of data quality problem</td>
<td>For example, for the Eclipse SWT component, there are 32% unlinked bugs (bugs that do not reflect in the CVS logs) in Eclipse 3.0 and 21% unlinked bugs in Eclipse 3.1. The existence of the unlinked bugs indicates that the defect data collected via MSR is noisy. We also noticed that the noise level decreased in Eclipse 3.4, where 92.27% SWT bugs are recorded in CVS logs.</td>
</tr>
<tr>
<td>What is the kind of data used?</td>
<td>Defect data</td>
</tr>
<tr>
<td>What is the type of data set used?</td>
<td>Open source</td>
</tr>
<tr>
<td>List the data sets</td>
<td>Eclipse, Debug</td>
</tr>
<tr>
<td>What method/technique was used?</td>
<td>1. Bayes Net classifier</td>
</tr>
<tr>
<td></td>
<td>2. A method to measure noise resistance in software defect prediction</td>
</tr>
<tr>
<td>Any software metrics / predictors used?</td>
<td>NA</td>
</tr>
<tr>
<td>Any pre-processing process to clean the data sets?</td>
<td>None</td>
</tr>
<tr>
<td>DQ Dimension</td>
<td>Accuracy, Incomplete</td>
</tr>
<tr>
<td>RDLC</td>
<td>Process</td>
</tr>
</tbody>
</table>
B.1 Dataset metamodel
C.1 Formal procedures for modelling data set

Table C.1: Procedure for identifying the physical structure elements.

<table>
<thead>
<tr>
<th>No.</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Identify <em>label</em>. Colour the <em>label</em> blue.</td>
</tr>
<tr>
<td>2.</td>
<td>Identify <em>values</em>. Colour the <em>values</em> light green.</td>
</tr>
<tr>
<td>3.</td>
<td>Identify elements that represent <em>metadata</em>. Colour the <em>metadata</em> elements dark grey.</td>
</tr>
<tr>
<td>4.</td>
<td>Identify elements that represent <em>ancillary</em>. Colour the <em>ancillary</em> elements purple.</td>
</tr>
<tr>
<td>5.</td>
<td>Check that all elements in the Dataset sheet have been coloured. Then, look in the WebpageDataset sheet. Identify elements that represent <em>metadata</em>. Colour the <em>metadata</em> elements dark grey.</td>
</tr>
</tbody>
</table>
C.1. FORMAL PROCEDURES FOR MODELLING DATA SET

Table C.2: Procedure for interpreting the physical structure elements.

<table>
<thead>
<tr>
<th>No.</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent measurement values. Colour the measurement values green.</td>
</tr>
<tr>
<td>2.</td>
<td>In the <strong>Dataset</strong> sheet, look for <strong>labels</strong> (already coloured blue) that correspond to metric labels. Check whether the <strong>labels</strong> are associated with the measurement values. If yes, they are metric labels. If no, check whether there are metadata (already coloured with dark grey) that describe a metric and are associated with the labels. If yes, the <strong>labels</strong> are metric labels. Colour the metric labels red.</td>
</tr>
<tr>
<td>3.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent metric metadata. Colour the metric metadata dark blue.</td>
</tr>
<tr>
<td>4.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent record identifiers. Colour the record identifiers light blue.</td>
</tr>
<tr>
<td>5.</td>
<td>In the <strong>Dataset</strong> sheet, look for <strong>labels</strong> (already coloured blue) that represent entity labels. Check whether the <strong>labels</strong> are associated with the record identifiers. If yes, they are entity labels. If no, check whether there are metadata (already coloured dark grey) that describe an entity and are associated with the labels. If yes, the <strong>labels</strong> are entity labels. Colour the entity labels light green.</td>
</tr>
<tr>
<td>6.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent entity metadata. Colour the entity metadata pink.</td>
</tr>
<tr>
<td>7.</td>
<td>In the <strong>Dataset</strong> sheet, identify values (already coloured light green) that represent ancillary values. Colour the ancillary values blue.</td>
</tr>
<tr>
<td>8.</td>
<td>In the <strong>Dataset</strong> sheet, look for <strong>labels</strong> (already coloured blue) that represent ancillary labels. Check whether the <strong>labels</strong> are associated with the ancillary values. If yes, they are ancillary labels. If no, check whether there are metadata (already coloured dark grey) that describe a property that is not a metric label or entity label and are associated with the <strong>labels</strong>. If yes, the <strong>labels</strong> are ancillary labels. Colour the ancillary labels purple.</td>
</tr>
<tr>
<td>9.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metadata (already coloured dark grey) that represent ancillary metadata. Colour the ancillary metadata light grey.</td>
</tr>
<tr>
<td>10.</td>
<td>In the <strong>Dataset</strong> sheet and the <strong>WebpageDataset</strong> sheet, look for metric metadata (already coloured dark blue) that represent data type metadata. If the data type metadata is present, draw a brown rectangle to highlight the data type metadata. Next, look for metadata (already coloured dark grey) that represent data type metadata. Draw a brown rectangle to highlight the data type metadata.</td>
</tr>
<tr>
<td>11.</td>
<td>In the <strong>Dataset</strong> sheet, identify records. Draw an orange rectangle to highlight the records. In each record, look for record identifier (already coloured light blue). If the record identifier is present, ensure each record has a unique record identifier. If there are two or more records with the same record identifiers, then these records potentially have quality issues. Draw a red rectangle to highlight these records.</td>
</tr>
</tbody>
</table>
D.1 Matrix for applying the modelling data sets process
### Modelling the data set process

<table>
<thead>
<tr>
<th>No.</th>
<th>Measurement data set (count of similar data sets)</th>
<th>Value</th>
<th>Ancillary Label</th>
<th>Metadata Record</th>
<th>Measurement value</th>
<th>Metric label</th>
<th>Entity label</th>
<th>Ancillary label</th>
<th>Record identifier</th>
<th>Ancillary value</th>
<th>Data type metadata (X/Y)</th>
<th>Entity metadata (X/Y)</th>
<th>Ancillary metadata (X/Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PROMISE AM1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>4/18</td>
<td>12/18</td>
<td>1/18</td>
<td>13/18</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Qualitas Weka 3.7.5</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>16/26</td>
<td>11/26</td>
<td>2/26</td>
<td>7/26</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PROMISE Nas93</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>20/24</td>
<td>24/24</td>
<td>1/24</td>
<td>3/24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>PROMISE ISBSG10</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>92/94</td>
<td>94/94</td>
<td>1/94</td>
<td>1/94</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PROMISE Cosmic</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>84/86</td>
<td>86/86</td>
<td>1/86</td>
<td>1/86</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PROMISE Mozilla4</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>1/6</td>
<td>6/6</td>
<td>1/6</td>
<td>4/6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PROMISE Reuse</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>25/28</td>
<td>28/28</td>
<td>1/28</td>
<td>2/28</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PROMISE Miyazaki94</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>8/9</td>
<td>9/9</td>
<td>1/9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PROMISE KC2</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>19/22</td>
<td>22/22</td>
<td>1/22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>PROMISE Datatrieve</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>8/9</td>
<td>9/9</td>
<td>1/9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>PROMISE JM1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>19/22</td>
<td>22/22</td>
<td>1/22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>PROMISE Nickle</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>5/10</td>
<td>10/10</td>
<td>1/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>PROMISE Xfree</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>5/10</td>
<td>10/10</td>
<td>1/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>PROMISE Xorg</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>5/10</td>
<td>10/10</td>
<td>1/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>PROMISE Coc81-dem</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>4/27</td>
<td>27/27</td>
<td>1/27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>PROMISE Comosodr</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>24/25</td>
<td>25/25</td>
<td>1/25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>PROMISE Nas93-dem</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>4/27</td>
<td>27/27</td>
<td>1/27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>PROMISE MB2-silap211</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>5/26</td>
<td>26/26</td>
<td>1/26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>PROMISE MW1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>PROMISE KC3</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>40/40</td>
<td>40/40</td>
<td>1/40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>PROMISE Cooomos81</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>19/19</td>
<td>19/19</td>
<td>1/19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>PROMISE PCI1 - PC5 (5)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>PROMISE MC1-MC2 (2)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>39/39</td>
<td>39/39</td>
<td>1/39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>PROMISE Kitchenham</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>10/10</td>
<td>10/10</td>
<td>1/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>PROMISE Maxwell</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>27/27</td>
<td>27/27</td>
<td>1/27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>PROMISE Albrecht</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>8/8</td>
<td>8/8</td>
<td>1/8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>PROMISE Kemerer</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>8/8</td>
<td>8/8</td>
<td>1/8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>PROMISE China</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>19/19</td>
<td>19/19</td>
<td>1/19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>PROMISE NFR</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>3/3</td>
<td>3/3</td>
<td>1/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>PROMISE Bike</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>11/11</td>
<td>11/11</td>
<td>1/11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>PROMISE Coc81lnh</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>19/19</td>
<td>19/19</td>
<td>1/19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>PROMISE AR (7)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>30/30</td>
<td>30/30</td>
<td>1/30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>PROMISE CM1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>PROMISE MB2-silap433</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>PROMISE Midjedit</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>PROMISE PITSA - PITSF (6)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>PROMISE Ant to Zuzel (33)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>PROMISE Bugreport</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>PROMISE Boetticher</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>PROMISE NrpClassic (5)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>38/38</td>
<td>38/38</td>
<td>1/38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend**

- √: Element present in the data set.
- : Element not present in the data set.

**x/y Number of <element> present in the data set/ Total number of labels in the data set**
## E.1 Formal procedures for quality assessment

Table E.1: Step 1: Assess the data set

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the elements of dataset category that are present and not present in the model of the data set.</td>
<td>Search for dataset category elements that are present and not present in the model of the data set. List them.</td>
</tr>
<tr>
<td>2.</td>
<td>Describe the entities.</td>
<td>If record identifiers and entity label are present in the model of the data set, find entity metadata. If found, determine the entities. Describe them. If not found, make an assumption for the entities based on the record identifiers and entity label. Describe the entities.</td>
</tr>
</tbody>
</table>
### Table E.2: Step 2: Identify data quality issues

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Identify duplicate data</td>
<td>Search for identical record identifiers in the records. If not found, go to procedure No. 3. If found, check whether the measurement values for all metric labels are identical. If this is the case, then these records are duplicate data. List the record identifiers for these records.</td>
</tr>
<tr>
<td>2.</td>
<td>Identify inconsistent data</td>
<td>Search for identical record identifiers in the records. If not found, go to the next procedure. If found, check whether the measurement values for all metric labels are identical. If this is not the case, then these records are inconsistent data. List the record identifiers for these records.</td>
</tr>
<tr>
<td>3.</td>
<td>Identify missing data</td>
<td>Search for records that have either an empty string or null value for a given label. If not found, go to the next procedure. If found, list the record identifiers. If there is no record identifier, list the associated label.</td>
</tr>
<tr>
<td>4.</td>
<td>Identify incorrect data</td>
<td>Search for records that have implausible measurement values for a given metric label. If not found, go to the next procedure. If found, find data type metadata and check whether the implausible measurement values represent the legal measurement values for the metric label. If this is not the case, then these records contain incorrect data. For each record that contains incorrect data, list the record identifier. If there is no record identifier, list the associated label.</td>
</tr>
<tr>
<td>5.</td>
<td>Identify incomplete metadata for a metric</td>
<td>Search for metric labels that do not have metric metadata. If not found, go to the next procedure. If found, list the metric labels.</td>
</tr>
<tr>
<td>6.</td>
<td>Identify incomplete metadata for an entity</td>
<td>Search for entity labels that do not have entity metadata. If not found, go to the next procedure. If found, list the entity labels.</td>
</tr>
<tr>
<td>7.</td>
<td>Identify imprecise metadata for a metric</td>
<td>Search for metric metadata. If found, check whether the metric metadata explicitly describe the metric used to measure the attribute of an entity. If this is not the case, then this metric metadata contains imprecise metadata for the metric. List the metric labels. If metric metadata is not found, this data set has a risk of misinterpretation.</td>
</tr>
<tr>
<td>8.</td>
<td>Identify imprecise metadata for an entity</td>
<td>Search for entity metadata. If found, check whether the entity metadata explicitly describe the property whose values distinguish between entities. If this is not the case, then this entity metadata contains imprecise metadata for the entity. List the entity label. If entity metadata is not found, this data set has a risk of misinterpretation.</td>
</tr>
</tbody>
</table>
Table E.3: Step 3: Evaluate the metadata of data set

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedure</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the entity metadata. Evaluate.</td>
<td>Find the entity metadata that describe the entity label. Determine whether the entity metadata explicitly describe the meaning of the property of the entity. Identify the location of metadata. Assign and list a rating.</td>
</tr>
<tr>
<td>2.</td>
<td>List the metric metadata. Evaluate.</td>
<td>Find the metric metadata that describe the metric label. Determine whether the metric metadata explicitly describe the meaning of the attribute of the entity. Identify the location of metadata. Assign and list a rating.</td>
</tr>
<tr>
<td>3.</td>
<td>Evaluate the overall metadata rating. List.</td>
<td>Review the ratings for entity and metric metadata. Assign the minimum rating to indicate the overall rating of the metadata. List.</td>
</tr>
</tbody>
</table>

Table E.4: Step 4: Prepare the assessment report

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedures</th>
<th>Type of report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>List the dataset category elements that are present in the model of the data set.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>2.</td>
<td>List the dataset category elements that are not present in the model of the data set. Describe them.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>3.</td>
<td>For each dataset category elements that are not present in the model of the data set, describe suggestions to improve the quality of the data set.</td>
<td>Full</td>
</tr>
<tr>
<td>4.</td>
<td>List identified and unidentified data quality issues. Describe them.</td>
<td>Full and Summary</td>
</tr>
<tr>
<td>5.</td>
<td>Describe the overall results for the metadata evaluation scale.</td>
<td>Summary</td>
</tr>
<tr>
<td>6.</td>
<td>Describe the rating of each element of the metadata, and the overall rating.</td>
<td>Full</td>
</tr>
</tbody>
</table>
F.1 Quality assessment for Ant 1.3 data set
## Dataset Name: Ant 1.3

### Instructions for completion:
Answer the following questions. You can refer to the Example & Dataset Model, Modeling Process, Dataset (1), Dataset (2) and WebpageDataset(2). For each process, write the relevant description in the appropriate box if applicable.

### 1. Assess the data set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedures</th>
<th>Task to perform (advice)</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List the elements of dataset category that are present and not present in the model of the data set.</td>
<td>Search for dataset category elements that are present and not present in the model of the data set. List them.</td>
<td>Elements that are present</td>
</tr>
<tr>
<td>2</td>
<td>Describe the entities.</td>
<td>Search for dataset category elements that are present in the model of the data set.</td>
<td><strong>No entity metadata in the data set.</strong> The record identifiers represent the conventions for class names in the Ant project. <strong>We assume the entities are the classes of Ant project.</strong></td>
</tr>
</tbody>
</table>

### 2. Identify data quality issues in the data set.

#### Search for metric labels that do not have metric metadata.

- If not found, go to the next procedure. If found, list the metric labels.

#### Search for record identifiers and entity labels.

- If found, determine the entities. If not found, make an assumption for the entities based on the record identifiers and entity label. Describe the entities.

#### Search for record identifiers.

- If not found, make an assumption for the classes of Ant project.

### 3. Evaluate the metadata of data set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedures</th>
<th>Task to perform (advice)</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List the entity metadata.</td>
<td>Find the entity metadata that describe the entity label. Determine whether the entity metadata explicitly describe the meaning of the property of the entity. Identify the location of metadata. Assign and list a rating.</td>
<td>No star</td>
</tr>
<tr>
<td>2</td>
<td>List the metric metadata.</td>
<td>Find the metric metadata that describe the metric label. Determine whether the metric metadata explicitly describe the meaning of the attribute of the entity. Assign and list a rating.</td>
<td>No star</td>
</tr>
<tr>
<td>3</td>
<td>Evaluate the overall metadata rating.</td>
<td>Review the ratings for entity and metric metadata. Assign the minimum rating to indicate the overall rating of the metadata.</td>
<td>No star</td>
</tr>
</tbody>
</table>

#### Scale of metadata criteria.

<table>
<thead>
<tr>
<th>Score</th>
<th>Criteria</th>
<th>Accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Star</td>
<td>Metadata are not visible except labels available to represent the property or attribute of entity.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>1 Star</td>
<td>Metadata are visible, but do not explicitly describe the meaning of the property or attribute of entity.</td>
<td></td>
</tr>
<tr>
<td>2 Star</td>
<td>Metadata are visible, explicitly describe the meaning of the property or attribute of entity, but are not precise.</td>
<td></td>
</tr>
<tr>
<td>3 Star</td>
<td>Metadata are visible, explicitly describes the meaning of the property or attribute of entity, and are precise.</td>
<td></td>
</tr>
</tbody>
</table>

### 4. Prepare the assessment report.

<table>
<thead>
<tr>
<th>No.</th>
<th>Procedures</th>
<th>Type of report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List the dataset category elements that are present in the model of the data set.</td>
<td>Full and summary</td>
</tr>
<tr>
<td>2</td>
<td>List the dataset category elements that are not present in the model of the data set.</td>
<td>Full and summary</td>
</tr>
<tr>
<td>3</td>
<td>For each dataset category elements that are not present in the model of the data set, describe suggestions to improve the quality of the data set.</td>
<td>Full</td>
</tr>
<tr>
<td>4</td>
<td>Identify and unidentified data quality issues.</td>
<td>Full and summary</td>
</tr>
<tr>
<td>5</td>
<td>Describe the overall results for the metadata evaluation scale.</td>
<td>Summary</td>
</tr>
</tbody>
</table>

| Summary | |
G.1 Summary report for Data Quality Assessment Framework
<table>
<thead>
<tr>
<th>Assessment results: Summary report</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elements of dataset category that are present in the model of data set</strong></td>
</tr>
<tr>
<td>Entity label, metric label, anciallir label, measurement values, record identifiers, ancillary values and record.</td>
</tr>
</tbody>
</table>

| **Elements of dataset category that are not present in the data set** | Description |
| Metric metadata, entity metadata and ancillary metadata. |

We describe the details of dataset categories that are not present in the model of data set as follows.

<table>
<thead>
<tr>
<th>Element of dataset category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity metadata</td>
<td>1 entity label does not have entity metadata. The entity label is ‘name’.</td>
</tr>
<tr>
<td>Metric metadata</td>
<td>21 metric labels do not have metric metadata. The metric labels are ‘wmc, dit, noc, cbo, rfc, lcom, ca, ce, npm, lcom3, loc, dam, moa, mfa, cam, ic, cbm, amc, avg_cc, max_cc, bug’.</td>
</tr>
<tr>
<td>Ancillary metadata</td>
<td>2 ancillary labels do not have ancillary metadata. The ancillary labels are ‘name’ and ‘version’.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Result for identifying data quality issues</strong></th>
<th><strong>Category of data quality issues</strong></th>
<th><strong>List of DQ issues</strong></th>
<th><strong>Explanation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified data quality issues</td>
<td>Incomplete metadata for an entity and incomplete metadata for a metric</td>
<td>This data set does not contain entity metadata and metric metadata.</td>
<td></td>
</tr>
<tr>
<td>Unidentified data quality issues</td>
<td>Imprecise metadata for an entity and imprecise metadata for a metric.</td>
<td>This data set does not contain entity metadata and metric metadata.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Result for evaluation scale of metadata</strong></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The evaluation scale of metadata for this data set is 0 star because there is no metric metadata and entity metadata.</td>
<td></td>
</tr>
</tbody>
</table>
G.2 Full report for Data Quality Assessment Framework
Elements of dataset category that are present in the model of data set

- Metric metadata
- Entity metadata
- Ancillary metadata

We describe the details of dataset categories that are not present in the model of data set and suggestions to improve this data set as follows.

<table>
<thead>
<tr>
<th>Element of dataset category</th>
<th>Description</th>
<th>Suggestions to improve the data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity metadata</td>
<td>1 entity label does not have entity metadata. The entity label is 'name'.</td>
<td>Add description to describe the entity label.</td>
</tr>
<tr>
<td>Metric metadata</td>
<td>21 metric labels do not have metric metadata. The metric labels are ‘wmc, dit, noc, cbo, rfc, lcom, ca, ce, npm, lcom3, loc, dam, moa, mfa, cam, ic, cbm, amc, avg_cc, max_cc, bug’.</td>
<td>Add description to describe each metric label.</td>
</tr>
<tr>
<td>Ancillary metadata</td>
<td>2 ancillary labels do not have ancillary metadata. The ancillary labels are 'name' and 'version'.</td>
<td>Add description to describe each ancillary label.</td>
</tr>
</tbody>
</table>

We describe the details of dataset categories that are not present in the model of data set as follows.

- Incomplete metadata for an entity and incomplete metadata for a metric.
- Imprecise metadata for an entity and imprecise metadata for a metric.

<table>
<thead>
<tr>
<th>Category of data quality issues</th>
<th>List of DQ issues</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified data quality issues</td>
<td>Incomplete metadata for an entity and incomplete metadata for a metric</td>
<td>This data set does not contain entity metadata and metric metadata.</td>
</tr>
<tr>
<td>Unidentified data quality issues</td>
<td>Imprecise metadata for an entity and imprecise metadata for a metric.</td>
<td>This data set does not contain entity metadata and metric metadata.</td>
</tr>
</tbody>
</table>

For the evaluation scale of metadata, the scale of metadata for each metric label is 0 stars because there are no metric metadata for each metric label. The following shows the outcome for scale of metadata for each metric label.

<table>
<thead>
<tr>
<th>Metric label</th>
<th>Scale of metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>wmc</td>
<td>0 star</td>
</tr>
<tr>
<td>dit</td>
<td>0 star</td>
</tr>
<tr>
<td>noc</td>
<td>0 star</td>
</tr>
<tr>
<td>cbo</td>
<td>0 star</td>
</tr>
<tr>
<td>rfc</td>
<td>0 star</td>
</tr>
<tr>
<td>lcom</td>
<td>0 star</td>
</tr>
<tr>
<td>ca</td>
<td>0 star</td>
</tr>
<tr>
<td>ce</td>
<td>0 star</td>
</tr>
<tr>
<td>npm</td>
<td>0 star</td>
</tr>
<tr>
<td>lcom3</td>
<td>0 star</td>
</tr>
<tr>
<td>loc</td>
<td>0 star</td>
</tr>
<tr>
<td>dam</td>
<td>0 star</td>
</tr>
<tr>
<td>moa</td>
<td>0 star</td>
</tr>
<tr>
<td>mfa</td>
<td>0 star</td>
</tr>
<tr>
<td>cam</td>
<td>0 star</td>
</tr>
<tr>
<td>ic</td>
<td>0 star</td>
</tr>
<tr>
<td>cbm</td>
<td>0 star</td>
</tr>
<tr>
<td>amc</td>
<td>0 star</td>
</tr>
<tr>
<td>max_cc</td>
<td>0 star</td>
</tr>
<tr>
<td>avg_cc</td>
<td>0 star</td>
</tr>
<tr>
<td>bug</td>
<td>0 star</td>
</tr>
</tbody>
</table>

Assessment results: Full report

<table>
<thead>
<tr>
<th>Outcome description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment results: Full report</td>
<td>Entity label, metric label, ancillary label, measurement values, record identifiers, ancillary values and record.</td>
</tr>
<tr>
<td><strong>Elements of dataset category that are present in the model of data set</strong></td>
<td>Metric metadata, entity metadata and ancillary metadata.</td>
</tr>
<tr>
<td><strong>Elements of dataset category that are not present in the data set</strong></td>
<td>We describe the details of dataset categories that are not present in the model of data set as follows.</td>
</tr>
<tr>
<td><strong>Result for identifying data quality issues</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Result for evaluation scale of metadata</strong></td>
<td>The evaluation scale for metadata for this data set is 0 stars because there are no metric metadata for each metric label. The following shows the outcome for scale of metadata for each metric label.</td>
</tr>
</tbody>
</table>

2 ancillary labels do not have ancillary metadata. The ancillary labels are 'name' and 'version'.
G.3 Matrix for applying the Quality Assessment Process
<table>
<thead>
<tr>
<th>No.</th>
<th>Measurement data set</th>
<th>Record</th>
<th>Measurement value</th>
<th>Metric label</th>
<th>Entity label</th>
<th>Record identifier</th>
<th>Auxiliary label</th>
<th>Auxiliary value</th>
<th>Ancillary label</th>
<th>Ancillary value</th>
<th>Ancillary label</th>
<th>Ancillary value</th>
<th>Ancillary label</th>
<th>Ancillary value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PROMISE: AM1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>2</td>
<td>Quality Weka 3.7.5</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>3</td>
<td>PROMISE: Nasar93</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>4</td>
<td>PROMISE: ISBSG10</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>5</td>
<td>PROMISE: Cosmic</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>6</td>
<td>PROMISE: Mozilla4</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>7</td>
<td>PROMISE: Renes</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>8</td>
<td>PROMISE: miyazaki</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>9</td>
<td>PROMISE: KC2</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>10</td>
<td>PROMISE: Datatrivie</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>11</td>
<td>PROMISE: JMI</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>12</td>
<td>PROMISE: Nickle</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>13</td>
<td>PROMISE: Xfuture</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>14</td>
<td>PROMISE: Xorg</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>15</td>
<td>PROMISE: Coo81-dem</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>16</td>
<td>PROMISE: cocoon50dr</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>17</td>
<td>PROMISE: Nasar93-dem</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>18</td>
<td>PROMISE: MB2-silag211</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>19</td>
<td>PROMISE: MB1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>20</td>
<td>PROMISE: KC3</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>21</td>
<td>PROMISE: Cocomom81</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>22</td>
<td>PROMISE: PC1 - PC5</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>23</td>
<td>PROMISE: MC1-MC2 (2)</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>24</td>
<td>PROMISE: Kitchenham</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>25</td>
<td>PROMISE: Maxwell</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>26</td>
<td>PROMISE: Abraham</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>27</td>
<td>PROMISE: Kemerer</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>28</td>
<td>PROMISE: China</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>29</td>
<td>PROMISE: NFR</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>30</td>
<td>PROMISE: Bike</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>31</td>
<td>PROMISE: Coc811h</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>32</td>
<td>PROMISE: AR (7)</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>33</td>
<td>PROMISE: CM1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>34</td>
<td>PROMISE: MB2-silag433</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>35</td>
<td>PROMISE: Midedit</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>36</td>
<td>PROMISE: PITS - PITSF (6)</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>37</td>
<td>PROMISE: Anto to Zuel (33)</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>38</td>
<td>PROMISE: Bugreport</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>39</td>
<td>PROMISE: Boeticher</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>40</td>
<td>PROMISE: NpClassic (5)</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

Legend:
- ✔️: Present in the data set.
- ×/¥: Not present in the data set.
- 1: The quality issue is identified in the data set.
- 0: The quality issue is unidentified due to the absence of specific dataset category elements.
- [A] (B): A is a star rating for a category metadata and [B] is number of a category label.
- R: Risk of misinterpretation is identified in the data set.
H.1 Full report for Example 1: PROMISE Boetticher dataset
We made assumptions to determine the dataset category and dataset concept elements because there is no metadata in the data set. The elements are: population, record, entity, measurement value and metric label. We notice that the web page of data set do provide information about the publication that uses this data set. However, we not able to evaluate the information in the publication because it is two steps further from the location that we access the data set. Furthermore, the information in the publication does not provide clear and explicit description about the data set.

This data set does not contain the following dataset category elements: record identifier, entity label, entity metadata and metric metadata. We describe the implications of these dataset category elements to the quality assessment and suggestions to improve this data set as follow.

### Data quality issues

We describe the detail as follows:

**Identified data quality issues**

- Incomplete metadata for metric

**Unidentified data quality issues**

- Duplicate data, inconsistent data, incomplete metadata for an entity, imprecise metadata for a metric and imprecise metadata for an entity

### Overall result for evaluation scale of metadata

The evaluation scale of metadata for this data set is 0 star because there is no metric metadata for each metric label. The following shows the outcome for scale of metadata for each metric label.

### Metric label description

<table>
<thead>
<tr>
<th>Metric label</th>
<th>Description</th>
<th>Scale of metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. Sc. Undergrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Comp. Sc. Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>MIS Undergrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>MIS Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Proj Mgmt Undergrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Proj Mgmt Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Soft. Eng. Undergrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Soft. Eng. Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Tech Undergrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Tech Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Mgmt UnderGrad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Mgmt Grad Courses</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Comp. Sc. Workshops</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>MIS Workshops</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>MIS Conferences</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Proj Mgmt Workshops</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Proj Mgmt Conferences</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Soft. Eng. Workshops</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Total Workshops</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Total Conferences</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Ada Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>ASP Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Assembly Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>C Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>C++. Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>C# Experience</td>
<td></td>
<td>0 star</td>
</tr>
<tr>
<td>Cobol Experience</td>
<td></td>
<td>0 star</td>
</tr>
</tbody>
</table>


Add metadata for each metric label.

Add record identifiers that represent the entities.

Add entity label to represent a property that values distinguish between entities.

Add entity metadata for the entity label.

Add metric metadata for each metric label.
<table>
<thead>
<tr>
<th>07</th>
<th>Database Experience</th>
<th>0 star</th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>Delphi Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>09</td>
<td>Fortran Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>10</td>
<td>Java Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>11</td>
<td>Pascal Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>12</td>
<td>Perl Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>13</td>
<td>PHP Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>14</td>
<td>PowerBuilder Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>15</td>
<td>SAP Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>16</td>
<td>TCL Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>17</td>
<td>VB Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>18</td>
<td>Other Program Lang Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>19</td>
<td>TotalLangExp</td>
<td>0 star</td>
</tr>
<tr>
<td>20</td>
<td>Names of other Lang</td>
<td>0 star</td>
</tr>
<tr>
<td>21</td>
<td>Hardware Work Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>22</td>
<td>Software Work Experience</td>
<td>0 star</td>
</tr>
<tr>
<td>23</td>
<td>Hardware Proj Mgmt Exp</td>
<td>0 star</td>
</tr>
<tr>
<td>24</td>
<td>Software Proj Mgmt Exp</td>
<td>0 star</td>
</tr>
<tr>
<td>25</td>
<td>No Of Hardware Proj Estimated</td>
<td>0 star</td>
</tr>
<tr>
<td>26</td>
<td>No Of Software Proj Estimated</td>
<td>0 star</td>
</tr>
<tr>
<td>27</td>
<td>Avg Size Of Software Proj Estimated</td>
<td>0 star</td>
</tr>
<tr>
<td>28</td>
<td>Client Exp</td>
<td>0 star</td>
</tr>
<tr>
<td>29</td>
<td>Procurement Industry Exp</td>
<td>0 star</td>
</tr>
<tr>
<td>30</td>
<td>Scale Factor</td>
<td>0 star</td>
</tr>
<tr>
<td>31</td>
<td>Correlation</td>
<td>0 star</td>
</tr>
</tbody>
</table>
H.2 Full report for Example 2: PROMISE Datatrieve dataset
**Determined elements of dataset category and dataset concept**

Elements of dataset category that are not present in the data set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Element of dataset category</th>
<th>Description</th>
<th>Suggestions to improve the data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Record identifier</td>
<td>We are not able to identify values that represent the entities and not able to distinguish between the records in the data set.</td>
<td>Add record identifiers that represent the entities.</td>
</tr>
<tr>
<td>2</td>
<td>Entity label</td>
<td>We are not able to determine the property that values distinguish between entities.</td>
<td>Add entity label to represent a property that values distinguish between entities.</td>
</tr>
<tr>
<td>3</td>
<td>Entity metadata</td>
<td>We are not able to determine the metadata that describe the entity label.</td>
<td>Add entity metadata for the entity label.</td>
</tr>
</tbody>
</table>

**Data quality issues**

We describe the detail as follows:

<table>
<thead>
<tr>
<th>Category of data quality issues</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified data quality issues</td>
<td>Imprecise metadata for metric LOC60, LOC61, ReusedLOC, AddedLOC, DelLoC, ModRate, ModKnow.</td>
</tr>
<tr>
<td>Unidentified data quality issues</td>
<td>Duplicate data, inconsistent data, incomplete metadata for entity and imprecise metadata for entity.</td>
</tr>
</tbody>
</table>

**Overall result for evaluation scale of metadata**

The scale of metadata for metric is 1 star because the metadata is visible, but not explicitly described the meaning of the property or attribute of entity. The scale of metadata for entity is 0 star because the metadata is not visible. We list the following scale of metadata for each metric label.

<table>
<thead>
<tr>
<th>No</th>
<th>Metric label</th>
<th>Scale of metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LOC6_0</td>
<td>1 star</td>
</tr>
<tr>
<td>2</td>
<td>LOC6_1</td>
<td>1 star</td>
</tr>
<tr>
<td>3</td>
<td>Added_LoC</td>
<td>1 star</td>
</tr>
<tr>
<td>4</td>
<td>Del_LoC</td>
<td>1 star</td>
</tr>
<tr>
<td>5</td>
<td>Diff_Block</td>
<td>1 star</td>
</tr>
<tr>
<td>6</td>
<td>Mod_Rate</td>
<td>1 star</td>
</tr>
<tr>
<td>7</td>
<td>Mod_Know</td>
<td>2 stars</td>
</tr>
<tr>
<td>8</td>
<td>Reused_LOC</td>
<td>1 star</td>
</tr>
<tr>
<td>9</td>
<td>FaultyE_1</td>
<td>2 stars</td>
</tr>
</tbody>
</table>
I.1 Participant Information Sheet
PARTICIPANT INFORMATION SHEET

Participant

Name of Researcher: Marshima Rosli

Researcher Introduction
I am Marshima Rosli, a PhD student in the Department of Computer Science at The University of Auckland under the supervision of Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly.

Project description and invitation
I am conducting a research developing a framework to evaluate the quality of software engineering data sets for analysis in empirical research. The framework consists of a dataset metamodel and a quality assessment. Part of our research involves an evaluation of this framework regarding its usability and effectiveness for describing the content of data sets, and identifying data quality issues in data sets.

If you are a researcher who has conducted any research in software engineering or computer science using data sets, you are invited to participate in this evaluation. Your comments and assistance would be greatly appreciated.

Project procedures
You will be requested to fill out an online survey, which will take 30-40 minutes of your time. In the survey, you will be asked a number of questions relating to the effectiveness and usability of part of the framework for three different kinds of data sets. I will be using your answers from this survey to compare and detect any issues with the effectiveness and usability of the framework.

Data storage/retention/destruction/future use
The digital data collected will be archived and kept indefinitely on a password protected computer and server in The University of Auckland, and will be used for future research about data quality assessment. The data will be analysed and the results of this research will be used for a PhD thesis and other academic publications. The results and publications resulting from this study will be available from the researcher upon request.
Right to withdraw from participation
Participation in this research is completely voluntary. You have the right to withdraw from the survey at any stage while you fill-in the survey. Please be aware that once you have submitted the survey, the collected data cannot be withdrawn.

Anonymity
The data that will be collected will only be used for academic purposes. The responses will be completely anonymous as you will not be asked to write your name on the responses. The survey will be conducted using Qualtrics and only can be accessed by the researcher as it will be protected by a password. The researcher will not be able to trace you through your responses, and no one will know whether or not you participated in the study.

Contact details
If you have any queries regarding this survey, please do not hesitate to contact me. You can email me at mmoh603@aucklanduni.ac.nz or call me (+6493737599 ext 82128). You may also contact my supervisors, Associate Prof. Ewan Tempero (+6493737599 ext 83765) and Dr. Andrew Luxton-Reilly (+6493737599 ext 85654).

For any queries regarding ethical concerns you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland 1142. Telephone 09 373-7599 extn. 83711. Email ro-ethics@auckland.ac.nz

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 5 October 2015 for 3 years, Reference Number 015880.
I.2 Advertisement
Email subject: A survey on evaluation of framework for data quality assessment.

Dear respondent,

I am Marshima Rosli, a PhD student in the Department of Computer Science at The University of Auckland. I am conducting a survey on evaluation of framework for data quality assessment as part of my PhD research on evaluating the quality of data sets. The purpose of this survey is to evaluate the usability and effectiveness of the framework for describing the content of data sets, and identifying data quality issues in data sets.

If you have conducted research in software engineering or computer science using data sets, you are invited to participate in this survey. To fill out the survey, please visit the following link: [http://tinyurl.com/qexcvuo](http://tinyurl.com/qexcvuo)

Your participation in the survey is completely voluntary and all your responses will be completely anonymous. Information provided will be stored electronically on a password protected computer and server in The University of Auckland and will be accessible to the researchers only. The data from this survey will be analysed and the analysis results will be used in a PhD thesis, journal articles and conference papers.

If you would like further information on this survey, please contact me at mmoh603@aucklanduni.ac.nz. If you know any potential participants for this survey, feel free to pass this advertisement.

Thank you very much for your time and cooperation.

Marshima Rosli  
PhD Student  
The University of Auckland

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 5 October 2015 for 3 years, Reference Number 015880.
J.1 Dataset Quality Assessment Survey
Dataset Quality Assessment Survey

Thank you for taking part in this research project. Please try to answer every question in this survey.

Section I Demographic Information

Q1(a) Are you ....
☐ Masters student
☐ PhD student
☐ Academic researcher
☐ Industry researcher
☐ Other ____________________

Q1(b) Have you used or analysed datasets from software repositories?
☐ Yes
☐ No

Q1(c) Which of the following software repositories contain datasets that you used for research?
☐ PROMISE (PRedictOr Models In Software Engineering)
☐ ISBSG (International Software Benchmarking Standards Group)
☐ Qualitas Corpus
☐ SIR (Software-artifact Infrastructure Repository)
☐ Eclipse Bug Data
☐ Helix Data Set
☐ SRDA (SourceForge Research Data Archive)
☐ Other ____________________

Q1(d) Which of the following quality issues that you have ever encountered with datasets?
☐ Duplicate data: Two or more records that have same measurement values associated with the same metric for the same entity.
☐ Missing data: A record that does not have a measurement value for a given metric.
☐ Inconsistent data: Two or more records that have different measurement values associated with the same metric for the same entity.
☐ None

Q1(e) Please provide any comments on quality issues in datasets.


Q1(f) Which of the following programming languages do you know?

- Java
- C++
- C#
- Python
- PHP
- C
- Other ____________________
Section II Dataset Framework Information
Software engineering datasets contain data pertaining to entities from various populations. Every entity has data associated with it, which are known as the characteristics of the entity. This study distinguishes between attributes, which are characteristics associated with metrics and properties, which are characteristics not associated with metrics.

Software engineering datasets may contain data other than attributes and properties. Examples of such data are column headers that provide names for the characteristics, and comments that provide more detailed explanations of the characteristics. Additionally, datasets may also contain text that is used solely to provide structure for the data.

In this study, we define dataset as consisting of a collection of elements. An element can be a single token, unit, word or phrase. We classify the elements as one of value, label, metadata, structural, and other. The definition of these elements will be provided later in the survey.

In this section, you will be asked a same number of questions for three different datasets.

Q2 Figure 1 shows part of the PROMISE Ant 1.7 dataset, which is presented in a comma separated values file that has columns and rows. It has column headers that represent different properties or attributes of entities. Each row except the first row is a record that contains a list of values for a particular entity and each value corresponds to different properties or attributes of entities. (You may access the full version of this dataset from here)
Q2(a) Approximately how many times have you analysed comma separated values (CSV) datasets?
- None
- 1-5
- 6-10
- 11-20
- More than 20

Q2(b) Record identifier is defined as an element that uniquely identifies an entity. Please select all the elements that meet the definition of record identifier.

Q2(c) How long does it take to identify the record identifier in Figure 1?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q2(d) Entity label is defined as an element that represents a property that identifies the entity. Please select all the elements that meet the definition of entity label.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q2(e) How long does it take to identify the entity label in Figure 1?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q2(f) Metric label is defined as an element that represents a metric that measures the attribute of entity. Please select all the elements that meet the definition of metric label.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know
Q2(g) How long does it take to identify the metric label in Figure 1?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q2(h) Measurement value is defined as an element that obtained through the process of measurement for attribute of entity. Please select all of the elements that meet the definition of measurement value.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q2(i) How long does it take to identify the measurement value in Figure 1?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q2(j) Metric metadata is defined as a set of elements that describes the metric used to measure an entity. Please select all the elements that meet the definition of metric metadata.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q2(k) How long does it take to identify the metric metadata in Figure 1?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)
Q2 (l) Which of the following is the entities in Figure 1?

- Java classes
- Source code files
- Java objects
- Don't know
- None

Q2(m) How did you identify the entities in Figure 1?

- I know this dataset.
- It is stated in the dataset.
- I made a reasonable guess based on my experience.
- Other ____________________

Q2 (n) Below is definitions of data quality issues. Please select the quality issues that you think appear in Figure 1.

- Duplicate data: Two or more records that have same measurement values associated with the same metric for the same entity.
- Missing data: A record that does not have a measurement value for a given metric.
- Inconsistent data: Two or more records that have different measurement values associated with the same metric for the same entity.
- Incomplete metadata: A label represents a property or an attribute of entity that does not have metadata.
- Inconsistent metadata: Two or more labels that are identical but have different values correspond to properties or attributes of entity.
- Imprecise metadata for metric: The metric metadata is not explicitly describes the metric used to measure the attribute of entity.
- None.
- Don't know.

Q2(o) How long does it take to identify the quality issues in Figure 1?

- Short (e.g. 1-14 seconds)
- Medium (e.g. 15-30 seconds)
- Long (e.g. more than 30 seconds)
Q3 Figure 2 shows part of the KC2 data set, which is presented in an Attribute-Relation File Format (ARFF). In this dataset, there are no columns and it is organized differently from the previous dataset. There are rows that contain properties or attributes of entities and values. The elements after @attribute represent properties or attributes of entity. Every row after the @data contains values corresponding to different properties or attributes of entities.
Q3(a) Approximately how many times have you analysed attribute-relation file format (ARFF) data sets?
- None
- 1-5
- 6-10
- 11-20
- More than 20

Q3(b) Record identifier is defined as an element that uniquely identifies an entity. - Please select all the elements that meet the definition of record identifier.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q3(c) How long does it take to identify the record identifier in Figure 2?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q3(d) Entity label is defined as an element that represents a property that identifies the entity. - Please select all the elements that meet the definition of entity label.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q3(e) How long does it take to identify the entity label in Figure 2?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)
Q3(f) Metric label is defined as an element that represents a metric that measures the attribute of entity. Please select all the elements that meet the definition of metric label.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q3(g) How long does it take to identify the metric label in Figure 2?

- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q3(h) Measurement value is defined as an element that obtained through the process of measurement for attribute of entity. Please select all of the elements that meet the definition of measurement value.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q3(i) How long does it take to identify the measurement value in Figure 2?

- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)
Q3(j) Metric metadata is defined as a set of elements that describes the metric used to measure an entity. Please select all the elements that meet the definition of metric metadata.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q3(k) How long does it take to identify the metric metadata in Figure 2?

- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q3(l) Which of the following are the entities in Figure 2?

- Java classes
- Change requests
- Source code files
- Don’t know
- None

Q3(m) How did you identify the entities in Figure 2?

- I know this dataset.
- It is stated in the dataset.
- I made reasonable guess based on my experience.
- Other ____________________

Q3(n) Below are definitions for data quality issues. Please select the quality issues that you think appear in Figure 2.

- Duplicate data: Two or more records that have same measurement values associated with the same metric for the same entity.
- Missing data: A record that does not have a measurement value for a given metric.
- Inconsistent data: Two or more records that have different measurement values associated with same metric for the same entity.
- Incomplete metadata: A label represents a property or an attribute of entity that does not have metadata.
- Inconsistent metadata: Two or more labels that are identical but have different values correspond to properties or attributes of entity.
- Imprecise metadata for metric: The metric metadata is not explicitly describes the metric used to measure the attribute of entity.
- None.
- Don’t know.
Q3(o) How long does it take to identify the quality issues in Figure 2?
- Short (e.g. 1-14 seconds)
- Medium (e.g. 15-30 seconds)
- Long (e.g. more than 30 seconds)
Q4 Figure 3 shows an example of a dataset in tabular format that has rows and columns. As similar to Figure 1, it has column headers in the first row that represent different properties or attributes of entities and each subsequent row is a record.

![Figure 3: An example of tabular format data set.](image)

Q4(a) Approximately how many times have you analysed tabular format data sets?
- None
- 1-5
- 6-10
- 11-20
- More than 20

Q4(b) Record identifier is defined as an element that uniquely identifies an entity. Please select all the elements that meet the definition of record identifier.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know
Q4(c) How long does it take to identify the record identifier in Figure 3?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q4(d) Entity label is defined as an element that represents a property that identifies the entity. Please select all the elements that meet the definition of entity label.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q4(e) How long does it take to identify the entity label in Figure 3?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q4(f) Metric label is defined as an element that represents a metric that measures the attribute of entity. Please select all the elements that meet the definition of metric label.
- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don't know

Q4(g) How long does it take to identify the metric label in Figure 3?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)
Q4(h) Measurement value is defined as an element that obtained through the process of measurement for attribute of entity. Please select all of the elements that meet the definition of measurement value.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q4(i) How long does it take to identify the measurement value in Figure 3?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q4(j) Metric metadata is defined as a set of elements that describes the metric used to measure an entity. Please select all the elements that meet the definition of metric metadata.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- None
- Don’t know

Q4(k) How long does it take to identify the metric metadata in Figure 3?
- Short (e.g. 1-10 seconds)
- Medium (e.g. 11-30 seconds)
- Long (e.g. more than 30 seconds)

Q4(l) Which of the following are the entities in Figure 3?
- Java classes
- Change requests
- Source code files
- Don't know
- None
Q4(m) How did you identify the entities in Figure 3?
☑ I know this dataset.
☑ It is stated in the dataset.
☑ I made reasonable guess based on my experience.
☑ Other ______________________

Q4(n) Below are definitions for data quality issues. Please select the quality issues that you think appear in Figure 3.
☑ Duplicate data: Two or more records that have same measurement values associated with the same metric for the same entity.
☑ Missing data: A record that does not have a measurement value for a given metric.
☑ Inconsistent data: Two or more records that have different measurement values associated with same metric for the same entity.
☑ Incomplete metadata: A label represents a property or an attribute of entity that does not have metadata.
☑ Inconsistent metadata: Two or more labels that are identical but have different values correspond to properties or attributes of entity.
☑ Imprecise metadata for metric: The metric metadata is not explicitly describes the metric used to measure the attribute of entity.
☑ None.
☑ Don’t know.

Q4(o) How long does it take to identify the quality issues in Figure 3?
☑ Short (e.g. 1-14 seconds)
☑ Medium (e.g. 15-30 seconds)
☑ Long (e.g. more than 30 seconds)

Q5 Please use the space below to provide any comments that you have regarding our framework.

Thank you for participating in this survey. We really appreciate your help.
Marshima Mohd Rosli, Dept. of Computer Science, The University of Auckland
J.2 Ethics application for an online survey
Updated By: Ewan D Tempero @ 25-Aug-2015 01:18:31 PM

**PREAMBLE**

**UNIVERSITY PERSONNEL**

**OTHER PERSONNEL**

**RESEARCH TYPE**

**RESEARCH PROCEDURES**

**PARTICIPANTS & CONSENT**

**STORAGE & RESULTS**

**CULTURAL ISSUES**

**RISKS & BENEFITS**

**HUMAN REMAINS/TISSUE**

**CLINICAL TRIALS & FUNDING**

**ETHICAL SUMMARY & ADVISOR REVIEW**

**ATTACHMENTS & CHECKLIST**

**FEEDBACK**

**ALL PAGES**

---

**eFORM VERSION**

Version 4.1 18-Jul-2015

**PROTOCOL**

Protocol No: 015880

Title: Evaluation of a framework for data quality assessment

**GENERAL**

Prior to completing your application:

- Read the Guiding Principles for conducting research with human participants.
- Go through the Applicants' Reference Manual.
- Check if an exemption applies (see Guiding Principles section 6.1 and Applicants' Reference Manual Section 3.8).
- Check if the matter needs to be referred to a Health and Disability Ethics Committee (HDEC) (see Guiding Principles section 3.1 and 6.1 and Applicants' Reference Manual Section 4.5).

For creating and submitting your application refer to the Human Ethics Module Quick Guide.

All documentation to support your application is on the UAHPEC web page.

For any queries please log a call with the Staff Service Centre at ext. 86000 or staffservice@auckland.ac.nz and it will be referred to the relevant team in the Research Office or ITS.

Please Note: Internet Explorer should not be used when working in InfoEd as many functions are not compatible with this browser. We recommend using the following browsers: Google Chrome, Mozilla Firefox, or Safari (for Mac).

**SECTION A: PERSONNEL**

* Are you a Student?

Yes [x] No [ ]

* Please add the degree you are studying towards.

PhD Computer Science

If the PI listed below is not your supervisor, please use the 'Change Project Information' button on the protocol screen to change the PI to your supervisor before continuing with the rest of the form. You should also change the department to your PI's department. Instructions on how to do this can be found [here](#).

PI: Tempero, Ewan D*

[ ]* I confirm that the PI listed above is my supervisor

Department: Computer Science*

[ ]* I confirm that the department listed above is my supervisor’s department

---

1. List all personnel, including co-investigators and ethics advisors, by selecting their name from the UNIVERSITY PERSONNEL page and add their Role from the dropdown list.
2. If you are a student, you must add your own name to allow access after closing the form.

3. To change the Principal Investigator, tick the ‘PI’ checkbox to the left of the relevant person’s name.

   Note: A student cannot be a PI. See Applicants’ Reference Manual Section 5.0 that states: “For Doctoral, Masters and Honours research, applications should be submitted by the primary supervisor who will be the Principal Investigator (PI)”.
<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>PI</th>
<th>Date</th>
<th>Co-Investigator</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tempero, Ewan D</td>
<td>PI</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohd Rosli, Marshima B</td>
<td>PI</td>
<td></td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luxton-Reilly, Andrew J</td>
<td>Student</td>
<td></td>
<td>18-Aug-2015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Any University of Auckland personnel and other personnel that cannot be found in the lookup list should be added here.

If you are a student, please add your University of Auckland ID

| Full Name (and ID) | Institution / Department | Project Role | Email address |
* Is this a Research Project or Coursework Application?
Research
Note: Because you are a student, please select Research
SECTION B: RESEARCH PROCEDURES

B:1 Title.
Evaluation of a framework for data quality assessment

*B:2 Aims/objectives of the project.*
The main objective of this project is to study the effectiveness and usability of a framework for evaluating the quality of data sets. This framework also includes a dataset metamodel to describe the structure of data sets, and a quality assessment process to evaluate the quality of data sets. The outcome will assist researchers in understanding how effective the framework is.

Please note: all acronyms must be written out in full the first time they appear in the application, recruiting materials, Participant Information Sheet (PIS) and Consent Form (CF).

*B:3 Summary of the project.*
The quality of empirical software engineering results and the validity of results interpretation essentially rely on the quality of the analysed data sets. Researchers have indicated that the quality of analysed data sets is critical to the results of empirical studies (Shepperd, M., 2011). Although data sets play a central role in empirical research, few studies consider the quality of the data sets used (Zhang, W., Yang, Y., & Wang, Q., 2011; Mockus, A., 2008). If the quality is poor, then the results of empirical studies cannot be trusted, hence, any models or conclusions based on the data sets are questionable. Ensuring good data quality is therefore, fundamentally important to the field of empirical software engineering.

We have developed a framework that researchers can use to evaluate the quality of data sets in empirical software engineering. The framework includes precise definitions of data sets and their potential elements, a metamodel to describe the structure and concepts in a data set, and the relationships between each concept. We anticipate that the definitions and the metamodel will allow researchers to clearly understand what data sets are actually intended to represent and identify whether a data set has sufficient information to be usable for analysis in empirical research.

We have also constructed a data quality assessment process to identify quality issues in data sets. This assessment contains a formal definition for data quality issues and allows researchers to clearly identify whether a data set has quality issues that includes issues related to the interpretation of data.

We therefore seek approval to investigate the effectiveness and usability of the definitions in the dataset metamodel and definitions for data quality issues in the quality assessment using an online survey. The findings from this study will inform suggestions and recommendations for future improvement of the framework.

*B:4 Project duration (in months).*
24 months
Note: The start date is when the proposal is approved

*B:5 Describe the study design.*
We will conduct an online survey using Qualtrics for comparative analysis, particularly in detecting any issues related to the understanding and application of the definitions in the dataset metamodel, and the definitions for data quality issues in the quality assessment.

The online survey consists of closed and open questions to obtain qualitative and quantitative data from participants. There will be two main sections in the online survey. The first section contains demographic questions and the second section contains questions about the effectiveness and usability of the framework. The questions in the second section are applications of the definitions in the dataset metamodel, and the definitions of data quality issues for three different data sets.

We will collect both qualitative and quantitative data. Quantitative analyses will be descriptive and will examine the distribution of all answers instead of only calculating medians. Qualitative analyses of free text answers will be conducted as thematic analysis.

*B:6 List all the methods used for obtaining information.*
* Interviews
  Yes ☐ No ☑
* Focus groups
* Questionnaires
Yes ☐ No ☐

Please attach the questionnaire(s) in the "Attachment" section at the end of the form.

* Observations
Yes ☐ No ☐

* Other
Yes ☐ No ☐

* B:7 Does the research involve processes that involve EEG, ECG, MRI, TMS, FMRI, EMG, radiation, invasive or surface recordings?
Yes ☐ No ☐

* B:8 Does the research involve processes that are potentially disadvantageous to a person or group (for example, the collection of information which may expose the person/group to discrimination)?
Yes ☐ No ☐

* B:9 Who will carry out the research procedures?

Marshima Mohd Rosli

If necessary, please provide more details of the role(s) of each member of the research team. If the research procedures will be carried out by a third party other than the researcher or co-investigators, please attach a copy of the confidentiality agreement to the "Attachment" section at the end of this form.

* B:10a Where will the research procedures take place?

Potentially anywhere as it is anonymous on-line survey.

If permission is required to conduct the study at a specific location, please attach an appropriate PIS and Consent form, or a support letter, in the "Attachment" section at the end of the form.

* B:10b Will the research be conducted overseas?
Yes ☐ No ☐

* Which countries are involved?

Data analysis will be conducted in New Zealand however data may be gathered from potentially anywhere as it is an anonymous on-line survey.

Please provide local contact details, as well as those of contacts at the University in the PIS.

* B:10c If the study is based overseas, explain what special circumstances arise and how they will be dealt with. Include any special requirements of the country (e.g. research visa) and/or the community with which the research will be carried out.

Please also provide an undertaking to abide by any local laws relating to research privacy and data collection. For more information, see Applicants' Reference Manual Section 13.6.

* B:11a Is the questionnaire web-based?
Yes ☐ No ☐

* B:11b Is it an anonymous questionnaire?
Yes ☐ No ☐

Explain and indicate on the PIS how anonymity will be preserved:
The survey will be conducted through an online survey using Qualtrics tool. The researcher will turn off the option for IP address tracking in the survey settings of the Qualtrics tool. The survey only can be accessed by the researcher as it will be protected by a password and it is located on a password protected server in The University of Auckland. The responses will be completely anonymous as participants will not be asked to providing identifying information. The researcher will not be able to trace the participants through their responses.

* B:12 How much time will participants need to give to the research?

Less than 40 minutes.

* B:13 Will information on the participants be obtained from third parties?
Yes ☐ No ☐

* B:14 Will any identifiable information on the participants be given to third parties?
Yes ☐ No ☐

* B:15 Does the research involve evaluation of the University of Auckland services or organisational practices where information of a personal nature may be collected and where participants may be identified?
Yes ☐ No ☐
<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>B:16 Does the research involve a conflict of interest or the appearance of a conflict of interest for the researcher?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:17 Does the research involve matters of commercial sensitivity?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:18 Has the study design or the use of the data been influenced by an organisation outside the University of Auckland?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:19 Are you intending to conduct the research in the University of Auckland class time?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:20 Does the research involve deception of the participants, including concealment or covert observations?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:21 Is there any koha, compensation or reimbursement of expenses to be made to participants?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:22a Is this an intervention study?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:22b Does this research involve potentially hazardous substances?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SECTION C: PARTICIPANTS

* C:1 Who are the participants in the research?

- Adults
  - Yes [ ] No [ ]

- Own colleagues
  - Yes [ ] No [ ]
  
  Explain in Section C7.

- Own students
  - Yes [ ] No [ ]

- Persons whose capacity to give informed consent (other than children) is compromised
  - Yes [ ] No [ ]

- Persons who are in a dependent situation, such as people with a disability, residents of a hospital, nursing home or prison, or patients highly dependent on medical care
  - Yes [ ] No [ ]

- Persons aged less than 16 years old where parental consent is NOT being sought
  - Yes [ ] No [ ]

- Persons aged less than 16 years old where parental consent is sought
  - Yes [ ] No [ ]

- Other
  - Yes [ ] No [ ]

* C:2 How many organisations and departments within the organisations will participate in your project?

Not applicable

If you have letters of support, please attach these in the 'Attachment' section at the end of the form.

* C:3 How many individual participants (research participants) will participate in your project?

Approximately 50 participants.

* C:4 How will you identify potential participants and by which method are participants invited to take part in the research?

The participants will be researchers who have background or experience in empirical software engineering research. We used three different methods for participants recruitment:

a) An advertisement will be posted to mailing lists and groups likely to have members with relevant expertise, such as from Yahoo and LinkedIn groups.

b) We will approach people who we know are involved in empirical software engineering research. This is because there are limited numbers of people who are engaged in this research that have the expertise to answer this survey. We will send out an advertisement via email that provides the online survey link for them to participate in the survey and contains the researcher contact information. (We are aware that this is a direct contact but because it is an anonymous survey, which will be conducted online, we believe that there will be no power imbalance or undue inducement for people to participate.)

c) In addition to contacting people directly, we will be using the snowballing method and participants will be asked to forward the advertisement to other participant. Participants who are interested to participate in the survey will be able to access the PIS and the survey through the forwarded link in the advertisement, and may contact the researcher for any further information.

Using a direct approach to recruit potential participants is not recommended.

Please attach the advertisement, media release, or notice, etc. and the letter of permission from the agency supplying them (if applicable) in the 'Attachment' section at the end of the form.

* C:5 Who will make the initial approach to potential participants?

Researchers

* C:6 Will access to participants be gained with consent of any organisation?

Yes [ ] No [ ]

* C:7 Is there any special relationship between participants and researchers?

Yes [ ] No [ ]

* Explain:
It is possible that some of the participants that might answer the survey may be other PhD students within the Computer Science department or other researcher that we met at the conferences. However, since the survey is anonymous, we will have no idea whether or not the participants are colleagues.

* C:8 Does the research involve the University of Auckland staff or students where information of a personal nature may be collected and where participants may be identified?
   Yes [ ] No [x]

* C:9 Does the research involve participants who are being asked to comment on employers?
   Yes [ ] No [x]

* C:10 Are there any potential participants who will be excluded?
   Yes [ ] No [x]

SECTION D: INFORMATION AND CONSENT

* D:1 By whom and how will information about the research be given to participants?
   The PIS is distributed online using Qualtrics before the participants can access the online survey.

* D:2a Will the participants have difficulty giving informed consent on their own behalf?
   Yes [ ] No [x]

* D:3a If a questionnaire is used, will the participants have difficulty completing the questionnaire on their own behalf?
   Yes [ ] No [x]

* D:4 Does the research involve participants giving oral consent rather than written consent?
   Yes [ ] No [x]

* D:5 Does the research use previously collected information or biological samples for which there was no explicit consent?
   Yes [ ] No [x]

* D:6 Is access to the Consent Forms restricted to the Principal Investigator and/or the researcher?
   n/a
   * Explain, justify and indicate in the PIS:
     Not applicable as survey is anonymous

* D:7 Will Consent Forms be stored by the Principal Investigator, in a secure manner?
   n/a
   * Explain, justify and indicate in the PIS:
     Not applicable

* D:8 Are Consent Forms stored separately from data and kept for six years?
   n/a
   * Explain, justify and indicate in the PIS:
     Not applicable
SECTION E: STORAGE AND USE OF RESULTS

* E:1 Will the participants be audio-recorded, video-recorded, or recorded by any other electronic means such as Digital Voice Recorders?
   Yes ☐ No ☑

* E:3 For the questionnaire, is any coding scheme used to identify the respondent?
   Yes ☐ No ☑

* E:4a Explain how and how long the data (including audio-recordings, video-recordings and electronic data) will be stored.
   The archival data will be stored indefinitely on a password protected computer and server in The University of Auckland, and will be used for future research about data quality assessment.

* E:4b Explain how data will be used. (For example, in a thesis/dissertation, publications, and/or conference presentations etc.)
   The data will be used in preparation of thesis and potentially for publications in journals or conference proceedings.

* E:4c Explain how data will be destroyed.
   Not applicable.

* E:5 Describe any arrangements to make results available to participants.
   The results and publications resulting from this study will be available from the researcher upon request.

* E:6a Are you going to identify the research participants in any publication or report about the research?
   Yes ☐ No ☑

* E:6b Is there any possibility that individuals or groups could be identified in the final publication or report?
   Yes ☐ No ☑
SECTION F: TREATY OF WAITANGI
* F:1 Does the proposed research have impact on Māori persons as Māori?
Yes ☐ No ☒

SECTION G: OTHER CULTURAL ISSUES
* G:1 Are there any aspects of the research that might raise any specific cultural issues?
Yes ☐ No ☒
SECTION H: RISKS & BENEFITS

* H:1 What are the possible benefits to research participants of taking part in the research?
Upon the completion of this research, participants may have learned more about data quality and will be able to apply the definitions in the dataset metamodel to any data set and identify the potential of quality issues a data set may have by using the definitions for data quality issues in the quality assessment.

* H:2 Is the research likely to place the researcher at risk of harm?
Yes [□] No [X]

* H:3 Is the research likely to cause any possible harm to the participants, such as physical pain beyond mild discomfort, embarrassment, psychological or spiritual harm?
Yes [□] No [X]

* H:4 Does the research involve collection of information about illegal behaviour(s) which could place the research or participants at risk of criminal or civil liability or be damaging to their financial standing, employability, professional or personal relationships?
Yes [□] No [X]

* H:5 Is the research likely to give rise to incidental findings?
Yes [□] No [X]
SECTION I: HUMAN REMAINS, TISSUE AND BODY FLUIDS

* I:1 Does the research involve use of human blood, body fluids, or tissue samples?

Yes ☐ No ☑
SECTION J: CLINICAL TRIALS
* J:1 Is this project a Clinical Trial?
Yes ☐ No ☑

SECTION K: FUNDING
* K:1 Have you applied for, or received funding for this project?
Yes ☐ No ☑
SECTION L: ETHICAL SUMMARY

* L:1 Have you made any other related applications?
   Yes [ ] No [x]

* L:2 Is there any relevant information from past applications or interaction with UAHPEC?
   Yes [ ] No [x]

* L:3 Please provide a summary of all the ethical issues arising from this project and explain how they are to be resolved:

   **Anonymity**
   This is an anonymous survey that will be conducted online and only the researcher has the access to the survey as it will be protected by a password. The researcher will not be able to identify who has and has not participated in the survey. All data will be kept secured on a password protected server in The University of Auckland.

   **Informed consent**
   All interested participants will be provided with a full explanation of the study. They will receive the Participant Information Sheet (PIS) explaining the study. The participants have the right to withdraw from the survey at any stage while they fill-in the survey, but once they have submitted the survey, the collected data cannot be withdrawn.

   **Recruitment**
   We are aware that some of the recruitment in this survey involves a direct contact with the participant, but this is an anonymous survey that we will not know who has or has not participated in the survey. In addition, we are also not in the position of authority to influence people become participant.

   UAHPEC expects applicants to identify the ethical issues in the project and explain in the documentation how they have been resolved. The application will not be considered if this is not answered adequately. A 'Not applicable' response is not acceptable.

SECTION M: ETHICS ADVISOR REVIEW

* M:1 Will this Application be reviewed by an Ethics Advisor after you submit it?
   Yes [x] No [ ]

* M:2 Has an Ethics Advisor been consulted in the preparation of this Application?
   Yes [x] No [ ]
ATTACHMENTS

Files of the following formats can be uploaded: .doc, .docx, .xls, .xlsx, .pdf, .gif and .jpg

<table>
<thead>
<tr>
<th>Document title</th>
<th>Upload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant Information Sheet</td>
<td></td>
</tr>
<tr>
<td>Email advertisement</td>
<td></td>
</tr>
<tr>
<td>Dataset Framework Survey</td>
<td></td>
</tr>
</tbody>
</table>

SECTION N: APPLICATION CHECKLIST

Please tick below to confirm that you have considered whether the following documents are required for your application and that you have attached them in the attachments section where necessary:

- N:1 Participant Information Sheet
  (see Applicants' Reference Manual Sections 6.3 and 6.4 for explanation and sample)
  ✔

- N:2 Consent Form
  (see Applicants' Reference Manual Section 6.5 for explanation and sample)
  ☐

- N:3 Advertisement
  ✔

- N:4 Questionnaire
  ✔

- N:5 List of Interview Questions
  ☐

- N:6 Confidentiality Agreement
  ☐

- N:7 Observation Schedule
  ☐

- N:8 Any other supporting documents (for example: approval from Course Coordinator, debriefing sheet)
  ☐
FEEDBACK/COMMENTS
If you wish to provide feedback on the usability of this e-form, please do so here.

Please do not use this section to make any comments related to the application itself, as the comments made here will not be included in the application.

Having text areas that are only 3 lines makes it very difficult to review content. Please extend boxes to at least 5 lines (ideally more).

WHEN YOU ARE FINISHED AND HAVE ANSWERED ALL QUESTIONS, TICK THE 'COMPLETE' BOX AT THE TOP OF THE FORM.
K.1 Participants responses of correct answers for the definitions of elements in three data sets in the survey

<table>
<thead>
<tr>
<th>Participants</th>
<th>Dataset 1: CSV</th>
<th>Dataset 2: ARFF</th>
<th>Dataset 3: Tabular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q3(b) Record identifier</td>
<td>Q3(d) Entity label</td>
<td>Q3(f) Metric label</td>
</tr>
<tr>
<td>P1</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P2</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P3</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P4</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P5</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P6</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P7</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P8</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P9</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P10</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P11</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P12</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P13</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure K.1: Distribution of correct answers for definition of elements for each participant

<table>
<thead>
<tr>
<th>Participants</th>
<th>Record identifier</th>
<th>Entry label</th>
<th>Metric label</th>
<th>Measurement value</th>
<th>Metric metadata</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>16</td>
</tr>
<tr>
<td>P2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>P3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>15</td>
</tr>
<tr>
<td>P4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>17</td>
</tr>
<tr>
<td>P5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>7</td>
</tr>
<tr>
<td>P6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>15</td>
</tr>
<tr>
<td>P8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>P9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>P10</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>9</td>
</tr>
<tr>
<td>P11</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>11</td>
</tr>
<tr>
<td>P12</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>15</td>
</tr>
<tr>
<td>P13</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure K.2: Distribution of correct answers for definition of elements for each participant by elements of data set
L.1 Survey: Evaluation of data quality framework for software engineering data sets
Survey: Evaluation of data quality framework for software engineering data sets

Part 1: Task list and observation

Purpose: To allow the participant to understand the quality of data sets using the data quality framework. Please take note, that participant will be observed on how he/she use the framework. Participant can ask question while doing the task. Observation data will be collected during participant doing his/her task.

Instruction: Please read and perform the following task steps.

Task 1: Explore the data quality framework that was given to you.

1. Definitions of dataset elements for the framework is on the Dataset definitions sheet.
2. Formal definitions of data quality issues for the framework is on the Formal definitions sheet.
3. An example of applying the definitions of dataset elements is on the Example 1 sheet.
4. An example of identifying the data quality issues in on the Example 2 sheet.

Task 2: Identify the dataset category elements.

1. Identify the dataset elements by applying the definition of dataset category elements in the given datasets.
2. List the elements (A,B,C,D,E,F,G,H,I) that meet the definition of the dataset category element in the following table.

<table>
<thead>
<tr>
<th>No.</th>
<th>Element category</th>
<th>Definition</th>
<th>Dataset A</th>
<th>Dataset B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Measurement value</td>
<td>An element that obtained through the process of measurement for attribute of entity.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Metric label</td>
<td>A label that represents a metric that measures an attribute of an entity.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Record identifier</td>
<td>A value in a record that represents an entity. (Record is defined as a list of values that are associated with a particular entity)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Entity label</td>
<td>A label that represents a property whose values distinguish between entities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Metric metadata</td>
<td>Metadata that describe the metric used to measure an attribute of an entity.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Entity metadata</td>
<td>Metadata that describe the property whose values distinguish between entities.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Task 3: Identify data quality issues.

1. Identify the data quality issues in the given datasets.

<table>
<thead>
<tr>
<th>No.</th>
<th>Data quality issue</th>
<th>Procedure</th>
<th>Dataset A</th>
<th>Dataset B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Duplicate data</td>
<td>Search for identical record identifiers in the records. If not found, go to procedure No. 3. If found, check whether the measurement values for all metric labels are identical. If this is the case, then these records are duplicate data. List the record identifiers for these records.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Inconsistent data</td>
<td>Search for identical record identifiers in the records. If not found, go to the next procedure. If found, check whether the measurement values for all metric labels are identical. If this is not the case, then these records are inconsistent data. List the record identifiers for these records.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Missing data</td>
<td>Search for records that have either an empty string or null value for a given label. If not found, go to the next procedure. If found, list the record identifiers. If there is no record identifier, list the associated label.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Incomplete metadata for a metric</td>
<td>Search for metric labels that do not have metric metadata. If not found, go to the next procedure. If found, list the metric labels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Incomplete metadata for an entity</td>
<td>Search for entity labels that do not have entity metadata. If not found, go to the next procedure. If found, list the entity labels.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Task 4: Evaluate the metadata of data set

1. Evaluate the metric metadata and entity metadata in the given datasets.

<table>
<thead>
<tr>
<th>No.</th>
<th>Task to perform</th>
<th>Dataset A</th>
<th>Dataset B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Find the entity metadata that describe the entity label.</td>
<td>Visible / Not visible</td>
<td>Visible / Not visible</td>
</tr>
<tr>
<td>2.</td>
<td>Do you think the entity metadata explicitly describe the meaning of the property of the entity?</td>
<td>Yes/ No</td>
<td>Yes/ No</td>
</tr>
<tr>
<td>3.</td>
<td>Find the metric metadata that describe the metric label.</td>
<td>Visible / Not visible</td>
<td>Visible / Not visible</td>
</tr>
<tr>
<td>4.</td>
<td>Do you think the metric metadata explicitly describe the meaning of the property of the entity?</td>
<td>Yes/ No</td>
<td>Yes/ No</td>
</tr>
</tbody>
</table>
Part 2- Background Information.

Instruction: Please answer the following questions.

1. Are you …
   a. Academic researcher
   b. Industry researcher
   c. Postgraduate student
   d. Other ______________

2. How many years doing research?
   a. Less than 5 years
   b. Between 5 to 10 years
   c. More than 10 years

3. Have you used or analysed datasets for your research?
   a. Yes
   b. No

4. If yes, which of the following data repositories contain datasets that you used for research?
   a. PROMISE (PRedictOr Models In Software Engineering)
   b. ISBSG (International Software Benchmarking Standards Group)
   c. SIR (Software Artifact Infrastructure Repository)
   d. SRDA (Source Forge Research Data Archive)
   e. Qualitas Corpus
   f. Eclipse Bug Data
   g. Other ______________

5. Which of the following data quality issues that you have ever encountered with datasets?
   a. Duplicate data
   b. Missing data
   c. Inconsistent data
   d. Incorrect data

6. Please provide any comments on quality issues in data sets.

7. After completing this survey, can you think of obvious ways that the data quality framework could be improved? What are they?

Thank you for your time!

Please let us know if you have any queries about the survey we are conducting. Questions or concerns can either be directed to the researcher, Marshima (marshima@aucklanduni.ac.nz)
L.2 The observation schedule
The observation schedule

<table>
<thead>
<tr>
<th>Task</th>
<th>Data to be collected</th>
<th>Yes/No</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Explore the data quality framework</td>
<td>a. Does participant look at the page containing the definitions of dataset category elements, the definitions of data quality issues and the examples of applying part of the data quality framework? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Identify the dataset category elements</td>
<td>a. Does participant communicate with the researcher? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Does participant look at the page containing the metamodel and the definitions of dataset category elements while performing the task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Does participant complete the given task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Identify the data quality issues</td>
<td>a. Does participant communicate with the researcher? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Does participant look at the page containing the definitions of data quality issues while performing the task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Does participant complete the given task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Evaluate the metadata of dataset</td>
<td>a. Does participant communicate with the researcher? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Does participant look at the page containing the definitions of data quality issues while performing the task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Does participant complete the given task? (Yes/No)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
L.3 Ethics application for an observational study
MEMORANDUM TO:

Dr Andrew Luxton-Reilly
Computer Science

Re: Application for Ethics Approval (Our Ref. 020546): Approved with comment

The Committee considered your application for ethics approval for your study entitled Evaluation of Framework for Data Quality Assessment.

Ethics approval was given for a period of three years with the following comment(s):

1. PIS (Participant) – please proofread for punctuation.

The expiry date for this approval is 01-Dec-2020.

If the project changes significantly you are required to resubmit a new application to UAHPEC for further consideration.

If you have obtained funding other than from UniServices, send a copy of this approval letter to the Activations team in the Research Office, at ro-awards@auckland.ac.nz. For UniServices contracts, send a copy of the approval letter to the Contract Manager, UniServices.

The Chair and the members of UAHPEC would be happy to discuss general matters relating to ethics approvals if you wish to do so. Contact should be made through the UAHPEC Ethics Administrators at ro-ethics@auckland.ac.nz in the first instance.

Please quote Protocol number 020546 on all communication with the UAHPEC regarding this application.

(This is a computer generated letter. No signature required.)

UAHPEC Administrators
University of Auckland Human Participants Ethics Committee
c.c. Head of Department / School, Computer Science
    Assoc Prof Ewan Tempero
    Mrs Marshima Mohd Rosli

Additional information:

1. Do not forget to fill in the 'approval wording' on the Participant Information Sheets, Consent Forms and/or advertisements, giving the dates of approval and the reference number. This needs to be completed, before you use them or send them out to your participants.

2. At the end of three years, or if the study is completed before the expiry, you are requested to advise the Committee of its completion.

3. Should you require an extension or need to make any changes to the project, please complete the online Amendment Request form associated with this approval number giving full details along with revised documentation. If requested before the current approval expires, an extension may be granted for a further three years, after which time you must submit a new application.
**PREAMBLE**

**UNIVERSITY PERSONNEL**

**OTHER PERSONNEL**

**RESEARCH TYPE**

**RESEARCH PROCEDURES**

**PARTICIPANTS & CONSENT**

**STORAGE & RESULTS**

**CULTURAL ISSUES**

**RISKS & BENEFITS**

**HUMAN REMAINS/TISSUE**

**CLINICAL TRIALS & FUNDING**

**ETHICAL SUMMARY & ADVISOR REVIEW**

**ATTACHMENTS & CHECKLIST**

---

**eFORM VERSION**


**PROTOCOL**

Protocol No: 020546

Title: Evaluation of Framework for Data Quality Assessment

**GENERAL**

Prior to completing your application:

- Read the Guiding Principles for conducting research with human participants.
- Go through the Applicants' Reference Manual.
- Check if an exemption applies (see Guiding Principles section 6.1 and Applicants' Reference Manual Section 3.8).
- Check if the matter needs to be referred to a Health and Disability Ethics Committee (HDEC) (see Guiding Principles section 3.1 and 6.1 and Applicants' Reference Manual Section 4.5).

For creating and submitting your application refer to the Human Ethics Module Quick Guide.

All documentation to support your application is on the UAHPEC web page.

For any queries please log a call with the Staff Service Centre at ext. 86000 or staffservice@auckland.ac.nz and it will be referred to the relevant team in the Research Office or ITS.

Please Note: Internet Explorer should not be used when working in InfoEd as many functions are not compatible with this browser. We recommend using the following browsers: Google Chrome, Mozilla Firefox, or Safari (for Mac).

**SECTION A: PERSONNEL**

*Is this a student project to be completed as part of an honours, masters, or doctoral qualification?*

Yes ☑ No ☐

*Please add the degree you are studying towards.*

PhD

If the PI listed below is not your supervisor, please use the 'Change Project Information' button on the protocol screen to change the PI to your supervisor before continuing with the rest of the form. You should also change the department to your PI's department. Instructions on how to do this can be found here.

**PI:** Luxton-Reilly, Andrew J *

☑ I confirm that the PI listed above is my supervisor

**Department:** Computer Science *

☑ I confirm that the department listed above is my supervisor's department

---

1. List all University of Auckland personnel, including the PI, co-investigators, students and ethics advisors, by selecting their name from the UNIVERSITY PERSONNEL page and add their Role from the dropdown list.

2. If you are a University of Auckland student, you must add your own name to the UNIVERSITY PERSONNEL section to have access after
3. To change the Principal Investigator, tick the ‘PI’ checkbox to the left of the relevant person’s name.

Note: A student cannot be a PI. See Applicants’ Reference Manual Section 5.0 that states: “For Doctoral, Masters and Honours research, applications should be submitted by the primary supervisor who will be the Principal Investigator (PI)”.
<table>
<thead>
<tr>
<th>UNIVERSITY PERSONNEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luxton-Reilly, Andrew J</td>
</tr>
<tr>
<td>Mohd Rosli, Marshima B</td>
</tr>
<tr>
<td>Tempero, Ewan D</td>
</tr>
</tbody>
</table>
List below any personnel that cannot be found in the lookup list above. University of Auckland students who are **not** in the lookup list must include a UoA ID number.

<table>
<thead>
<tr>
<th>Full Name (and ID)</th>
<th>Institution / Department</th>
<th>Project Role</th>
<th>Email address</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Is this a Research Project or Coursework Application?

Research

Note: Because you are a student, please select Research
SECTION B: RESEARCH PROCEDURES

B:1 Title.
Evaluation of Framework for Data Quality Assessment

* B:2 Aims/objectives of the project.
The main objective of this project is to study the effectiveness and usability of a framework for evaluating the quality of data sets. This framework also includes a dataset metamodel to describe the structure of data sets, and a quality assessment process to evaluate the quality of data sets. The outcome will assist researchers in understanding the quality of data sets for analysis in empirical research.

Please note: all acronyms must be written out in full the first time they appear in the application, recruiting materials, Participant Information Sheet (PIS) and Consent Form (CF).

* B:3 Summary of the project.
The quality of empirical software engineering results and the validity of results interpretation essentially rely on the quality of the analysed data sets. Researches have indicated that the quality of analysed data sets is critical to the results of empirical studies (Shepperd, M., 2011). Although data sets play a central role in empirical research, few studies consider the quality of the data sets used (Zhang, W., Yang, Y., & Wang, Q., 2011; Mockus, A., 2008). If the quality is poor, then the results of empirical studies cannot be trusted, hence, any models or conclusions based on the data sets are questionable. Ensuring good data quality is therefore, fundamentally important to the field of empirical software engineering.

We have developed a framework that researchers can use to evaluate the quality of data sets in empirical software engineering. The framework includes precise definitions of data sets and their potential elements, and a metamodel to describe the structure and concepts in a data set, and the relationships between each concept. We anticipate that the definitions and the metamodel will allow researchers to clearly understand what data sets are actually intended to represent and identify whether a data set has sufficient information to be usable for analysis in empirical research.

We have also constructed a data quality assessment process to identify quality issues in data sets. This assessment contains a formal definition for data quality issues and allows researchers to clearly identify whether a data set has quality issues that includes issues related to the interpretation of data.

We therefore seek approval to investigate the effectiveness and usability of the definitions in the dataset metamodel and definitions for data quality issues in the quality assessment using a user study. The findings from this study will inform suggestions and recommendations for future improvement of the framework.

* B:4 Project duration (in months).
12

Note: The start date is when the proposal is approved

* B:5 Describe the study design.
This study will be conducted at Universiti Teknologi MARA, Malaysia where the PhD student (Marshima Mohd Rosli) holds a teaching staff position and is currently working in the Universiti Teknologi MARA. Permission to conduct the study will be obtained from the Dean of the Faculty of Computer and Mathematical Sciences.

The participants will be recruited from researchers who have a background or experience in computer science or software engineering research. We will post an advertisement via posters around the Computer Science department (in University Teknologi MARA) to allow the participants to contact the researcher. We will also send an email invitation to the mailing list of potential participants (e.g., mailing list of academic staff). People who are interested in the research project will be asked to reply to the email address provided in the email invitations.

The study will be conducted individually with the volunteer participants. The participants will be given a description about the framework, and examples on how to use the framework. We will ask participants to perform four different tasks in the study. The four tasks are:

a) Explore the data quality framework
b) Identify the dataset category elements
c) Identify the data quality issues
d) Evaluate the metadata in the data sets

To complete these tasks, participants will be given a typical data set (e.g. an Excel spreadsheet of data) and asked to perform activities such as identifying different elements of the data set (e.g. record identifiers, metric names etc.).
We will observe how the participants respond (comments/suggestions) when they are applying the framework to complete the tasks. After performing the tasks, we will ask participants to fill in a very short questionnaire (7 questions) on generic demographic information and research experience. The participants may fill out the questionnaire at their own pace without supervision. Each study session will be timed to not exceed 1 hour. Ideally, we plan to have the session completed within 45 minutes, with an additional 5-10 minutes of informal debriefing before participants leave the session.

We will collect both qualitative and quantitative data. Quantitative analyses will be descriptive and will examine the distribution of all answers instead of only calculating medians. Qualitative analyses of free text answers will be conducted as thematic analysis.

*B:6 List all the methods used for obtaining information.

*Interviews
Yes ☐ No ☑

*Focus groups
Yes ☐ No ☑

*Questionnaires
Please attach the questionnaire(s) in the 'Attachment' section at the end of the form.
Yes ☐ No ☑

*Observations
Please attach the observation schedule(s) in the 'Attachment' section at the end of the form.
Yes ☐ No ☑

*Explain:
We will observe in how the participants use the framework, whether the participants could use the framework in an easy and efficient way. The aspects that we want to observe are:
1) how participants apply the definitions of dataset category elements, the definitions for data quality issues and the evaluation scale of metadata?
2) do participants manage to complete the tasks for identifying dataset category elements, identifying the data quality issues in the datasets and evaluating the metadata of datasets?

The observation will be conducted while the participant perform the tasks. We will ask participants what they think about the framework while they are performing the given tasks in a manner similar to a standard “think aloud” protocol.

The observational data will be recorded via the researcher taking written notes. No recording of video or audio will be conducted.

No personal information about the participant will be collected.

*Other
Yes ☐ No ☑

*B:7 Does the research involve processes that involve EEG, ECG, MRI, TMS, FMRI, EMG, radiation, invasive or surface recordings?
Yes ☐ No ☑

*B:8 Does the research involve processes that are potentially disadvantageous to a person or group (for example, the collection of information which may expose the person/group to discrimination)?
Yes ☐ No ☑

*B:9 Who will carry out the research procedures?
Marshima Mohd Rosli (PhD student).
If necessary, please provide more details of the role(s) of each member of the research team. If the research procedures will be carried out by a third party other than the researcher or co-investigators, please attach a copy of the confidentiality agreement to the "Attachment" section at the end of this form.

*B:10a Where will the research procedures take place?
Please use a maximum of 100 characters (including spaces)
We will book a private room at University Teknologi MARA
If permission is required to conduct the study at a specific location, please attach an appropriate PIS and Consent form, or a support letter, in the "Attachment" section at the end of the form.

*B:10b Will the researcher be travelling overseas to conduct this research?
Yes ☑ No ☐

*Which countries are involved?
Malaysia
Please provide local contact details, as well as those of contacts at the University in the PIS.

B:10c If the study is based overseas, explain what special circumstances arise and how they will be dealt with. Include any special requirements of the country (e.g. research visa) and/or the community with which the research will be carried out.
The research will be conducted in University Teknologi MARA where the PhD student (Marshima Mohd Rosli) holds a teaching staff position. University Teknologi MARA is a governent institution and thus all policy relating to research, privacy and data collection are determined by the institution itself.
Verbal approval has been given by the Dean of Universiti Teknologi MARA for the study to be conducted at the institution. An additional PIS and consent form will be given to the Dean for her formal approval.

Please also provide an undertaking to abide by any local laws relating to research privacy and data collection. For more information, see Applicants' Reference Manual Section 13.6.

*B:11a Is the questionnaire web-based?
Yes ☐ No ☐

*B:11b Is it an anonymous questionnaire?
Yes ☐ No ☐

*B:12 How much time will participants need to give to the research?
Approximately 1 hour

*B:13 Will information on the participants be obtained from third parties?
Yes ☐ No ☐

*B:14 Will any identifiable information on the participants be given to third parties?
Yes ☐ No ☐

*B:15 Does the research involve evaluation of the University of Auckland services or organisational practices where information of a personal nature may be collected and where participants may be identified?
Yes ☐ No ☐

*B:16 Does the research involve a conflict of interest or the appearance of a conflict of interest for the researcher, particularly in a power relationship? Please see the help box for examples.
Yes ☐ No ☐

*B:17 Does the research involve matters of commercial sensitivity?
Yes ☐ No ☐

*B:18 Has the study design or the use of the data been influenced by an organisation outside the University of Auckland (excluding questionnaires developed at other research institutions)?
Yes ☐ No ☐

*B:19 Are you intending to conduct the research in the University of Auckland class time?
Yes ☐ No ☐

*B:20 Does the research involve deception of the participants, including concealment or covert observations?
Yes ☐ No ☐

*B:21 Is there any koha, compensation or reimbursement of expenses to be made to participants?
Yes ☐ No ☐

*B:22a Is this an intervention study?
Yes ☐ No ☐

*B:22b Does this research involve potentially hazardous substances?
Yes ☐ No ☐
**SECTION C: PARTICIPANTS**

*C:1* Who are the participants in the research?

- **Adults**
  - Yes [ ]
  - No [x]

- **Own colleagues**
  - Yes [x]
  - No [ ]

  Explain in Section C7.

- **Own students**
  - Yes [ ]
  - No [x]

- **Persons whose capacity to give informed consent (other than children) is compromised**
  - Yes [ ]
  - No [x]

- **Persons who are in a dependent situation, such as people with a disability, residents of a hospital, nursing home or prison, or patients highly dependent on medical care**
  - Yes [ ]
  - No [x]

- **Persons aged less than 16 years old where parental consent is NOT being sought**
  - Yes [ ]
  - No [x]

- **Persons aged less than 16 years old where parental consent is sought**
  - Yes [x]
  - No [ ]

- **Other**
  - Yes [ ]
  - No [x]

*C:2* How many organisations and departments within the organisations will participate in your project?

- None

If you have letters of support, please attach these in the ‘Attachment’ section at the end of the form.

*C:3* How many individual participants (research participants) will participate in your project?

- Approximately 30

*C:4* How will you identify potential participants and by which method are participants invited to take part in the research?

The participants will be from researchers who have a background or experience in empirical software engineering research. We used two different methods for participants recruitment:

a) An advertisement via posters will be posted around Computer Science department (in University Teknologi MARA).

b) We will approach people who we know are involved in software engineering research. This is because there are limited numbers of people who are engaged in this research that have the expertise to participate in this study. We will send out an email invitation that provides the researcher contact information to allow participants to contact the researcher. (We are aware that this is a direct contact but as the potential participants are experienced researchers, we believe that there will be no power imbalance or undue inducement for people to participate.)

Using a direct approach to recruit potential participants is not recommended.

Please attach the advertisement, media release, or notice, etc. and the letter of permission from the agency supplying them (if applicable) in the ‘Attachment’ section at the end of the form.

*C:5* Who will make the initial approach to potential participants?

- Researcher(s)

*C:6* Will access to participants be gained with consent of any organisation?

- Yes [x]
  - No [ ]

*C:7* Is there any special relationship between participants and researchers?

- Yes [ ]
  - No [x]

  * Explain:
    It is possible that some of the participants may be colleagues of the researcher, or other PhD students engaged in software engineering research.

*C:8* Does the research involve participants in the same organisation as the researcher, where information of a personal nature may be collected and where participants may be identified?

- Yes [x]
  - No [ ]

*C:9* Does the research involve participants who are being asked to comment on employers?

- Yes [ ]
  - No [x]

*C:10* Are there any potential participants who will be excluded?

- Yes [ ]
  - No [x]

* Explain and state the criteria for excluding participants:
This study is intended to be completed by participants who have expertise in software engineering research. Potential participants who are not engaged in such research will be excluded.

SECTION D: INFORMATION AND CONSENT

* D:1 By whom and how will information about the research be given to participants?
The PIS and consent form are distributed before the participants can start the session of the study. The PIS will be distributed either via email or in person.

* D:2a Will the participants have difficulty giving informed consent on their own behalf?
Yes ☐ No ☑

* D:3a If a questionnaire is used, will the participants have difficulty completing the questionnaire on their own behalf?
Yes ☐ No ☑

* D:4 Does the research involve participants giving oral consent rather than written consent?
Yes ☐ No ☑

* D:5 Does the research use previously collected information or biological samples for which there was no explicit consent (excluding already de-identified or anonymous data)?
Yes ☐ No ☑

* D:6 Is access to the Consent Forms restricted to the Principal Investigator and/or the researcher?
Yes ☑

* D:7 Will Consent Forms be stored by the Principal Investigator, in a secure manner?
Yes ☑

* D:8 Are Consent Forms stored separately from data and kept for six years?
Yes ☑
**SECTION E: STORAGE AND USE OF RESULTS**

- **E:1** Will the participants be audio-recorded, video-recorded, or recorded by any other electronic means such as Digital Voice Recorders?
  - Yes
  - No

- **E:3** For the questionnaire, is any coding scheme used to identify the respondent?
  - Yes
  - No

- **E:4a** Explain how and how long the data (including audio-recordings, video-recordings and electronic data) will be stored.
  
  The archival data will be stored indefinitely on a password protected computer and server in The University of Auckland. The initial data collection will occur in Malaysia, so electronic data will be transferred from Malaysia to Auckland via the University Web Dropoff box. Electronic data will be used for future research about data quality assessment.

  The hard-copy data (e.g., consent form) will be stored in a locked cabinet in the students office while she is in Malaysia. When she returns to Auckland, the Consent Forms will be stored in a locked cabinet in the supervisors office at the University of Auckland.

- **E:4b** Explain how data will be used. (For example, in a thesis/dissertation, publications, and/or conference presentations etc.)
  
  The data will be used for the PhD thesis and potentially for publications in journals or conference proceedings.

- **E:4c** Explain how data will be destroyed.
  
  The data will be kept indefinitely.

- **E:5** Describe any arrangements to make results available to participants.
  
  The participants may state on the consent form that they wish to receive a summary of the results via email. Once the study has been completed, a summary will be emailed to all participants that have required a summary.

- **E:6a** Are you going to identify the research participants in any publication or report about the research?
  - Yes
  - No

- **E:6b** Is there any possibility that individuals or groups could be identified in the final publication or report?
  - Yes
  - No
SECTION F: TREATY OF WAITANGI
*F:1 Does the proposed research have impact on Māori persons as Māori?  
Yes ☐ No ☒

SECTION G: OTHER CULTURAL ISSUES
*G:1 Are there any aspects of the research that might raise any specific cultural issues?  
Yes ☐ No ☒
SECTION H: RISKS & BENEFITS

H:1 What are the possible benefits to research participants of taking part in the research?
Upon the completion of this research, participants will have been exposed to a process for identifying potential data quality issues in data sets. If the participants find this to be a useful process then they may benefit by applying the process in their own research to identify any quality issues with the data they use.

H:2 Is the research likely to place the researcher at risk of harm?
Yes [ ] No [X]

H:3 Is the research likely to cause any possible harm to the participants, such as physical pain beyond mild discomfort, embarrassment, psychological or spiritual harm?
Yes [ ] No [X]

H:4 Does the research involve collection of information about illegal behaviour(s) which could place the research or participants at risk of criminal or civil liability or be damaging to their financial standing, employability, professional or personal relationships?
Yes [ ] No [X]

H:5 Is the research likely to give rise to incidental findings?
Yes [ ] No [X]
SECTION I: HUMAN REMAINS, TISSUE AND BODY FLUIDS

*1:1 Does the research involve the use, collection or storage of human tissue, as defined by the Human Tissue Act 2008?

Yes ☐ No ☑
SECTION J: CLINICAL TRIALS

*J:1 Is this project a Clinical Trial?
   Yes [ ] No [x]

SECTION K: FUNDING

*K:1 Have you applied for, or received funding for this project?
   Yes [ ] No [x]
**SECTION L: ETHICAL SUMMARY**

*L:1* Have you made any other related applications?
Yes [x] No [ ]

*Approval reference number(s):*
015880

*L:2* Is there any relevant information from past applications or interaction with UAHPEC?
Yes [ ] No [ ]

*L:3* Please provide a summary of all the ethical issues arising from this project and explain how they are to be resolved:

**Anonymity**
This is a user study that will be conducted face to face, so it is not anonymous. However, the questionnaire does not have any questions that would allow the participant to be identified, and will not be coded or linked to the consent form in any way. Once the questionnaire has been handed in, there will be no way to link the data to the participant, so it will be de-identified at the point of data collection. Participants will be warned not to provide any personal information in their responses. If a participant reveals any identifying information, it will be removed from any analysis or reporting. We will analyse and report the information provided by the participants anonymously.

The observation will focus only on difficulties applying the processes in the framework. The identity of the user will not be recorded with the observation, so the data remains de-identified and may not be withdrawn. This observation is only intended to help the researcher identify flaws in the framework they have developed and no personal or sensitive information will be collected.

All data will be kept secured on a password protected server in the University of Auckland. The consent forms will be distributed and collected separately from the questionnaire.

**Informed consent**
All interested participants will be provided with a full explanation of the study. They will receive the Participant Information Sheet (PIS) explaining the study. The participants have the right to withdraw from the study at any stage while they fill-in the tasks in the study, but once they have completed the study, the collected data cannot be withdrawn because it is not identifiable.

**Possible conflict of interest**
We are aware that the recruitment in this study involves a direct contact with the participant, and that some of the participants may be colleagues of the student researcher. However, the researchers are not in a position of authority that may unduly influence a person to become a participant, and the HoD has provided her assurance that neither participation nor non-participation will have an impact on employment or relationship with the department.

**ETHICAL SUMMARY & ADVISOR REVIEW**

*UAHPEC expects applicants to identify the ethical issues in the project and explain in the documentation how they have been resolved. The application will not be considered if this is not answered adequately. A 'Not applicable' response is not acceptable.*

**SECTION M: ETHICS ADVISOR REVIEW**

*M:1* Will this Application be reviewed by an Ethics Advisor after you submit it?
Yes [x] No [ ]

*M:2* Has an Ethics Advisor been consulted in the preparation of this Application?
Yes [x] No [ ]
ATTACHMENTS

Files of the following formats can be uploaded: .doc, .docx, .xls, .xlsx, .pdf, .gif and .jpg

<table>
<thead>
<tr>
<th>Document title</th>
<th>Upload</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIS for Participants</td>
<td>✔️</td>
</tr>
<tr>
<td>CF for Participants</td>
<td>✔️</td>
</tr>
<tr>
<td>Advertisement</td>
<td>✔️</td>
</tr>
<tr>
<td>Observation Schedule</td>
<td>✔️</td>
</tr>
<tr>
<td>Email advertisement</td>
<td>✔️</td>
</tr>
<tr>
<td>PIS for HoD</td>
<td>✔️</td>
</tr>
<tr>
<td>CF for HOD</td>
<td>✔️</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>✔️</td>
</tr>
</tbody>
</table>

SECTION N: APPLICATION CHECKLIST

Please tick below to confirm that you have considered whether the following documents are required for your application and that you have attached them in the attachments section where necessary:

N:1 Participant Information Sheet
(see Applicants' Reference Manual Sections 6.3 and 6.4 for explanation and sample)

N:2 Consent Form
(see Applicants' Reference Manual Section 6.5 for explanation and sample)

N:3 Advertisement

N:4 Questionnaire

N:5 List of Interview Questions

N:6 Confidentiality Agreement
(see Applicants' Reference Manual Sections 6.8 for explanation and sample.)

N:7 Observation Schedule

N:8 Any other supporting documents (for example: approval from Course Coordinator, debriefing sheet)
FEEDBACK/COMMENTS
If you wish to provide feedback on the usability of this e-form, please do so here.

Please do not use this section to make any comments related to the application itself, as the comments made here will not be included in the application.

WHEN YOU ARE FINISHED AND HAVE ANSWERED ALL QUESTIONS, TICK THE 'COMPLETE' BOX AT THE TOP OF THE FORM.
M.1 Participant Information Sheet and Consent Form for participant
PARTICIPANT INFORMATION SHEET

Participant

Name of Researcher: Marshima Rosli

Researcher Introduction
I am Marshima Rosli, a PhD student in the Department of Computer Science at The University of Auckland under the supervision of Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly.

Project description and invitation
I am conducting research developing a framework to evaluate the quality of software engineering data sets for analysis in empirical research. The framework consists of a dataset metamodel and a quality assessment. Part of our research involves an evaluation of this framework regarding its usability and effectiveness for describing the content of data sets, and identifying data quality issues in data sets. If you are a researcher (postgraduate or academic) who has conducted any research in software engineering or computer science, you are invited to participate in this survey. Your comments and assistance would be greatly appreciated.

Project procedures
This research involves the use of a written task list and questionnaire, which will take approximately 1 hour to complete. You will be informed about the time and the location where the user study session will be held in the email. A task list and questionnaire sheet will be given to you before you start using the framework. You will be requested to perform a number of tasks on the framework and once you completed the task, you will be asked to answer the questions in the questionnaire sheet given to you. You also will be observed to allow researcher to learn whether the framework is easy to use and also to know more about the usefulness and acceptance of the framework. The observations will take place only while you perform the tasks on the framework. There will be note-taking while you perform the tasks and also while you are responding or commenting when using the framework. However, no personal information will be collected in this observation process. Audio-tape, vide-tape and any other electronic means such as digital voice recorders are not used in this evaluation.

After completing the tasks you will be asked to answer the questions in the questionnaire sheet. Once you have completed the questionnaire, you need to put it in the box that will be placed in the room. There will be no coding to your questionnaire as it is treated anonymously.
Data storage/ retention/destruction/future use
The observation and questionnaire data will be archived and kept indefinitely on a password protected computer and server in The University of Auckland via University Web Dropoff box, and will be used for future research about data quality assessment. The data will be analysed and the results of this research will be used for a PhD thesis and other academic publications. The results and publications resulting from this study will be available from the researcher upon request.

Right to withdraw from participation
Participation in this survey is on a voluntary basis and there will be no financial compensation. The survey is performed in an anonymous way. No personal information will be collected during the survey. You can be assured that neither your grades nor employment performance be affected by either refusal or agreement to participate. This assurance is given by the Computer Science Head of Department. You can withdraw yourself from the survey at any time. Completing the required tasks in the survey and submitting the evaluation is an indication of consent but as the evaluation is anonymous, no answers can be withdrawn once the evaluation is submitted.

Anonymity
The data that will be collected will only be used for academic purposes. The responses will be completely anonymous as you will not be asked to write your name on the responses. The responses will be kept on a computer and only can be accessed by the researcher as it will be protected by a password. The researcher will not be able to trace you through your responses, and no one will know whether or not you participated in the study.

Contact details
If you have any queries regarding this survey, please do not hesitate to contact me. You can email me at marshima@tmsk.uitm.edu.my or call me (+60355442000 ext 1192). You may also contact my supervisors, Associate Prof. Ewan Tempero (+6493737599 ext 83765) and Dr. Andrew Luxton-Reilly (+6493737599 ext 85654).

For any queries regarding ethical concerns you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland 1142. Telephone 09 373-7599 extn. 83711. Email ro-ethics@auckland.ac.nz

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 1 DEC 2017 FOR 3 YEARS, REFERENE NUMBER 020546.
PARTICIPANT CONSENT FORM

THIS FORM WILL BE HELD FOR A PERIOD OF 6 YEARS

Name of Researcher: Marshima Rosli
Supervisors: Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly.

I have read the Participant Information Sheet, have understood the nature of the research and why I have been selected. I have had the opportunity to ask questions and have them answered to my satisfaction. I understand that I can withdraw at any time but that data already recorded cannot be withdrawn. I agree to take part in this research.

- I understand that this consent form will be stored separately from the questionnaire for a period of 6 years before it is destroyed.
- I understand that this consent form will not be used to associate my handwriting with the survey.
- I understand that all of the data collected from the survey will be non-identifying.
- I understand that I will be observed while doing a task on the framework if I agree to participate in this survey. No audio-tape, video-tape or any other electronic means such as Digital Voice Recorders is used in this survey.
- I understand that only the researcher and her main supervisor will have access to the questionnaire and observation data.
- I understand that the Computer Science Head of Department have provides assurance that neither my grades nor employment performance will be affected by either refusal or agreement to participate
- I understand that data will be stored and used for the PhD research of the student for 6 years, after which they will be destroyed.
- I wish to receive a summary of findings, which can be emailed to me at this email address:___________________________

Name __________________________
Signature _________________________
Date ____________________________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 1 DEC 2017 FOR 3 YEARS, REFERENCE NUMBER 020546.
M.2 Participant Information Sheet and Consent Form for Organisation
PARTICIPANT INFORMATION SHEET
Head of Department
Faculty of Computer and Mathematical Sciences
University Technology MARA, Malaysia

Name of Researcher: Marshima Rosli
Supervisors: Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly.

Re�eacher Introduction
My name is Marshima Rosli, and I am a PhD student in the Department of Computer Science at the University of Auckland under the supervision of Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly. I am also a teaching staff under your department (Computer Science Department, Faculty of Computer and Mathematical Sciences, University Technology MARA, Malaysia.)

I am conducting research developing a framework to evaluate the quality of software engineering data sets for analysis in empirical research. The framework consists of a dataset metamodel and a quality assessment. Part of our research involves an evaluation of this framework regarding its usability and effectiveness for describing the content of data sets, and identifying data quality issues in data sets.

I seek your consent to conduct this study at your department. Our study will involve participation from researchers (postgraduates or lecturers) who has conducted any research in software engineering or computer science. Participation from the researchers involves:

a) Completing a task list
The task list consists of four tasks relating to the data quality framework. This task list will take approximately 40-45 minutes. I will observe on how the participants use the framework. I will also ask participants to express their comments or suggestions while performing the tasks.

b) Completing a questionnaire
The questionnaire consists of seven questions about participant background and research experience. This questionnaire will take approximately 10-15 minutes.

The observations will take place only while participants perform the tasks on the framework. There will be note-taking while participants perform the tasks and also while participants are responding or commenting when using the framework. However, no personal information will be collected in this observation process.
I will be using data from the task list, questionnaire and observation to learn whether the framework is easy and efficient to use and to know more about the usefulness and acceptance of the framework.

Participation in this study is on a voluntary basis and there will be no financial compensation. The study is performed in an anonymous way. No personal information will be collected during the study. We would like you to provide us the assurance that participation or non-participation from the postgraduate students and department staff members will not impact their grades or employment. Your support would be greatly appreciated.

The data collected will be archived and kept indefinitely and will be securely destroyed after three years. At the conclusion of the study, a summary of the findings will be available from the researchers for the participants to view upon request. Any publications resulting from this study will also be made available upon request. Participants will not be identified in any publications that result from this study.

Thank you very much for your time and help in making this study possible. If you have any questions at any time, you can email me at marshima@tmsk.uitm.edu.my or call me (+60355442000 ext 1192). You may also contact my supervisors, Associate Prof. Ewan Tempero (+6493737599 ext 83765) and Dr. Andrew Luxton-Reilly (+6493737599 ext 85654).

For any queries regarding ethical concerns you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland 1142. Telephone 09 373-7599 extn. 83711.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 1 DEC 2017 FOR 3 YEARS, Reference Number 020546.
CONSENT FORM
Head of Department
Faculty of Computer and Mathematical Sciences
University Technology MARA, Malaysia.

THIS FORM WILL BE HELD FOR A PERIOD OF 6 YEARS

Name of Researcher: Marshima Rosli
Supervisors: Associate Prof. Ewan Tempero and Dr Andrew Luxton-Reilly.

I have read the Participant Information Sheet, and I have understood the nature of the research. I have had the opportunity to ask questions and have them answered to my satisfaction.

- I agree to allow the researcher to have access to the postgraduate students and academic staff members in the Computer Science department.
- I give my consent for this study to be conducted at my department between 1 Dec 2017 and 1 Dec 2020.

Name ___________________________ Position __________________
Department ______________________________________________________

Signature ___________________________ Date _______________________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 1 DEC 2017 FOR (3) YEARS REFERENCE NUMBER 020546
Email subject: A survey on evaluation of framework for data quality assessment.

Dear respondent,

I am Marshima Rosli, a PhD student in the Department of Computer Science at The University of Auckland. I am conducting a survey on evaluation of framework for data quality assessment as part of my PhD research on evaluating the quality of data sets. The purpose of this survey is to evaluate the usability and effectiveness of the framework for describing the content of data sets, and identifying data quality issues in data sets.

If you have conducted research in software engineering or computer science, you are invited to participate in this survey. Please reply to this email and I will contact you to arrange a session to fill out the survey.

Your participation in the survey is completely voluntary and all your responses will be completely anonymous. Information provided will be stored electronically on a password protected computer and server in The University of Auckland and will be accessible to the researchers only. The data from this survey will be analysed and the analysis results will be used in a PhD thesis, journal articles and conference papers.

If you would like further information on this survey, please contact me at mmoh603@aucklanduni.ac.nz. If you know any potential participants for this survey, feel free to pass this advertisement.

Thank you very much for your time and cooperation.

Marshima Rosli
PhD Student
The University of Auckland

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 1 DECEMBER 2017 for 3 years, Reference Number 020546.
Bibliography


