

Legal Metadata for Semantic Web Applications: Case Creative Commons

Herkko Hietanen¹, Melanie Dulong de Rosnay²

¹ Helsinki Institute for Information Technology,
Lappeenranta University of Technology Department of Business Administration,
Creative Commons Finland
Tammasaarencatu 3, P.O. Box 9800, FIN-02015 HUT, Finland

² CERSA CNRS University Paris II,
Creative Commons France,
10 rue Thénard, 75005 Paris, France

{herkko.hietanen@hiit.fi, melanie.dulong-de-rosnay@cersa.org}

Abstract. Sharing of documents on the networks and information retrieval rely on metadata. Making creative works searchable is not the only use for metadata. Legal metadata contributes in lowering content distribution transaction costs but the use of legal metadata leaves several legal issues unaddressed. Creative Commons, an open content licensing system with its flexible copyright licenses has demonstrated the potentialities of legal metadata to encourage access and creativity. Creative Commons still faces challenges and it needs complementary applications and policy changes in order to ensure users' legal security.

Keywords: copyright, law, metadata, semantic interoperability, standard, access.

1 Introduction

The legal profession is just starting to understand the potential of metadata as means of expressing rights related to content. European Union 2001 Copyright Directive urges the content industry to acknowledge legal metadata as one of their priorities.

"Technological development will facilitate the distribution of works, notably on networks, and this will entail the need for right-holders to identify better the work or other subject-matter, the author or any other right-holder, and to provide information about the terms and conditions of use of the work or other subject-matter in order to render easier the management of rights attached to them. Right-holders should be encouraged to use markings..." EUCD article 55

The use of legal metadata has remained sparse. The "all rights reserved" paradigm of online distribution has fulfilled all the needs of the publishers. The open source licensing and its close companion open content licensing have changed the scenery and posed questions that have brought lawyers and semantic web researchers together.

This paper describes one of the most prominent open content licensing schemes Creative Commons (CC) and its approach to using metadata to harvest the benefits of semantic web. The first part of the paper describes the CC licensing system and the services it enables. The second part explains some of the restrictions and challenges that legal metadata expression faces. The paper uses normative regulation theory [1] as method. The paper describes the current situation, and suggests how the law and current licensing systems could be changed in order to better fulfill the needs of both licensors and licensees.

1.1 Generating Open Content Licenses

Creative Commons [2] is a nonprofit organization that has together with international community created a set of open content copyright licenses. CC-licenses have become *de facto* standard in open content publishing. All the CC-licenses permit non-commercial distribution of the works as long as the author is attributed. Rights owners can also choose a license that grants additional permissions like a right to make derivative works with or without copyleft terms and a right to perform and reproduce the work also for commercial use. Choosing a license is done with a simple web form. During the process the user can add additional metadata like the name and the contact information of the author. The web site generates a suitable CC-license according to the choices the user has made. It also generates corresponding resource description framework (RDF/XML) [3] tags that users can attach to their works. Creative Commons has also developed simple "human readable" licenses that summarize the legal license. The three tier approach for license presentation and the automated generation process are the heart of Creative Commons licensing system.

1.2 Legal Metadata and Data Mining

Legal metadata is searchable just like any other metadata. Creative Commons-metadata enables users to narrow their searches to relevant content according to their licenses. For example searches can be limited to images that can be used for non-commercial purposes. Yahoo and Google have incorporated CC-search as part of their advanced search pages [4] that indexes documents using the licenses as one of the attributes. This enables users to search the over fifty million works [5] that are currently licensed under CC-licenses. Sites like online digital image repository Flickr [6] have taken CC-licensing as part of their publishing work flow by enabling photographers to automatically attach the licenses at the time of publishing. Flickr users can also browse and search photos according to their licenses. [7]

1.3 The Regulative Power of Computer Code

Creative Commons is the first model to include data mining applications based on semantic web standards allowing rich search on the content also according to its legal reusability status. Social and political values are included in its design and code. [8] An observation of current legal practice shows abuses (e.g. copyright cartels, copyright term extension [9]). The whole open content ideology is promoting practices and tools that encourage sharing and foster creation at the same time.

Instead of digital rights restrictions that are designed for wealthy stakeholders of the dominant contractual practice system, CC-licenses grant users positive permissions to be exercised. The licenses don't pose additional restriction but rather grant access to rights that would otherwise be exclusive property of the rights owner. By granting the right to make copies, distribute and reuse the work in collective works, Creative Commons provides a tool to the authors who aim at granting more generous licensing conditions to the public that the average usages, and encourages everyone to use and build upon others' works. [10]

1.3 Toward Linguistics, Legal and Technical Interoperability

At the crossing between cyberspace law and knowledge engineering, digital rights expression languages [11] (REL) describe copyrighted works' licensing conditions. This legal knowledge embedded in intelligent agents needs to be represented in human-readable applications and interfaces. CC ensures accessibility for non lawyers and non computer scientists through user-friendly application and language¹, including standardized semiotics like icons that represent each licensing option. Icons also help the users to read and understand the metadata that is typically included in html-header in xml-form. Firefox browser's mozCC [12] extension displays the appropriate CC-icons if a web page includes CC-metadata.

International Commons-project [13] provides translation for access in local languages and adaptation maintaining interoperability between over thirty national legal systems. An international equivalency clause² fosters international collaboration to adapt and gather works from different countries. The international Commons has had its equivalence at technical field. The use of RDF W3C standard aims at facilitating technical interoperability with other standards, for example the expression of CC metadata with ODRL Rights Expression Language. [15]

Creative Commons licensing tools offer a bridge toward other semantic web-based applications. CC has hosted the development of open content research tools and interfaces to gather content and organize derivative works. A popular music remix service CCMixer [15] has enabled a community of music makers and re-mixers to document their works linkages between subsequent contributions and contributors.

Collecting societies need metadata of the copyrighted works in order to distribute royalties according to the works effective usage. Unfortunately existing rights

-
- 1 Creative Commons provides human readable texts that can be understood and executed by the public.
 - 2 According to the ShareAlike clause, derivative work shall be distributed only under the same license or its national adaptation developed under International Commons project.

management applications currently used by collecting societies leaves a lot to hope for. For example a member of French *Cour des Comptes*³ considered current society information management system as relic inherited from the Middle Age. [16] Standardized metadata could provide additional income for the societies and their members. CC metadata experience can bring an added value during negotiations toward the compatibility between CC licensing scheme and some collecting societies' statutes. Unfortunately European collecting societies' rules currently prevent their members from using CC licenses. One of the arguments of the collecting societies has been that such licensing scheme would make it more expensive to administrate their catalogue. Improved semantic web reporting systems, user authentication and the growth of e-commerce all speak on behalf of giving more flexibility and autonomy to authors who want to distribute their work using new business models. Currently creators still have to choose whether to use the collective management system or new innovative Internet based systems.

2 The Quest for Trusted Metadata

At first sight Creative Commons metadata seems to lower transaction cost. It makes easier for users to find relevant content and removes the need to negotiate and bargain terms of use in most cases. CC-metadata is legally protected against removal,⁴ but the metadata alone does not solve the copyright liability issues. The second part of the paper describes liability challenges and discusses whether public registries might establish legal security.

2.1 Metadata and Copyright Liability

Legal metadata may be valuable information for data mining but does it has any legal significance and it and does it offer a safeguard against infringement claims? The answer is yes and no. It does limit the infringement claim of the licensor if the licensor is also the rights owner. But the users and distributors can not fully rely on the metadata, because there are no safeguards to stop fraudulent or unsolicited attachment of the metadata to distributed works. Distributors cannot plead ignorance or to metadata that fraudulently grants permission to freely distribute works. The copyright system does not generally recognize good faith (*bona fide*) defense.

Let's take an example from the tangible world where good faith and registries can abolish the liability. Alice who buys a car from Bob can rely that the department of motor vehicles automobile registry is up-to-date when it says that the car is owned by Bob. Later on Charlie claims that the car was stolen and fraudulently registered to Bob.

³ Public accountings control body.

⁴ Legal protection against the removal or alteration of technical information measures removal is stated in WIPO 1996 Copyright Treaty article 12 which was implemented by 2001 European Union Copyright Directive article 7.

Alice can say that she was in good faith and relied on the information in the registry. Alice does not have to give the car back to Charlie.

Because there is no good faith defense in copyright the whole distribution chain is liable for infringements that happen when the work is first released. The liability means that the risk averse intermediaries may want to deal directly with the authors or large rights owners who can ensure that (rights have been cleared and) the content is licensable. Collecting societies have helped the users to limit the licensing risks. The collecting societies can license works that are owned by their members and in Europe in most cases statutory right to license even the works that are owned by non-members⁵. The intermediaries or users have no risk of liability when dealing with collecting societies. This solution has enabled easy licensing for the users and it has meant low administrative overheads to the societies.

2.2 Limiting Liability

In the Internet some of the actors are exempt from liability. Enterprises handling search engines are not liable for browsing and *indexing* the content even though they may make copies of the works. Internet service providers who merely enable *storage* and *access* services for their clients have limited liability. According to the e-commerce directive the service providers are not liable if they comply with the notice and takedown procedure.

Content service providers who actively filter content and use their editorial power to add value to the service are not exempted from the liability. Even the most diligent risk assessment will not release the distributor from the liability, if there is a problem in the distribution chain. Licensees who want to distribute works that carry the Creative Commons license have to carry the risk of an infringement even though the licenses may suggest that the permission is granted. This has led to a situation where services either automatically filter the content from the web or place the burden of selecting the works on the users. Big media houses and content distributors would be the likely targets for infringement lawsuits.

2.3 Strict Liability within Distribution Chain

Several situations could trigger the liability:

- 1) The work is licensed fraudulently by other than the rights owner or a party empowered to do this;
- 2) It turns out later that the rights owner did not have the authority to license it. This is the case when the rights owner has transferred all the rights for the exclusive collecting society supervision;
- 3) The work is modified but it still carries the same metadata than the original work;
- 4) The work is missing a part of the metadata that was originally attached to it;

⁵ Non-members have the right to collect the license royalties from the society but the practice shows this prerogative is not easily fulfilled.

- 5) The metadata has changed since the initial release of the work;
- 6) The metadata only represents the license text that can leave room for interpretation.

Regulation theory has for long concentrated on legal and economic regulation. Lawrence Lessig describes in his book "Code and other laws of cyberspace" how computer code can sometimes complement or replace the legal code. [9] One of the goals of the Creative Commons is that "machine-readable licenses will further reduce barriers to creativity". In the case of legal metadata the code is not law and the metadata has only restricted legal significance. Legal metadata is the first but certainly not the only step to secure legal sharing of protected works. The false sense of security may lead people to act carelessly just by relying to the attached license. Fraudulent metadata does not absolve the users from the infringement liability but it may generate false beliefs that could be spread throughout the distribution chain. Disconnecting the distribution chain and revoking the license metadata would benefit the rights owners and users alike. The problem of unlimited liability⁶ is not only typical to open content but to any content that is protected by copyright.

2.4 Metadata and Interpretation

Legal metadata is merely a link to the legal system and contracts in the real world. Regulating human behavior can never be done in completely water tight manner. Legal language always leaves room for ambiguity. The interpretation is mainly done by the courts and by the academics. The use of metadata is meant to reduce transaction costs and having terms that require further investigation of interpretation is contrary to the goal of using the metadata.

Creative Commons has tried to reduce legal friction but the licenses have suffered from the ambiguity of the license language. Especially the non-commercial licenses have caused problems for the institutional users like public broadcasters. [17] The non-commercial element is not derived from the copyright system or from the long praxis of the content licensing industry. Over 70 percent of the CC-licensed content is offered with terms that grant the license only to non-commercial use but the license text⁷ leaves the definition somewhat open to the judge interpretation. Creative Commons has facilitated a discussion within a community of artists and authors of what the meaning of non-commercial could be. The discussion reached a consensus and Creative Commons helped to write non-commercial guidelines. [18] But even with the consensus, the loose community lacks the legal means to impose those norms

⁶ Creative Commons France licenses include a limited liability clause applicable to the licensor. He guarantees that he secured all the rights involved in the work (copyright, privacy, defamation, tort injury...). Criminal liability is part of French law and order and licensor responsibility would be applicable even if not mentioned in the license.

⁷ "You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation." Section 4b of the CC non-commercial licenses.

The European Union Copyright Directive also refers to the expression "non commercial" to qualify certain activities and apply them specific processing. e.g. art 5. 2. e) "social institutions pursuing non-commercial purposes".

on users. However, fuzziness is not specific to Creative Commons licensing terms, it is a characteristic of the legal language.

2.5 What about DRM Systems?

Right expression languages can be used in a standalone manner as Technical Information Measures, explaining which actions user may perform and which actions are reserved. But Rights Expression languages can also be used in conjunction with Technical Protection Measures. In this case, rights expression will be enforced by a cryptography algorithm and actions which are outside the license grant will be impossible to perform: the computer will not play, copy or even access the work. The doctrine [19] has been largely criticized because technical regulation may override legal prerogatives and prevent the accomplishment of actions which were not regulated by copyright law. Fair use (exceptions to copyright law in civil law countries) can hardly be represented into DRM systems. [20] Some DRM architectures include a trusted third party who can verify, that the licensor truly is who she claims to be and, that the metadata is up-to-date and not tampered with. DRM languages are developed by and for the community whose goal has been to sell movies, music, and other primarily commercial digital materials. [21] These systems have been built to enforce the access and usage rights over the materials and not to foster free distribution. While DRM may help to revoke the licenses, they raise further issues and they don't offer legal relief to the liability risk before the revocation.

2.6 Voluntary Public Registries

Public registries have been used in cases where the control of the object is not possible or it is not a signal of the ownership. Motor vehicle or real estate registries provide up-to-date information about the owners and the legal servitudes that burden the registered assets. The registries enjoy public trust. Relying to the entries in the registries may reduce parties' legal liability in case there is a breach of property rights. Public licenses that last for the duration of the copyright are very close to being servitudes.

Verifying the metadata could be maintained by the collecting societies. The music collecting societies already have large databases of their clients' music metadata. [22] Attaching licensing metadata and opening the database for public queries would not require large additional investments. The problem is the interoperability of such registries. Representing the works metadata and making the queries would need to be standardized. The collecting societies have a need to control the information of licensed works for royalty collection. [23] At the same time the collecting society could carry the liability of handling the registry. The legal status of the collecting societies means that they are able to carry the liability of false metadata. It is hard to see that a commercial entity other than collecting society could provide "copyright insurance" to users.

Copyright system has lacked registries that enjoy strong public trust. Creative community has had to face among uncertain liability issues a growing problem of *orphan works*. [24] The works are orphan because they lack the relevant metadata. There is no way of knowing who owns the rights to them and sometimes even if there are any rights left because there is no public registry to refer to. The uncertainty and looming liability has meant that several works have been left unused.

Instead of centralized databases, that leads to administrative overheads and arbitrary controls, including metadata descriptors to works suits much better to the web decentralized architecture. Having a network of small, trusted and standardized databases that could discuss with each other would be most suitable for the purpose of creating voluntary copyright database. Lessig proposed a copyright registry that would function like distributed domain name registry. [25] Decentralized registry would ensure that the registry system would serve also the non-commercial niche markets.

The registry could also help to clarify terms that are unclear. Rights owner could define what the sections that are left to interpretation mean and license additional rights to the works. The development of a Rights Data Dictionary as a complement to a Rights Expression language may enable rich definitions and adaptation to national legislations.

3 Conclusions

While Creative Commons provides a legal metadata generator useful to distribute, search, organize and build upon digital content, it does not solve all the issues raised by the economy of sharing. Current legal framework was not designed for digital distribution and open content licenses. Some policy changes and semantic web technology are needed in order for licensing schemes like Creative Commons to fully utilize their potential. Most of the problems related to legal metadata are common to whole content industry. Orphan works, strong liability and the lack of interoperable registries of copyrighted works and their license terms are hurting the whole creative community by raising the transaction costs. Voluntary decentralized registries based on common standards would serve both industry and consumers.

References

- 1 See Robert Baldwin, Martin Cave, *Understanding Regulation: Theory, Strategy, and Practice* Oxford University Press, (1999)
- 2 <http://www.creativecommons.org>
- 3 <http://www.w3.org/RDF/>
- 4 <http://search.yahoo.com/web/advanced>
- 5 Statistics according to volatile web site engines sources are available at http://wiki.creativecommons.org/License_statistics and http://www.openbusiness.cc/cc_stat/
- 6 <http://www.flickr.com>
- 7 Flickr has over 11 million CC-licensed works. *See* <http://www.flickr.com/creativecommons/>
- 8 Lawrence Lessig, *Code And Other Laws of Cyberspace*, Basic Books, (2000).

- 9 Eldred v Ashcroft US Supreme Court case 537 U.S. 186
- 10 "Creative Commons aspires to cultivate a commons in which people can feel free to reuse not only ideas, but also words, images, and music without asking permission — because permission has already been granted to everyone." <http://creativecommons.org/about/legal>
- 11 Karen Coyle, Rights Expression Languages, a white paper for the Library of Congress, (2004) http://www.loc.gov/standards/Coylereport_final1single.pdf
- 12 <https://addons.mozilla.org/firefox/363/>
- 13 <http://creativecommons.org/worldwide/overview>
- 14 <http://odrl.net/Profiles/CC/SPEC.html>
- 15 <http://ccmixter.org/>
- 16 Article from French newspaper le Monde, 09-07-2005 based on 2005 Report from Collecting Societies Control Commission (*Rapport de la Commission de contrôle des sociétés de perception et de répartition des droits*, 2005).
- 17 For an insight of the questions raised by the definition on the notion of Non-Commercial, see: Mikael Pawlo, "What is the meaning of Non-Commercial", in Danièle Bourcier, Melanie Dulong de Rosnay (eds.), International Commons at the digital age, Romillat, (2004) <http://fr.creativecommons.org/articles/sweden.htm>
- 18 <http://lists.ibiblio.org/pipermail/cc-licenses/attachments/20060110/02d7a271/NonCommercialGuidelinesclean-0001.pdf>
- 19 Jessica Litman, Digital Copyright, Prometheus Books (2001)
- 20 Deirdre Mulligan, Aaron Burstein, Implementing Copyright Limitations in Rights Expression Languages, 2002 ACM Workshop on Digital Rights Management.
- 21 See e.g. MPEG-21 framework standard ISO/IEC TR 21000-1 "Information technology -- Multimedia framework (MPEG-21) -- Part 1: Vision, Technologies and Strategy" [http://standards.iso.org/ittf/PubliclyAvailableStandards/c040611_ISO_IEC_TR_21000-1_2004\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/c040611_ISO_IEC_TR_21000-1_2004(E).zip)
- 22 See e.g. ISWC and MI3P at <http://www.iswc.org/iswc/en/html/home.html> and <http://www.mi3p-standard.org/>
- 23 In February 2005 the Lower Court number six of Badajoz, a city in Extremadura, Spain, ruled that a bar owner did not have to pay license fees to the main Spanish collecting society for his use of Creative Commons-licensed music. See <http://creativecommons.org/press-releases/entry/5829>
- 24 See Report on Orphan Works, a report of the register of copyrights, January 2006 at <http://www.copyright.gov/orphan/orphan-report-full.pdf>
- 25 Lawrence Lessig's letter to the congress woman Zoe Lofgren March 6, 2006, p. 7. "In this way, a copyright registry could function analogously to the Internet's "domain name system" (DNS). As you know, to maintain a domain name, the owner must pay a fee for each year the domain name is held. That fee is paid to one of many DNS registrars. These registrars feed the necessary information to a central registry. That registry is then publicly available to resolve DNS addresses." at <http://www.lessig.org/blog/archives/20060306-lofgren.pdf>