

Reliable scholarly objects search and interchange framework

Victor Torres¹; Ruben Tous²; Jaime Delgado²

1 Departament de Tecnologies de la Informació i les Comunicacions,
Universitat Pompeu Fabra (UPF)

Carrer Roc Boronat 128, 08018 Barcelona, Spain
victor.torres@upf.edu;

2 Departament d'Arquitectura de Computadors,
Universitat Politècnica de Catalunya (UPC)

Campus Nord. Carrer Jordi Girona 1-3, 08034 Barcelona, Spain
{rtous, jaime.delgado}@ac.upc.edu

Abstract

Authors of scholarly objects might fear that there is a potential risk that the original material they publish in online sites or that they submit for evaluation to scientific journals or conferences is used by others as their own material. In such cases, it would not be easy for the original authors to prove authorship of the original contribution. In similar circumstances, it is very difficult to prove the authorship or origin of some materials that are being distributed amongst social networks, private or institutional websites or any other means through the Internet, namely documents, papers, images, data, etc. Those materials can be easily plagiarised (e.g. partially or totally translated) and redistributed without any control and with no means to prove authorship. In this context, we propose an online framework for the registration, search, interchange and trade of scholarly objects, which helps to overcome the potential drawbacks of online distribution and publishing. This framework acts as an intellectual property repository and sales point, where people is able to register content and determine the way they want to trade it, while providing innovative search capabilities based on the MPEG Query Format standard [1]. Creative Commons (CC) [2] limitations are identified and overcome by means of a licensing approach that combines Rights Expression Languages and the MPEG-21 Media Value Chain Ontology [3].

Keywords: intellectual property rights; scholarly objects; creative commons; multimedia information retrieval.

1. Introduction

In general, it is very difficult to prove the authorship or origin of some materials that are being distributed amongst social networks, private or institutional websites or any other means through the Internet, namely documents, papers, images, data, etc.

Although there are some initiatives focused to detect plagiarism [4] regarding well-known contributions to literature, it is very difficult to prove authorship for other minor or recent works that are not yet consolidated or present in global databases. Those materials can be easily plagiarised and redistributed without any control and even partially or totally translated.

In this paper, we analyse current approaches and initiatives that deal with intellectual property (IP) rights, determining up to which point they can be considered a secure means for protecting IP from the authors' perspective. After this analysis, we describe the desirable features that an ideal system would have. This framework would act as an intellectual property repository and sales point, where people would be able to register content and determine the way they want to trade it, while providing innovative search capabilities based on the MPEG Query Format standard [1]. Creative Commons (CC) [2] limitations will be identified in section 1.3 and overcome by means of a licensing approach that combines the flexibility of rights expression languages and the MPEG-21 Media Value Chain Ontology [3].

2. Intellectual property, services and initiatives

Intellectual property rights is the set of rights that correspond to authors and other entities (artists, producers, broadcasters, companies, etc.) with respect to works and other types of creations and inventions [5].

Copyright rights apply to literary and artistic works (e.g. written compositions, musical works, photographs, paintings, etc.) and they involve economic rights regarding the work reproduction, distribution, public performance, adaptation and translation and moral rights regarding the right to claim authorship and the right of integrity [6].

2.1. Copyright protection

In general, in most countries, any document, work, or creative project is protected by copyright by virtue of its creation from the date it is created. The inclusion of the author's name, date of creation and a copyright statement or

the symbol "©" within or accompanying the work is a valid means for declaring copyright. However, the presence of this statement does not fully protect the author in case of litigation. Other types of qualified proofs such as written or documentary evidence of the date and time of registration are needed to be sure that a work is safely protected.

2.2. Intellectual Property registry offices

Intellectual Property (IP) registry offices, which usually depend on national governments, provide a mechanism for registering and proving content authorship in both the analogue and digital world.

Although the inscription of content in such registries is not compulsory, they are useful to provide qualified proofs stating that copyright exists for a work and it belongs to someone. Some intellectual property registries already offer online registration facilities [7] [8], easing authors the tedious process of the traditional manual and on-site registration. However, those registries lack other functionalities than the mere registration, such as the interaction with other applications that build upon them via APIs, the possibility for authors to determine other licensing schemes than the de facto "all rights reserved", powerful searching facilities and even trading options.

2.3. Creative Commons licensing

Creative Commons (CC) [2] is a non-profit organisation that provides a set of reference licensing models that can be used by authors which hold IP rights to enable people to easily change their copyright terms from the default of "all rights reserved" to "some rights reserved", while being consistent with the rules of copyright.

Following the CC approach, authors can mark their content with some specific licenses that grant some permissions regarding copyright rights to anyone that accesses the content. It is relevant to remark the importance for customers of being capable to prove that they own the appropriate rights for using a specific content.

We could imagine an editor that is used to work with images subject to any of the CC licensing models, which are used for illustrating their online newspaper or blog with, let's say, photographs obtained from the Flickr [9] site. It may happen that in a certain moment in time an image is licensed under the CC Attribution (BY) model [10], which lets others copy, distribute, display, and perform the copyrighted work, and derivative works based upon it, but only if they give credit the way the author requests. Later in time, the

author of the image may decide to change the licensing model to the CC Attribution Non-Commercial (BY-NC) model [10], which lets others copy, distribute, display, and perform the work, and derivative works based upon it, but for non-commercial purposes only. What happens then? Which is the licensing model that applies to the image? According to the CC model, it would depend on the moment the content is accessed. That is, if the editor accessed the content after the change in the licensing model, the Attribution model would apply. Moreover, as stated in the CC BY-NC license [10], this license would be royalty-free and perpetual (for the duration of the applicable copyright). However, licensing in CC does not provide any proof on how or when content is accessed, since there is no link or association between content and its licensing model, so it would be the editor's responsibility to prove that the licensing model being offered in that moment was the appropriate one in case a legal dispute occurs. The main problems that arise from the usage of Creative Commons licenses are, thus:

- 1) The lack of protection for content authors or rights holders regarding the content commercial use. Enabling a commercial usage of content does not mean they resign a part of the income perceived by the party that exploits it. However, the possibilities for authors of perceiving any income are reduced, since CC licenses do not contemplate the possibility for stating such compensations. In general, content consumers do not have the initiative to reach an agreement with content authors. Thus, authors need to start a legal dispute, which is often a long, expensive and non-fruitful process.

- 2) The lack of protection for content authors or rights holders regarding the license duration. CC licenses grant perpetual rights for those that can prove the content was accessed under that specific licensing model. Any change in the licensing model being used for the content will not apply to the users that accessed the content under the previous model.

- 3) The lack of protection for content consumers, since they need to prove which is (or was) the licensing model applicable to the content they use in case of litigation.

Regarding the first problem, CC has defined the CCPlus (CC+) model [11], which enables authors to express where consumers can get rights beyond those granted by the CC license, which is a non-commercial license (e.g. CC BY-NC). The CCPlus license can include a link to an external site or service, which can be a specialized commercial license broker as e.g. gettyimages [12], or even an email address to be contacted by the consumer. The CCPlus approach is still not available at many sites.

Regarding the third problem, some initiatives already tackle it for specific fields. ImageStamper [13] is a free online tool that generates and keeps a

timestamp that includes the image, the license that applies and the date. However, it is only useful for images.

2.4. Safe Creative

SafeCreative [14] is a global, free, open and independent intellectual property registry that allows creators and rights holders to register their works and obtain a valid proof suitable to be used on court hearings.

One of the main differences between SafeCreative and other registries is the possibility to state the rights that apply to works by means of predefined or customized licenses. CC and GNU [15] licenses are included between the templates offered to users, whereas, for other customized models, users need to provide their own specific text. The licensing model being applied can be changed any time by the content author or rights holder.

SafeCreative also provides proofs for content consumers that can be used to certify the licensing model being applied to the content when accessed. In that sense, SafeCreative solves the third issue identified previously, since it provides proofs for content consumers, while keeping track of licensing changes. In order to have reliable proofs, SafeCreative uses officially recognised timestamping services and accepts some X.509 digital certificates issued by trusted issuers. API interfaces are also provided to enable the integration of their services in other web applications.

2.5. Copyright associations

The Writers' Copyright Association (WCA), Webmaster's Copyright Association (WMCA) and Musician's Copyright Association (MCA) [16] [17] [18] are different associations that provide registration services for the type of content they deal with. WCA accepts literary work for film and television, books, poems, artwork, lyrics, teleplays, game shows, storyboards, animations and cartoons, web pages etc. WMCA deals with websites, e.g. zipped entire sites, flash movies, custom java scripts, etc. Finally, MCA accepts music files and documents containing scores.

They all function in the same manner. The user uploads a file, pays a fee and receives a registration number that should be applied to the front page of the author's work. If necessary, a Registry employee may produce registration information or material as evidence if legal or official guild action is initiated. The processing fees, which are common for the three associations are available at their sites. Although they provide a simple interface for authors,

they act as an unofficial intellectual property registry, while providing some basic search functionalities to browse amongst their registry entries.

Some possible problems when relying on these 3 registries are related to its terms and conditions of use, available at their web sites: 1) any of the three associations provides a formal copyright; 2) they do not “verify the originality or authenticity of the material, make comparisons of registration deposits, provide any statutory protections, nor give legal advice”; 3) “In the unlikely event that said file is lost, corrupted, damaged or destroyed due to the WCA's failure to maintain reasonable care or by any other cause whatsoever, it is agreed between both parties that the liquidated damages for the loss of the manuscript shall be £1.00”.

3. Proposed System

In order to deal with the potential drawbacks of online distribution and publishing and respecting intellectual property rights, we propose an online framework for the registration, search and trade of scholarly objects.

A means to prove authorship is the first functionality needed for such framework. Thus, it would act as an intellectual property repository where people would be able to post their content prior to any other action they may want to do with them, such as submitting a paper for evaluation or publishing it elsewhere. Digital signatures applied to the XML [19] representation of works will be a reliable proof for authorship. Two approaches can be followed here. A digital signature from the framework will be trustable as long as we trust in the framework management. A digital signature from the user will require the usage of a recognised X.509 certificate and private key from the user's side and will be a more reliable proof of registration. The combination of both approaches would be the optimum solution. We must say that even adopting these mechanisms, still some legal disputes may arise regarding content ownership. However, the chances of happening so will be clearly reduced and limited to some active thefts.

Another desirable feature is the possibility to trade or share the registered content. In this way, the proposed framework acts as a sales point, easing the distribution and commercialisation of content, and always giving digital evidence of all the transactions being executed in the system, not only for authors but also for customers. The rights to be considered for being traded are, on one hand, those involved in content creation and distribution, which are defined in the MPEG-21 Media Value Chain Ontology (MVCO) [3]: make adaptation, make instance and make copy (useful for determining the type of

works that can be derived), distribute, produce, public communication and synchronization. On the other hand, we should consider those rights related to content consumption and fruition such as render/play, embed, extract, enlarge, diminish, enhance, etc., which are defined in the MPEG-21 Rights Expression Language (REL) [20]. With these two sets of rights we can refer to any action that can be exercised over the content both during creation and distribution and its consumption by final users. Apart from rights, we need to consider conditions, which restrict how rights can be exercised. MPEG-21 REL defines different types of conditions, amongst which we find the following: temporal (e.g. from/to or time interval), payment (e.g. flat or per use fees to be cleared), territorial (e.g. country or region) and the number of times the right can be exercised. Additionally, in order to cover a wider range of agreements, such as those offered in CC and others, some conditions should be added: attribution, exclusivity, non-territory. Finally, a new condition should be considered: the possibility to keep a percentage of the income generated by derived works.

Once authors can determine the rights they want to trade and the applicable conditions, the same framework can act as a sales point where other users (business or end users), to whom we call consumers, could get and clear licenses that grant them some rights or permissions under certain conditions. A license would, thus, formalise the ownership of rights by a consumer and the related conditions. In that context, both authors and consumers would need to be registered so as to be able to identify them and generate the corresponding licenses that act as proofs, since they are expressed as digitally signed XML documents.

Once we have authors and consumers identified in the system, we can also provide an advanced functionality that enables authors to trade content with everyone or just with a limited and selected set of trusted users. We could even decide to offer different conditions for different sets of selected users, depending on our needs or will. In fact, this can be seen as a social network functionality that empowers social relationships. It is important to remark that the author does not need for the permission of target users, since these targeted offers are not public and will be only accessible by target users when accessing that specific content in the framework or otherwise notified by the author, depending on the implementation. The only requirement is that the author is able to identify the target user by means of a nick-like identifier.

As we have seen with other initiatives, this framework can be managed by anyone, as long as security mechanisms are deployed so as to have a trustable system. In that sense, any of the following aspects will help: security audits, use of external timestamping services, use of digital signatures and

recognised X.509 certificates. The business model relying under the framework may be diverse, varying from the payment model to the free model. It is worth noting that the free model may be free for the general public but not free for massive registration through applications that make use of a specific API. Other approaches may include the payment for some value-added services such as the usage of pseudonyms, advanced statistics, registration with more than one author, use of advanced and/or customised licensing models, preview images, etc., for which authors may appreciate a real and useful value so that they may be willing to pay.

In order to deal with the problems identified in section 2.3 for the Creative Commons (CC) approach, the proposed framework separates licensing into two parts. First, authors decide how content is to be traded, by editing the rights they offer and determining the conditions that apply to them. This edition is done through an intuitive and simple interface which hides the complexity of legal texts. Pre-defined templates are also available for common licenses. After selecting rights and conditions, an equivalent legal text is automatically produced. Finally, any rights acquisition is formalized by means of a digitally signed license, expressed in a Rights Expression Language (REL) (e.g. MPEG-21 REL [20]), which links together consumer identity, consumer rights and conditions and the content identification. Thus, a license acts as a proof for both the author and customer. Whenever the license does not state any temporal conditions, it will be forever. In general, a license will apply as long as conditions are fulfilled. In this way, licenses equivalent to the CC models can be generated, but with the possibility of adding new conditions such as those previously mentioned. On the other hand, authors will still be able to modify the rights they offer from a given moment, but without affecting any licenses that might have been acquired prior to the change. Another important feature proposed for the framework is the possibility to search for content. Authors and consumers should be able to express complex conditions to filter the potentially huge amount of documents. Current information retrieval technologies allow extending the traditional search functionalities beyond the traditional keywords-based or metadata-based querying. New approaches allow, for instance, searching for research papers containing potential image copyright infringements (through content based image retrieval techniques). It would be desirable that both the traditional and the advanced search functionalities would be provided through an open query interface for search, providing high expressive power to allow users formulate sophisticated conditions over the scholarly objects' metadata and contents (textual or audiovisual). We envisage that this interface is based on the MPEG Query Format (MPQF) [1]. MPQF is a recent

standard of the MPEG standardization committee (i.e. ISO/IEC JTC1 SC29/WG11), which provides a standardized interface to multimedia document repositories, including but not limited to multimedia databases, documental databases, digital libraries, spatio-temporal databases and geographical information systems.

The MPEG Query Format offers a new and powerful alternative to the traditional scholarly communication model. MPQF provides scholarly repositories with the ability to extend access to their metadata and contents via a standard query interface, in the same way as Z39.50 [21], but making use of the newest XML querying tools (based in XPath 2.0 [22] and XQuery 1.0 [23]) in combination with a set of advanced multimedia information retrieval capabilities defined within MPEG. This would allow, for example, querying for journal papers by specifying constraints over their related XML metadata (which is not restricted to a particular format) in combination with similarity search, relevance feedback, query-by-keywords, query-by-example media (using an example image for retrieving papers with similar ones), etc. MPQF has been designed to unify the way digital material is searched and retrieved. This has important implications in the near future, when scholarly users' information needs will become more complex and will involve searches combining (in the input and the output) documents from different nature (e-prints, still images, audio transcripts, video files, etc.).

4. Results

In this section we present the system we have developed, which tackles some of the problems identified in section 2.3, and which can be used not only for dealing with scholarly objects, but also for musical compositions, audiovisual works and many other types of creations.

The Intellectual Property Operations System – Digital Shadow (IPOS-DS) [24] is a service-oriented architecture that consists of a main web application, accessible through a web browser, which interacts with different web services. It also includes a user desktop application which deals with the rendering of protected content. Figure 1 depicts the overall architecture. Further details can be found at [25] and [26].

IPOS-DS main features include: 1) Content registration and certification. The IPOS-DS system digitally signs an XML representation of any work registered in the system including the identification of the work and author. Content ownership is ensured and content lineage can be traced thanks to the presence of a reference to its ancestor (e.g. adaptation to work) in the

representation. 2) Content licensing according to MVCO and MPEG-21 REL capabilities. Authors decide how they want to trade content and they can modify it any time without affecting previous purchases. Conditions include not only standard temporal limitations, territory restrictions, fees to be cleared and limited number of executions of the right, but also specific IPOS-DS conditions such as keeping percentage of the income generated by derived content, and determining for whom rights will be available to be acquired. Customers formalise the rights' acquisition through personal user-specific licenses. 3) Content access and monitoring. Content is encrypted and can only be accessed by those who have purchased a license. For those users that are entitled to access content, it can be stored in clear so that they can use it without Digital Rights Management (DRM) restrictions. IPOS-DS keeps track of the licenses being purchased and when content is accessed so that the authors can have detailed usage information. 4) IPOS-DS provides search interfaces based on main content metadata fields.

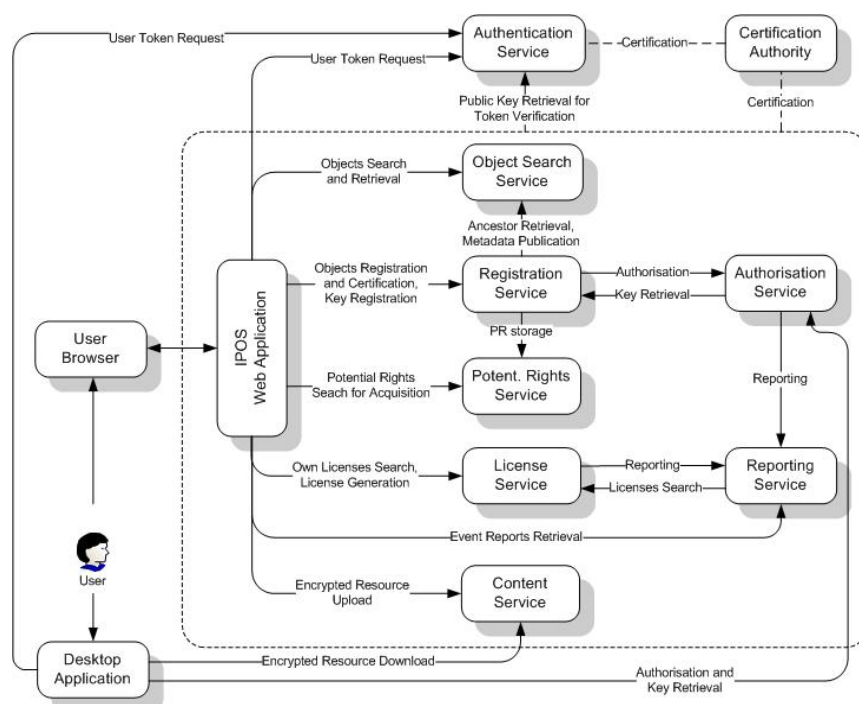


Figure 1: The IPOS-DS System

IPOS-DS is still being improved in some aspects in order to fulfil all the features proposed in section 3: 1) New conditions need to be added to generate Creative Commons-equivalent licenses: exclusivity, attribution. 2) A thorough usability analysis is needed (e.g. use of license templates). 3) Provide better searching capabilities by adopting the MPQF approach. 4)

Interface with official IP registries or recognised timestamping services in order to improve trust.

The IPOS-DS system was commissioned for development to the DMAG (Distributed Multimedia Applications Group) of the UPC (Universitat Politècnica de Catalunya) [27] by the company NetPortedItems S.L. (NPI) [24], which is responsible for its exploitation. It has been made accessible [24] for the public in a pre-exploitation phase.

Regarding the business model, IPOS-DS can be exploited independently by a private company, or even adopted by collecting societies, as it provides much added value by offering their constituents and other users that later may become members the benefit of their collective management services.

5. Conclusions

In this paper we have identified several initiatives that deal with the protection and management of intellectual property rights, which can be applied to scholarly objects.

After describing their main features and analysing their operation, we have identified the drawbacks of current systems and proposed a set of desirable functionalities that an intellectual property registry should have. Our proposal has been made with the aim not only to give protection to authors in terms of copyright but also to give them the freedom to trade their content and provide powerful and innovative searching capabilities in a standardised and automated manner.

Finally, we have also presented the IPOS-DS [24] system, which partially implements the proposed features and which will be extended to fulfil them.

Acknowledgements

Part of this work has been supported by the Spanish administration: Multimedia Content Management Life Cycle project, TEC2008-06692-C02-01.

References

- [1] ISO/IEC 15938-12:2008 "Information Technology -- Multimedia Content Description Interface -- Part 12: Query Format".
- [2] Creative Commons, <http://creativecommons.org/>.
- [3] ISO/IEC FDIS 21000-19. Information technology - Multimedia framework (MPEG-21) - Part 19: Media Value Chain Ontology.

- [4] Turnitin, <http://turnitin.com/>
- [5] World Intellectual Property Organization (WIPO). About Intellectual Property, <http://www.wipo.int/about-ip/en/>
- [6] World Intellectual Property Organization (WIPO). Understanding Copyright and Related Rights, http://www.wipo.int/freepublications/en/intproperty/909/wipo_pub_909.html
- [7] Spanish Intellectual Property Registry, <https://wwwws.mcu.es/RPIntelectual>
- [8] US Intellectual Property Registry, <http://www.copyright.gov/eco/>
- [9] Flickr, <http://www.flickr.com/>
- [10] Creative Commons licenses, <http://creativecommons.org/licenses/>
- [11] CCPlus, <http://wiki.creativecommons.org/CCPlus>
- [12] gettyimages, <http://www.gettyimages.com/>
- [13] ImageStamper, <http://www.imagestamper.com/>
- [14] SafeCreative, <http://www.safecreative.org/>
- [15] GNU licenses, <http://www.gnu.org/licenses/>
- [16] Writers' Copyright Association, <http://www.wcauk.com/>
- [17] Webmasters Copyright Association, <http://www.wmcaglobal.org/>
- [18] Musicians' Copyright Association, <http://www.mcaglobal.org/>
- [19] XML-Signature Syntax and Processing W3C recommendation, <http://www.w3.org/TR/xmlsig-core/>.
- [20] ISO/IEC, Information Technology – Multimedia framework (MPEG-21) – Part 5: Rights Expression Language, ISO/IEC 21000-5:2004.
- [21] ISO 23950. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.
- [22] XML Path Language (XPath) 2.0. W3C Recommendation. 23 January 2007, <http://www.w3.org/TR/xpath20/>.
- [23] XQuery 1.0: An XML Query Language. W3C Recommendation. 23 January 2007, <http://www.w3.org/TR/xquery/>.
- [24] IPOS-DS (Intellectual Property Operations System – Digital Shadow), <http://www.digitalmediavalues.com/>.
- [25] TORRES, V.; DELGADO, J. et al. A web-based rights management system for developing trusted value networks. Proc. of the 18th International World Wide Web Conference Developer's Track, p. 57-59.
- [26] TORRES, V.; DELGADO, J. et al. Enhancing rights management systems through the development of trusted value networks. Proc. of the 7th International Workshop on Security in Information Systems, pp. 26-35.
- [27] Distributed Multimedia Applications Group (DMAG), <http://dmag.ac.upc.edu/>.