An eScience tool for understanding Copyright in Data Driven Sciences

Richard Hosking[1,2], Mark Gahegan[1,2], Gillian Dobbie[1]
Dept of Computer Science[1], Centre for eResearch[2]
The University of Auckland
Auckland, New Zealand
{r.hosking; m.gahegan; g.dobbie}@auckland.ac.nz

Abstract— **Understanding the impacts of copyright is a challenge for the sharing and reuse of our research data. There is growing recognition of the problem, but the legal knowledge required to navigate through the minefield of restrictions and risks is often too difficult to uncover and understand. As of yet there are no appropriate tools to aid researchers, librarians and research policy makers. To address this gap we present Camden, an automated copyright reasoning tool designed to integrate into existing research workflows. At its core, Camden uses dynamically generated defeasible rules to reason over the legality of a situation of using, combining and publishing data, while additionally suggesting potential licenses by which to safely share derived research outputs. This functionality has been wrapped up into an embedded software library and offered as a web application. In this paper we introduce Camden, describe its model of computational reasoning and discuss how it can be included into existing and future eResearch tools and services.**

Keywords—Camden, eScience, Copyright, Licensing, Data Science, Research Data Management

I. INTRODUCTION

The research process is undergoing an explosion in the quantity of information available. Underlying the technological advances driving this change is a paradigm shift towards a networked way of doing science [1]. Ensuring our data is (re)usable is essential for reaching the potential offered by our diverse range of research tools and services. As data collections become established in key disciplines, we must consider the social obstacles to the effective cycle of learning and sharing.

One such obstacle overshadowing this cycle is copyright: copyright has the potential to stem the flow of essential information, creating a frustrating situation where each technological advance and data source is greeted by a range of seemingly distant legal questions. Copyright is a legal framework that creates exclusive usage rights for the authors of intellectual or creative works. As illustrated in Figure 1, for researchers this means that licensing is currently an essential, yet flawed, necessity for allowing our colleagues access to reuse and share the data we create.

Metadata and ontologies help overcome the problems of finding and interpreting data, and the development of more expressive and reusable workflows help us to share and reuse our methodologies. However the lack of clarity over licensing remains a real impediment to reproducibility and reuse of our scholarly outputs. Whereas freedom from legal restriction and uncertainty is our desired state, we must also deal with the realities of our current situation—which is data collections protected by various families of often-incompatible copyright licenses. This issue will only become more pressing with: (i) the relative speed of research practice outpacing our ability to gain reliable copyright advice; (ii) a greater emphasis on the sharing and reuse of data; (iii) increasing cross-discipline, and international collaborations (iv) a greater ambition for collaborative endeavors in: data generation, data publishing, data analytics and data archival and (vi) an increasing emphasis on the commercialization of research.

Effective licensing strategies have the potential to alleviate the issue, allowing our knowledge to flow through our systems and communities. But we observe the following tensions that force us to consider the current needs of researchers who are faced with navigating this issue: Firstly, institutions, not individual researchers, are often the owners of the copyright; thus we are somewhat reliant on policy changes as well as on individual changes of behavior. Secondly, we acknowledge that there is a legitimate role for commercial data suppliers, who need to assert licenses over their effort and intellectual property to stay in business. Lastly, we cannot focus purely on effectively licensing new research at the expense of the existing data caught up in a web of copyright issues; reuse of this existing data is essential. While we should endeavor to clear restrictions from this content, research projects cannot wait indefinitely until we do.

Understanding how the threads of rights, conditions and obligations combine and endure through derived research outputs is at the heart of this challenge[2]. The common act of combining several datasets during the course of an analysis presents a legal challenge, that of identifying the provenance (and thus licensing) of the original contribution and each separate, derived component. Our ability to discern the authors or copyright owners through chains of derived works is currently dubious at best. Our belief is that the expert knowledge required to navigate this issue should be provided to researchers in a meaningful way at pertinent times –during the course of planning, conducting and publishing research. We believe this will greatly aid researchers in determining their rights and enable them to make better-informed licensing decisions during the publication of their own work. Such a solution helps lower the barrier for gaining legal advice, provides comfort and assurance, and speeds up modern

research flows that are currently plagued by niggling copyright uncertainties.

In this paper we present *Camden*, a legal reasoning tool that aims to clearly describe the legality of reusing and sharing research data. We show its applicability to some common research activities, such as workflows, software tools and through a web interface. We promote the use of defeasible rules for this task, which in contrast to the strict subsumptive capabilities of description logics enable a richer and more nuanced understanding of licenses to be captured and used. By using constructs from argument theory such as schemes [3] and proof burdens [4], which have been realized in the Carneades argumentation framework [5], we are able to model the legal interpretation of copyright law in the context of academic practice. We illustrate the value of this tool with three use cases. The tool alongside its source code is freely available under the open source Mozilla public license.

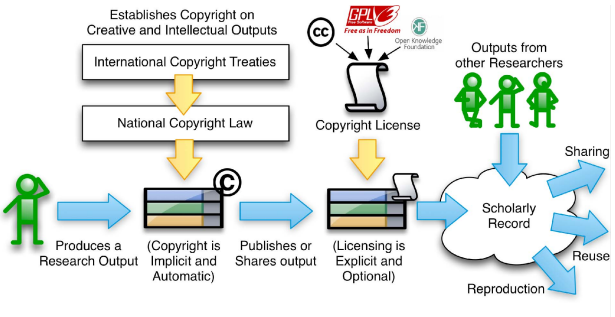

Fig. 1. Understanding the flow of Copyright and Licensing on research data

II. Objectives

The challenge we set for ourselves in this research is to provide meaningful copyright advice in-step with common research workflows. We observe that the terminology commonly used within legislation and licenses is abstract (and legally terse), thus requiring careful interpretation to relate back to everyday research tasks – such as merging two datasets and publishing the results. In making these interpretations, researchers are thus asked to synthesize understanding from a variety of sources using a set of unfamiliar legal skills. In this research we propose to address the following questions plaguing data reuse, recombination and subsequently sharing:

(1) Does the current state of rights permit the desired actions?

(2) Can I legally share the results of this work, and if so under what licenses may I do so?

(3) How can this guidance be integrated with existing and future research workflows?

A. *Determining the legality of obtaining and reusing data*

The process and key decisions for determining the legality of reusing data produced by others is illustrated by the flowchart in Figure 2. We aim to capture this process in order to provide automated guidance.

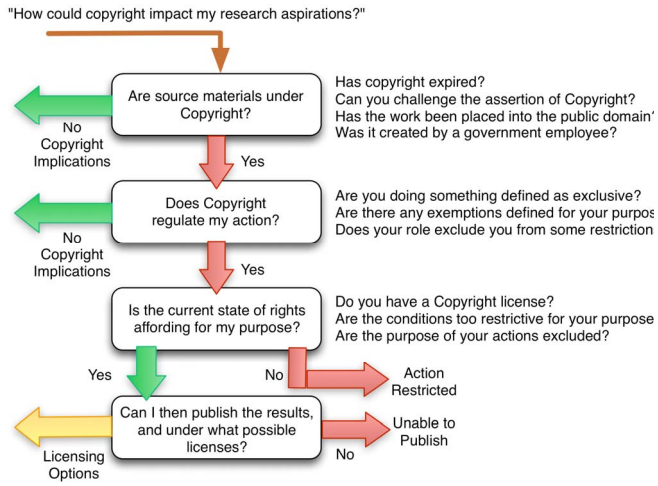

Fig. 2. A flowchart of decisions relating to the use and publication of copyrighted data

B. *Determining the legality of sharing the products of your research, and under what possible licences*

Not only does copyright protect unauthorized use of data, it also places constraints on the publication of derived products. Thus, before we can share our work, we must also consider our own licensing options and responsibilities, which are constrained by the terms and conditions placed on the original data used. We aim to capture these constraints in order to provide guidance for this task.

C. *Effectively integrating this reasoning into research workflows*

Current research tools and services orchestrate quick and complex manipulation and combination of data resources. Unwelcome inertia will thus build when relevant copyright guidance cannot keep pace. As illustrated in Figure 3, pertinent guidance requires sensitivity to context – making an incorrect assumption may result in overly restrictive or potentially risky conclusions. To give an idea of the complexities, the legal interpretation of a license can differ depending on the jurisdiction (country) and the individual's purpose for undertaking the task. Any computational tool developed must appreciate the importance of the advice it provides, thus guidance needs to be trustworthy with transparent and defensible reasoning which can be extended to suit contextual requirements.

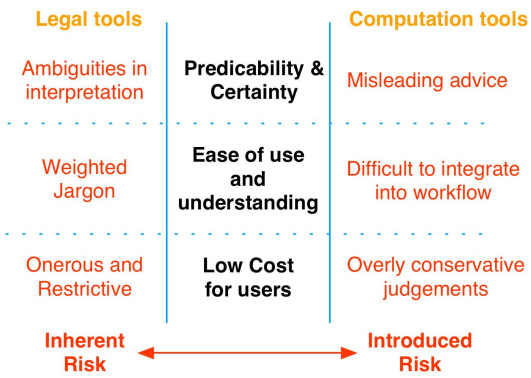

Fig. 3. Inherent and Introduced Legal Risk

III. USE CASES

We illustrate Camden's applicability through three common scenarios:

Searching for data: You have determined the analysis you wish to perform and are now searching for data fit for purpose. By gathering the rights-statements of the potential data, attached as metadata or by manual input (in the case where this knowledge is not appropriately provided), Camden can determine the set of data able to be legally used and combined, the data where rights must be negotiated with the copyright owners, and what constraints you may have on any derived work.

Planning collaborations: You are in the early stages of planning a collaborative project. You are arranging access for your colleague to your existing data and you wish to understand your licensing options. For the most effective use of your data, the license chosen will have to interoperate with licenses on your collaborator's existing data and support any required licensing terms on the derived data.

Creating an automated workflow: As part of your research, and utilizing a community infrastructure, you wish to create a repeatable workflow that will ingest new data as it becomes available. As part of the design requirements you want this automation to avoid creating legal troubles. You may want to exclude data not under license or with incompatible licenses to your requirements – perhaps marking these for manual consideration at a later date.

IV. INTRODUCING CAMDEN

Camden is an open-source, cross-platform tool that provides legal guidance on issues concerning copyright. It's namesake, Charles Pratt, 1st Earl of Camden, led the rejection of common law (essential perpetual) copyright in the House of Lords in the UK in 1774, and is perhaps one of the earliest proponents of what we now call the Open Access movement. The design and functionality of Camden is illustrated in Figure 4.

One of the initial tasks in developing Camden was to determine a concrete and relevant set of support it could provide. It was determined that in order to be an effective tool it must provide advice in 3 key areas, that is: obtaining, using and sharing research data, and importantly taking into account situation with multiple interacting licenses.

We dismissed developing the required reasoning in an OWL ontology due to incompatible qualities of the reasoning produced. What we require is not a shared medium to communicate the domain facts and their relationships but rather a tool to determine the legal truth of a given situation. As noted by Gordon [6] and many others, legal arguments are not primarily deductive, but rather a modeling process of shaping an understanding of the facts, based on evidence and an interpretation of the legal sources, to construct a theory for some legal conclusion. In response we turned to defeasible logics[7], which are a non-monotonic formalism suitable for modeling situations with incomplete and contradictory information.

The value of this choice is best understood through an example. In making decisions Camden must thread multiple sources of contradictory information. The default position of copyright law dictates restrictions on use, but in the same situation a license may state it permits that same use - perhaps conditionally. The outcome may further be contradicted by a relevant exemption (often dubbed fair use). Defeasible logics allow us to explicitly prioritize these competing rules. In contrast without this feature we would reach an impasse in determining the resulting legality.

Camden is built on the Carneades argument framework [5], which has been developed into an open-source reasoning tool. As noted by Governatori [8], Carneades implements a type of defeasible logic, but through an abstraction of argumentation. Constructs such as Schemes [3], burden of proof [4] and the idea of dialectic reasoning [9] provide a rich tool set for modeling complex legal and social issues.

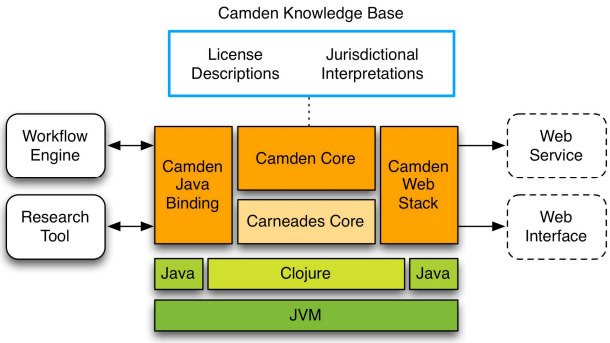

Fig. 4. Camden System Diagram

Written in Clojure (extending Carneades) Camden is compatible with applications written in Java. A feature of the Clojure language is that it runs on the Java Virtual Machine (JVM), providing two-way compatibility with Java; it can both make calls to existing Java libraries, and Java applications can make calls to the denoted Clojure methods. Camden has also been integrated with a web stack that can provide a rich browser interface backed by a reusable web service. The web stack is realized by utilizing the Clojure Ring and Compojure libraries alongside the Angular and Bootstrap Javascript libraries for building the dynamic HTML5 web interface.

A. *Example Usage*

We introduce the functionality of Camden via a tour of the web interface. We present this as a series of steps, along with accompanying use cases taken from the perspective of the user in order to make it more concrete.

Step 1: User input. The initial task when interacting with the web interface is to collect the facts of the situation, illustrated with the examples below. A key contribution is the ability to choose an arbitrary number of differing licenses, connected with the nature and purpose of the action.

Step 2: Obtaining and Using. The main purpose of step two is clearly presenting the legality of obtaining and using data –given the situation described in the previous step.

Step 3: Sharing and Licensing. Step 3 recognizes an important aspect of research, that of sharing the results of investigations. To this end guidance is provided on the legality of sharing and the licenses under which this can be done.

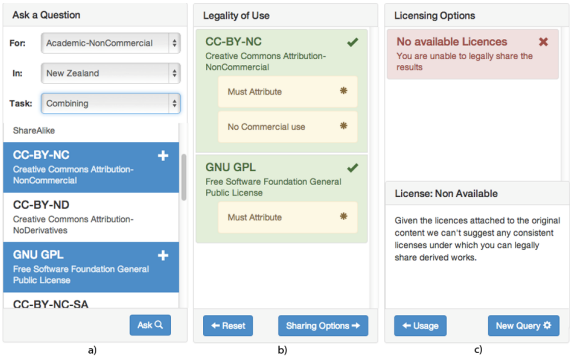

Fig. 5. Use Case One: Combining Data Under Open Licences

1) *Use case 1*

- Figure 5 shows the use case where the desired task is a routine data integration effort, combining two data sets for analysis and then publishing or sharing the result. The input for this situation, shown in a), involves two commonly considered 'open licenses' - the Creative Commons attribution non-commercial license (CC BY-NC), and the Free Software Foundations general public license (GPL).
- Shown in b), we see the results of the query. In this case, given the academic nature of the task, and the licenses used we have no problems –indicated by the styling on the interface. For further reassurance we present the conditions of use for each license – indicating that while they exist, given the circumstances, they do not pose an issue to the task.
- Presented in c), we see the possibly counter intuitive result of trying to share derived work from two 'open licenses'. In this case, the GPL license enforces all derived work to be shared under an identical license, something that is not supported by the (CC BY-NC). Understanding this outcome in advance allows the researchers to plan ahead, potentially negotiating with the copyright owners for different licensing terms, or perhaps just searching for replacement data.

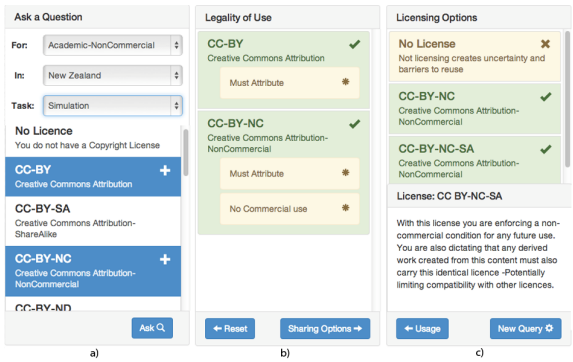

Fig. 6. Use Case 2: Using Data as into to a simulation

2) *Use Case 2*

- In Figure 6, a similar use case is presented, this time involving two Creative Commons licenses - the attribution license (CC BY), and the attribution non-commercial licenses (CC BY-NC). The purpose this time being to run a simulation, taking in data under the previously mentions licenses, to produce an output.
- In b), we also see a similar result to the previous case - the licenses and the non-infringing conditions are displayed appropriately. But, for example changing the purpose of the action from a purely academic simulation to include a commercial interest would result in a glaring red styling for the Creative Commons Attribution, Non-Commercial license, indicating an issue.
- Presented in c), shows different outcome than the previous scenario. In this case we have multiple options, as shown by the (CC BY-NC) and (CC BY-NC-SA) licenses –and the option to not license the products. Within Camden, this list is compiled by taking the set of known licenses, and filtering them with the constraints imposed by the licenses of the original content. In this case given that the user has a choice to make, we present the impact of each licensing choice –especially the conditions that will be imposed on further use.

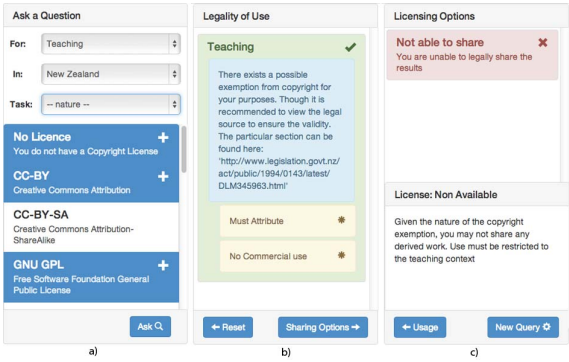

Fig. 7. Use Case 3: Creating a lesson plan that requires preparing data

3) *Use Case 3*

- In Figure 7, we observe a difference case; here a lesson plan is being prepared for an undergraduate lab. Part of the lesson will be using data from a variety of sources to process in a GIS application. The difficulty being the multiple licenses, and unlicensed desired content.
- In b), we see an interesting result. Given the situation, Camden has determined that New Zealand Legislation defines a possible exemption for the purpose of teaching (normally, content not licensed would have restricted use). This presents an example of the further information that Camden can present –the factors that could determine the final judgment could benefit from human involvement. We believe Camden is capable of making these calls, but this shifts the burden to a knowledge elicitation exercise.
- Lastly c), displays a warning indicating the inability to share the data. This result stems from the conditions attached to the exemption. Explaining this result, like the licensing choices, is an important factor in order to convince users of its validity, and more importantly the extent of its applicability - for instance the clarification that you can share to students, as they are part of the educational context.

V. Discussion and Evaluation

A. Supporting research workflows

In order to fit in with a variety of research workflows, we offer Camden as both a stand-alone Java library, and a stand-alone web server able to provide a JSON endpoint, utilized by the web application shown in Figures 5, 6 and 7. The functionality of the web and java interfaces, is identical and illustrated in Figure 8. The difference is the programming interface offered.

By wrapping this reasoning in a simple and easy-to-use query structure we can provide insight into: (1) the allowance of each license for creating derived content, (2) the combined legality of generating derived data, (3) the legality of sharing this data, and (4) the set of available licenses by which it can be shared. Future versions will likely include additional functionality, but we believe any further requirements can only be usefully developed through experience and community feedback.

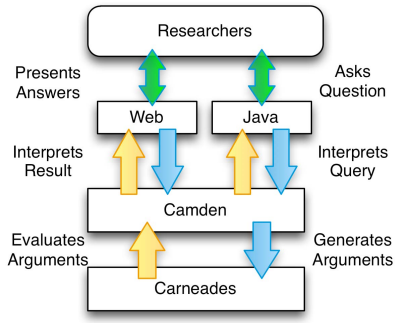

Fig. 8. Illustrating Camden's Query interfaces

B. Facets of Reasoning

Development of Camden is ongoing, but by focusing on the most important aspects we have constructed a useful tool to help understand the copyright currently attached to our research data. We handle the following key factors, as shown in the examples, which aid in determining the legality of situations involving copyright:

- The *Purpose* underlying what the researcher is doing – an important facet when understanding exemptions offered under many legal jurisdictions.
- The *Country* or legal jurisdiction in which tasks are being undertaken. This dictates the national laws that govern the specific legality of the task.
- The *Nature* of usage –also an important determinant in understanding fair usage rights.

The choices offered for each option are a representation of the knowledge currently understood by Camden. Populating the knowledge base is an ongoing process. This is a task likely beyond the time and knowledge of any one individual, thus crowd sourcing this knowledge may be an interesting avenue. But this route will not be a trivial undertaking; the quality of the submitted facts is crucial to the value of the system.

C. Using Defeasible Rules

We make use of the defeasible properties of the legal rules we create and evaluate within Camden. Take for example the case presented in Figure 7 (preparing a lesson plan that involves the use of data from multiple sources). We can see an illustrative slice of the reasoning shown in Figure 9, chosen to illustrate the importance of the defeasible logics and the role priority plays in producing the results.

As we know, the default position of copyright prohibits unauthorized use. But as we can see, both the CC-BY and GPL licenses entail the rights to be available. Though given we still wish to use data from an unlicensed source this undercuts the availability of the required rights (that is, the priority of unlicensed content takes priority over the licenses).

In this example we have a second undercutting argument - the exemption for teaching, made true by the stated purpose and jurisdiction. While the required rights are still not gained, their need is negated, thus allowing the action to legally proceed. An essential component of this process is presenting this to the user. We dynamically generate the user interface based on queries on the state of the argument graph generated, providing not only the final result but also important steps along the way.

A similar reasoning process occurs when determining the legality of sharing the result. In this case, the entailments indicates the user does not have the required rights in order to share any results, and the particular exemption provides no affordances to counter that (resulting in what is shown in figure 7 c).

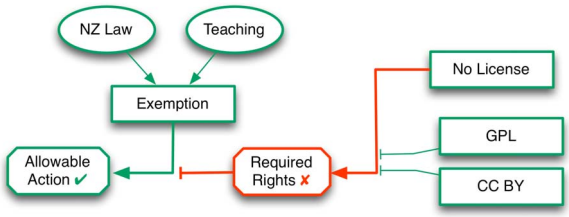

Fig. 9. Highlighting the use of Defeasible rules in Camden

D. Additional Questions

Through Camden we don't claim to be providing all the legal answers; many of these questions are currently being played out in the courts. But what we do provide is a place to capture the precedents and rulings once they become clear. By providing a programmatic framework, and the accompanying user interface we can begin exploring further questions that play an important role in understanding copyright. For example, capturing the type of artifact in question would open the way for greater accuracy. For instance, are there legal differences between, the use of an API, a data schema or ontology, or perhaps the output of sensory data?

Another important factor, and a question in need of answering is the debate over what qualifies for copyright and the duration of copyright exclusivities. For example, we can consider the following factors (not currently captured) in making this decision: 1) the cross-jurisdictional recognition of copyright, b) copyright durations across jurisdictions, c) the criteria of originality of the work, and lastly d) the required level of intellectual effort or creativity. Both c) and d) have sparse precedent and are especially untested in terms of the criteria for machines-generated data (i.e. the results of simulations or sensor networks). At issue is the differentiation between pure (not copyrightable) facts about the world and their creative interpretations.

VI. Related Work

A. Legal Understanding

Copyright was born out of the printing age, with the Statue of Ann, the first 'Copyright like act' passed by the British parliament in 1710 [10]. Originally argued as a means to ensure the quality of published books, it evolved driven by arguments to protect investments of capital [11]-[13]. But we can also view the history of copyright legislation as a means of controlling the disruptive potential of successive new technologies, such as the printing press, photocopiers, tapes and the recent digital revolutions [14] – a troubling idea for eScience tools.

Contemporary copyright Acts are often communicated as efforts to protect authors' rights. But the underlying premise (heavily lobbied for) is that print, radio broadcasting, television broadcasting, sound recording and movies are all capital-intensive to produce, and thus deserve special protections. Now almost universally, products of intellectual effort captured on durable media receive legal rights analogous to those given to physical property. These rights are conferred without any form of registration, and grant the author or creator the sole right to decide how, and for what reason (if any), the work can be used and shared. Of particular relevance to researchers are the economic rights to duplicate (copy), adapt, distribute and publicly communicate, together with the moral rights of integrity and attribution. This is because, *in the absence of any explicit authorization, these actions are by legal default exclusive to the owner and the rights must be transferred or waived in order for others to make use of the work.*

TABLE I. Quick reference to a modern understanding of Copyright in the Sciences

Copyright History	Kaplan[12], Patterson[11], Deazley et al[13], Eisenstein[15], [16]
Recent History	Harvey[14], Hamilton[17], Hargreaves[18], Tehranian[19], David[20], Rife[21], Hugenholtz[22]
Contemporary Science and Copyright	Wilbanks[23]-[25], Korn[26], Hagedorn[27], Hilty[28], [29], Reichman[30], [31], Stodden [32], [33]
Licensing Issues	Wilbanks[34]-[36], Stallman[37], Stodden[32], Mathews et al[38], Williams et al[39]
Copyright Advocacy	JISC, Open Knowledge Foundation, Creative Commons, Science Commons

A recent history of legislative amendments to copyright laws has seen longer durations of exclusivity [17], [40], larger punitive damages and greater reach given to copyright owners; while inadequately adapting to technological and cultural shifts, and ignoring freedoms for scientific research [31]. Regardless of the technical ease, as it stands, legally utilizing the creative work of others require explicit permission from the author or copyright owner. This can be fundamentally at odds with the iterative nature of scientific discovery: in effect it means that in order to free our colleges from legal risk when sharing, we must license or waive copyright on our work. This growing conflict between private rights and public good [20] has only been further complicated in recent years by the enactment of so called database 'protection' laws and a series of multinational trade treaties.

Rather more positively, there is evidence that our legal tools are also advancing to meet our evolving conceptualization of openness. Public or Open licenses place greater emphasis on the commons - that is opening up our creative and scientific works for collective benefit. The distinguishing features of these licenses are their focus on the freedoms or affordances of the end-users, as opposed to the leveraging of rights. Though in practice, not all public licenses are created equal or necessarily interoperate with one another. The Free Software Movement's [37] virally 'open' licenses – challenging the closed and proprietary model of software development – includes explicit provisions that derived works must also carry similar affordances. This particular approach of propagating values brought with it a cultural movement. With the advent of the Creative Commons family of licenses we observe a shift towards facilitating the public domain, bringing a much-simplified licensing model while also employing the use of graphical notations of reuse conditions to improve clarity. Some affordances of control remain for authors, such as preventing commercial use, and – borrowing from the Free Software Movement – the share-alike clause to allow propagation of the intent to allow derivative works. Most recently the legal commitment to the commons was strengthened with the creation of waivers, such as CC-Zero[1], and the Public Domain Dedication and License (PDDL)[2], designed to release (to the fullest extent possible under the law) all conditions and reserved rights on content.

B. Computing with Legal Knowledge

Many existing metadata vocabularies provide scope to capture the state of legal rights. Dublin Core, for instance, defines fields for capturing rights statements, the date of creation and references to legal document in addition to the person or organization that holds these rights. Examples of vocabularies supporting similar fields are seen below in Table 2. These representations, while capturing minimal semantics, have been put to good use. The Australian National Data Service (ANDS), for instance, has utilized this additional information to allow filtering of data by license – a positive step towards recognition of this issue.

TABLE II. Approaches for Capturing Copyright

Metadata	Dublin Core [41], DCAT [42], DOAP [43], myExperiment [44]
Rights Expression Language	ODRL[45], MPEG-21 [46], METSRights [47], PRISM [48]
Rights Vocabulary	Creative Commons REL [49], L4LOD [50], Copyright Ontology [51]
Legal Defeasible Rules	LegalRuleML[52], CAF [53], LKIF

By contrast to the metadata initiatives, which aim to enable functionality, Rights Expression Languages (REL) aim to encode restrictions on the use of content. More specifically rights expression languages provide formal, machine-readable expressions of copyright usually through creating a controlled vocabulary of verbs standing in as restricted actions. Their use is usually tied with a digital rights management (DRM) system that technically imposes the restrictions. In practice the MPEG-21 and the Open Digital Rights Language (ODRL) are the most prominently used RELs.

[1] http://creativecommons.org/publicdomain/zero/1.0/

[2] http://opendatacommons.org/licenses/pddl/summary/

Collectively RELs have been criticized for lack interoperability [51], misrepresentation of copyright and their potential to harm the preservation of content [36], [54], [55]. Soft rights such as 'fair use' have been difficult to encode due to their ambiguous nature and most existing RELs take the "Everything not permitted is forbidden" approach. Interoperability concerns and the divergence from copyright law led to the development of copyright domain ontologies [51], designed to provide a layer of interoperability; extending notions of copyright law to existing RELs [56]. But this approach only captures the expression and transfer of rights and neglects obligations and conditions and, crucially, exemptions. Additionally, its narrow interpretation of copyright law, built upon the subsumption capabilities of OWL-DL, does not address the diversity and potential misalignment between the terms of various licenses.

There have been many attempts at authoring task-level vocabularies for capturing rights statements [49], [50]. The L4LOD vocabulary, for example, has been used to mediate between many existing vocabularies [50]. Compatibility and composition between representations of licenses have also been explored. Work from Speiser [57] has tackled the self-referential clauses often seen in copyleft licensing terms. Villiate [50] et al. have built on their work with the L4LOD vocabulary to develop logic to compose various licenses [58]. But the limitation with their approach is the narrow interpretation of licensing. We believe using established techniques from the artificial intelligence, and specifically from the legal and computational argument communities [59] is the most promising direction for capturing the social and legal factors within our research tools.

VII. Future Work

The sharing and effective reuse of the data and knowledge we generate, while driven by technological advancements, does not exist in isolation from our identities, our culture and the rules we create to govern our societies. Inevitably it seems, culture (and governance) shifts to match the assumptions and expectations created by new technology, but these changes take time. Within the sciences, knowledge production may be the overarching purpose, but the day-to-day practice is grounded in a social context. For eScience tools and services, failure to design for this very real context will inevitably slow uptake and reduce the realized value. We hope this ongoing research encourages others to consider the range of cultural and social rules as first class citizens in the tools they create.

We now suggest some areas in which this work can be improved and expanded upon.

A. Focus for Further Work

- It may not always be the case that licenses are known or adequately described, improved processes for making better assumptions in this case could be valuable, or perhaps the means of integrating a search for the license owners.
- Infringing copyright cannot be separated from the consequences. Presenting a legal interpretation of the range of potential consequences, ranging from financial, criminal or the reduced ability to commercialize research would be valuable.
- Additional constraints such as institutional copyright policy or personal licensing opinions in formulating potential licensing options must eventually be captured, as they play an important part of decision making.

B. Further tooling

- Development of automatic denotation of licenses when publishing/saving/exporting data would encourage better communication.
- Inclusion of metadata harvesting abilities into the tool to ingest rights statements from common vocabularies. This could enable more interesting integrations, such as browser plugins.
- Investigate closer integration into eScience workflow tools. This may provide a useful facet for the automated selection of suitable data, for example.

C. Governance

- Developing a higher-level set of functions to generate reports on a desired corpus of data.
- Integrating the reasoning engine into agent-based models to simulate emergent behaviour of a research community given certain data sharing policies and aspirations, for better understanding community licensing practices and policy decisions.

D. Broadening the Scope

- Investigating options for encoding privacy, ethics and policy concerns, which impact the sharing and reuse of research data.
- Emphasizing positive outcomes for sharing and reuse to encourage desirable actions, such as achieving recognition and reward.

VIII. Conclusions

This paper describes Camden, a novel eScience tool to aid researchers understanding of the constraints imposed by copyright on the reuse and sharing of research data. By adopting an approach to copyright reasoning based on defeasible arguments we have been able to richly represent and automate the decision-making process for determining copyright legality. Bundling this understanding into a lightweight tool has allowed us to make legal guidance widely available to researchers and research communities. By enabling integration with existing and future research tools, through the use of a Java interface and as a web service, we can enable a more socially situated and relevant research infrastructure. The challenge ahead is evolving Camden through user feedback while continuing to build appreciation of other potential factors, described above in Section VII, which play an important role in making decisions regarding copyright.

References

[1] J. Wilbanks, "I Have Seen the Paradigm Shift, and It Is Us," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, T. Hey and S. Tansley, Eds. Microsoft Research.

[2] R. Hosking and M. Gahegan, "The Effects of Licensing on Open Data: Computing a Measure of Health for Our Scholarly Record," in *link.springer.com*, vol. 8219, no. 28, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 432–439.

[3] T. F. Gordon and D. Walton, "Legal reasoning with argumentation schemes," presented at the the 12th International Conference, New York, New York, USA, 2009, p. 137.

[4] T. Gordon and D. Walton, "Proof Burdens and Standards," in *Argumentation in Artificial Intelligence*, no. 12, G. Simari and I. Rahwan, Eds. Boston, MA: Springer US, 2009, pp. 239–258–258.

[5] T. F. Gordon, H. Prakken, and D. Walton, "The Carneades model of argument and burden of proof," *Argumentation in Artificial Intelligence*, vol. 171, no. 10, pp. 875–896, Jul. 2007.

[6] T. F. Gordon, "Analyzing open source license compatibility issues with Carneades," presented at the the 13th International Conference, New York, New York, USA, 2011, pp. 51–55.

[7] D. Nute, "Defeasible logic," *Handbook of logic in artificial intelligence and logic …*, 1994.

[8] G. Governatori, "On the relationship between Carneades and Defeasible Logic," presented at the the 13th International Conference, New York, New York, USA, 2011, pp. 31–40.

[9] P. M. Dung, R. A. Kowalski, and F. Toni, "Dialectic proof procedures for assumption-based, admissible argumentation," *Argumentation in Artificial Intelligence*, vol. 170, no. 2, pp. 114–159.

[10] *Statute of Anne, London (1710)*. 1710.

[11] L. R. Patterson, "Copyright in historical perspective," 1968.

[12] B. Kaplan, "An Unhurried View of Copyright: Proposals and Prospects," *Columbia Law Review*, vol. 66, no. 5, pp. 831–854, May 1966.

[13] R. Deazley, M. Kretschmer, and L. Bently, "Privilege and Property: Essays on the History of Copyright Law," *eprints.gla.ac.uk*, 2010.

[14] D. Harvey, "Copyright | The IT Countrey Justice," *theitcountreyjustice.wordpress.com*. [Online]. Available: http://theitcountreyjustice.wordpress.com/category/copyright/. [Accessed: 11-Nov-2013].

[15] E. L. Eisenstein, *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press, 2009.

[16] E. L. Eisenstein, *The Printing Revolution in Early Modern Europe*, 2nd ed. Cambridge: Cambridge University Press, 2009.

[17] M. A. Hamilton, "Copyright Duration Extension and the Dark Heart of Copyright," *Cardozo Arts & Ent. LJ*, vol. 14, p. 655, 1996.

[18] I. Hargreaves, *Digital Opportunity*. 2011.

[19] J. Tehranian, *Infringement Nation: Copyright 2.0 and You*. Oxford University Press, 2011.

[20] P. A. David, "The economic logic of open science and the balance between private property rights and the public domain in scientific data and information: a primer," pp. 19–34, 2003.

[21] M. C. Rife, "The fair use doctrine: History, application, and implications for (new media) writing teachers," *Computers and Composition*, vol. 24, no. 2, pp. 154–178, Jan. 2007.

[22] P. B. Hugenholtz and E. J. Dommering, "The future of copyright in a digital environment: proceedings of the Royal Academy Colloquium organized by the Royal Netherlands Academy of Sciences (KNAW) and the Institute for Information Law,(Amsterdam, 6-7 July 1995)," vol. 4, 1996.

[23] J. Wilbanks, "Another reason for opening access to research," *BMJ*, vol. 333, no. 7582, pp. 1306–1308, Dec. 2006.

[24] J. Wilbanks, "Openness as infrastructure," *J Cheminf*, vol. 3, no. 1, p. 36, 2011.

[25] J. Wilbanks, "We need a Web for data," *Learn. Pub.*, vol. 23, no. 4, pp. 333–335, Oct. 2010.

[26] N. Korn, C. Oppenheim, and C. Duncan, "IPR and Licensing issues in Derived Data," *Report submitted to the JISC*, 2007.

[27] G. Hagedorn, D. Mietchen, R. A. Morris, D. Agosti, L. Penev, W. G. Berendsohn, and D. Hobern, "Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information," *ZooKeys*, no. 150, p. 127, 2011.

[28] R. Hilty, "Copyright law and scientific research," in *Copyright Law*, no. 13, Edward Elgar Publishing, 2009.

[29] R. M. Hilty, "Five Lessons about Copyright in the Information Society: Reaction of the Scientific Community to Over-Protection and What Policy Makers Should Learn," *J Copyright Soc'y USA*, 2005.

[30] J. H. Reichman and P. F. Uhlir, "A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment," *Law and Contemporary Problems*, vol. 66, no. 1, pp. 315–462, Jan. 2003.

[31] J. H. Reichman and R. Okediji, "When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale," *Minnesota Law Review*, vol. 96, no. 4, Jan. 2013.

[32] V. Stodden, "Enabling reproducible research: licensing for scientific innovation," *Int'l J. Comm. L. & Pol'y*, vol. 13, p. 1, 2009.

[33] V. C. Stodden, "Open science: policy implications for the evolving phenomenon of user-led scientific innovation," *Journal of Science Communication*, vol. 9, no. 1, p. A05.

[34] J. Wilbanks, "Licence restrictions: A fool's errand," *Nature*, vol. 495, no. 7442, pp. 440–441, Mar. 2013.

[35] J. Wilbanks, "Public domain, copyright licenses and the freedom to integrate science," *JCOM*, vol. 7, p. 2, 2008.

[36] J. T. Wilbanks and T. J. Wilbanks, "Science, Open Communication and Sustainable Development," *Sustainability*, vol. 2, no. 4, pp. 993–1015, Apr. 2010.

[37] R. M. Stallman and J. Gay, *Free Software, Free Society: Selected Essays of Richard M. Stallman*. CreateSpace, 2009.

[38] D. J. H. Mathews, G. D. Graff, K. Saha, and D. E. Winickoff, "Access to Stem Cells and Data: Persons, Property Rights, and Scientific Progress," *Science*, vol. 331, no. 6018, pp. 725–727, Feb. 2011.

[39] A. J. Williams, J. Wilbanks, and S. Ekins, "Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models," *PLoS Comput Biol*, vol. 8, no. 9, p. e1002706, Sep. 2012.

[40] R. A. Reese, "Reflections on the Intellectual Commons: Two Perspectives on Copyright Duration and Reversion," *Stanford Law Review*, vol. 47, no. 4, pp. 707–747, Jan. 1995.

[41] "Dublin Core," *purl.org*. [Online]. Available: http://purl.org/dc/terms/. [Accessed: 29-Apr-2014].

[42] "Data Catalog Vocabulary (DCAT)," *w3.org*. [Online]. Available: http://www.w3.org/TR/vocab-dcat/. [Accessed: 29-Apr-2014].

[43] "Description of a Project (DOAP)," *github.com*. [Online]. Available: https://github.com/edumbill/doap/wiki. [Accessed: 29-Apr-2014].

[44] "myExperiment base Ontology," *rdf.myexperiment.org*. [Online]. Available: http://rdf.myexperiment.org/ontologies/base/. [Accessed: 29-Apr-2014].

[45] "Open Digital Rights Language (ODRL)." [Online]. Available: http://www.w3.org/community/odrl/. [Accessed: 29-Apr-2014].

[46] "MPEG-21," *iso.org*. [Online]. Available: http://www.iso.org/iso/catalogue_detail?csnumber=36095. [Accessed: 29-Apr-2014].

[47] "METSRights," *loc.gov*. [Online]. Available: http://www.loc.gov/standards/rights/METSRights.xsd. [Accessed: 29-Apr-2014].

[48] "PRISM," *idealliance.org*. [Online]. Available: http://www.idealliance.org/specifications/prism-metadata-initiative. [Accessed: 29-Apr-2014].

[49] H. Abelson, B. Adida, M. Linksvayer, and N. Yergler, "ccREL: The Creative Commons Rights Expression Language," 2008.

[50] S. Villata and F. Gandon, "Licenses Compatibility and Composition in the Web of Data.," 2012.

[51] R. García, R. Gil, and J. Delgado, "A web ontologies framework for digital rights management," *Artif Intell Law*, vol. 15, no. 2, pp. 137–154, Feb. 2007.

[52] T. Athan, H. Boley, G. Governatori, M. Palmirani, A. Paschke, and A. Wyner, "OASIS LegalRuleML," presented at the ICAIL '13: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 2013.

[53] T. F. Gordon, "Introducing the Carneades web application," presented at the the Fourteenth International Conference, New York, New York, USA, 2013, p. 243.

[54] C. Barlas, "Digital Rights Expression Languages (DRELs)," *JISC Technology and Standards Watch*, 2006.

[55] K. Coyle, "Rights expression languages," *A Report for the Library of Congress*, 2004.

[56] R. García, R. Gil, I. Gallego, and J. Delgado, "Formalising ODRL semantics using web ontologies," *Proc. 2nd Intl. ODRL Workshop*, 2005.

[57] S. Speiser and R. Studer, "A Self-Policing Policy Language," in *link.springer.com*, vol. 6496, no. 46, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 730–746.

[58] G. Governatori, A. Rotolo, S. Villata, and F. Gandon, "One License to Compose Them All-A Deontic Logic Approach to Data Licensing on the Web of Data," *ISWC-12th International …*, 2013.

[59] T. J. M. Bench-Capon and P. E. Dunne, "Argumentation in artificial intelligence," *Argumentation in Artificial Intelligence*, vol. 171, no. 10, pp. 619–641, Jul. 2007.