

CoMa in Context

Aldo de Moor, Manfred Jeusfeld

Infolab, Dept. of Information Systems and Management

Tilburg University

v1 / December 19, 2003

Many initiatives are being launched in the domain of scientific copyright management, one of them the CoMa project. A core deliverable of CoMa is a knowledge base of copyright policies of the various publishers. Some important questions regarding the knowledge base need to be answered: what should be its requirements, what is the context of use of the proposed system, how does it relate to other systems, and what could be its future role in an (inter)national context of distributed scientific open archives?

To start addressing these questions, we need a conceptual view of the complex problematic we are trying to address. The purpose of this short paper is to give a high level overview of the context of the CoMa project. This overview may help project members in getting a better understanding of how the various components are related, as well as in the preparation of the final report and future project plans. It should be considered a working document that can be extended as the need arises.

We first mention some related projects, followed by the objectives and deliverables envisioned in the CoMa project itself. We then do a quick scan of the open access world out there. Key in open access is license management. Metadata play an important role in its optimization. To effectively use metadata in knowledge base matching, digital rights management is required. After having given an overview of this domain, we will focus specifically on digital rights expression languages. Using these components, we propose to follow the RoMEO mixed approach. We argue that rights disclosure and harvesting, as well as the services to be provided, need to be addressed in a future project. We end the paper with some conclusions and future research directions.

1. Related Projects

Two international projects, RoMEO and SHERPA, and two national projects, ARNO and DARE are closely related to CoMa.

1.1 RoMEO

The RoMEO (Rights Metadata for Open archiving) project aimed to find out how the rights status of open-access research papers might be communicated digitally through rights metadata [1]. A particular focus was to find out how rights metadata might be disclosed and harvested under the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) [2]. This protocol allows metadata about resources to be shared by Data Providers and harvested by Service Providers. The latter build services using these metadata that can be used by end-users such as researchers or librarians.

1.2 SHERPA

SHERPA (Securing a Hybrid Environment for Research Preservation and Access) focuses on the development of open access institutional digital repositories of research output [3, 4]. It is the successor of the RoMEO project. The focus is on institutionalization and scaling up open access experiments. There is much attention for investigating fundamental issues. One of the concrete services it provides that is directly related to CoMa is the database of

publisher copyright policies & self-archiving [5]. The database has can be queried on publisher name and also has an update-function per publisher record. Given the complexity and tentative nature of many of the copyright policy interpretations we do ourselves, it may be worthwhile having a similar interface on the CoMa database.

1.3 ARNO

The ARNO (Academic Research in the Netherlands Online) project aimed to develop and implement university document servers to make available the scientific output of participating institutions [6]. Its goals were (1) to couple document servers to international distributed digital archives and to the Dutch national information infrastructure, (2) to couple the developed infrastructure to the production processes of scientific publishers and form a basis for peer review, and (3) to connect it seamlessly to digital learning environments.

1.4 DARE

DARE, of which CoMa is part, is the successor of the ARNO project [7, 8]. The objective of this project is to make digitally accessible the research results of Dutch universities and research institutes by building distributed archives. These archives are to be connected with each other and with international initiatives using the OAI protocol. DARE will run from 2003-2006. In 2003, the main goals are to (1) implement a basic infrastructure of repositories and (2) to stimulate the supply of scientific research material.

2. CoMa

One of the main challenges in DARE is how to motivate researchers to submit their articles to an institutional repository, such as a university archive. Publishers are moving in the direction of (partially) allowing this. Main barriers, however, preventing individuals from doing so are, first, fear of breaching publishers' copyrights and, second, not willing to get involved in lengthy bureaucratic procedures. The CoMa project focuses on reducing these barriers [8].

Objectives of the CoMa project are the following: (1) to develop a central knowledge base filled by copyright experts that can be used to help staff place publications in the institutional repository; (2) to couple a copyright management module to the institutional repository entry module. The module can match the article data with the knowledge base and either generate a mail to the publisher, point to a generic agreement, or refer the case to a human expert; (3) to develop a Web interface on the knowledge base.

Concrete deliverables of the project are: (1) the specification of the copyright management knowledge base, (2) the development and testing of a pilot system at Tilburg University, and (3) a proposal for national deployment of the knowledge base.

3. Open Access Licenses

Open access is becoming a major paradigm in scientific publishing. Recent high-profile initiatives and declarations, such as the Budapest Open Access Initiative and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities have acted as a catalyst in focusing many smaller projects, give a rationale, visibility and a call-to-arms to the scientific community [9, 10]. However, many initiatives have a drawback, in that they are already very specific in proposing a particular protocol, and not doing complete justice to the wide variety of licensing conditions and applications possible, as described in, for instance [11]. As an example, the Berlin Declaration already specifically grants all users all rights to copy, use, distribute, transmit and display work publicly as well as to make a small number of printed copies for personal use. This is like putting speed limits in the constitution, instead of a specialized traffic law.

To make more flexible and scalable approaches possible, it is therefore better to argue from a set of principles that can be adapted to a particular context of use. A concrete starting point for CoMa development could be the work of the Zwolle group on Copyright Management in Higher Education [12]. They have defined the so-called Zwolle Principles, which aim to balance stakeholder interests in scholarship friendly copyright practices [13]. According to these principles, in copyright management, the primary focus should be on the allocation of specific rights to various stakeholders. These stakeholders include authors, publishers, librarians, universities and the public.

There are two basic approaches in building support systems for (open access) licenses: systems focusing on providing comprehensive license management facilities and those focusing on specifying and using the (machine-readable) rights that are to be used in licensing management. The first approach is represented by the Creative Commons initiative, the second by digital rights expression languages like ODRL. Both approaches complement and overlap each other. A rapid alignment and partial merging of these approaches can therefore be observed in practice.

3.1 License Management: Creative Commons

The OpenContent movement aims to provide licenses to protect open-access content. There are various forms of license regimes, ranging from the generic GNU Public Licence, which says that source code may be used and modified, but that all derivatives must be put back in the public domain, to licenses that regulate very specific aspects, such as attribution rights [2].

The Creative Commons project is one of the most important representatives of the OpenContent movement [2, 14, 15]. Creative Commons allows users to build custom licenses [16]. Its motto is 'some rights reserved'. An electronic questionnaire can be filled out to create a license. The license comes in three formats: the full legal text, an abbreviated human-readable text with clear symbols, and an electronic version containing machine-readable RDF/XML data that can be interpreted by software. A browser can then search on those documents that match certain licensing conditions, e.g. those that can be used as long as attribution is given.

Creative Commons licenses allow display, public performance, reproduction and distribution of a work. They also allow for four optional restrictions: attribution, non-commercial use, no derivative works, and permitting derivative works under a "sharealike" condition (subsequent works will be made available under the same terms as the original).

3.2 Digital Rights Management

Creative Commons provides full license packages, which are quite simple in their options. In general, such generic and often manual license management approaches do not suffice for more advanced applications, such as using rights knowledge bases for enforcing certain work practices or automatically calculating complex copyrights and exceptions. Digital Rights Management (DRM) techniques are better suited here. A DRM system should offer persistent content protection against unauthorized access. It should manage usage rights for different kinds of digital content across different platforms and control access to content delivered [17]. Commercial DRM applications are often focused on *protecting* against copyright violations whereas open access techniques are more aimed at communicating authorized use [2, 15].

However, caution is needed when applying DRM techniques, as they may lead to privacy violations, for instance. To develop satisfactory DRM approaches, several interests must be balanced, including copyright, privacy, commerce, and security [15].

The goal of the CoMa knowledge base is not to control access but to describe the parameters of control, which can then be implemented in (open access) DRM systems that actually do manage content, such as library systems. Still, CoMa must address DRM issues and techniques, since it is in fact a kind of expert system that *reasons about* digital rights management issues.

3.2.1 Digital Rights Expression Languages (DREL)

At the heart of DRM systems are digital rights expression languages (DREs), which allow for expressing rights over content. There are currently two main DREs: XrML and ODRL.

- XrML is the Extensible Rights Markup Language [18]. It has been built into actual products. However, its use is contentious as there are certain patents issues. The RoMEO project decided against using XrML as, first, it can only be used to describe rights that can be implemented by software, and thus not allow for the representation of more complex rights such as related to "non-commercial" or "educational" use. Second, there is no agreed data dictionary. Third, the licensing conditions of XrML were unclear [2].
- ODRL stands for Open Digital Rights Language [19, 20]. It is favoured by open source and educational communities [2]. Reasons are that it has no license and has some defined lists of terms. In ODRL, content can be defined at any appropriately level of granularity that can be uniquely identified. It distinguishes between three main classes of rights: permissions, restrictions, and conditions over works. A restriction (or constraint) is a limit on the extent of the permission being offered (e.g. max four prints). A condition is a prerequisite requirement that must be met before the permission may be performed (e.g. printing only allowed if you pay a fee).

3.3 The RoMEO approach: a mix of Creative Commons and ODRL

Advantages of ODRL are that it provides flexible and extensive solutions due to the comprehensive nature of the language and its own XML schema. One of the main benefits of Creative Commons (CC), however, is that it provides not just rights metadata but a complete rights system with human readable statements for end users, legal licenses for lawyers, and rights metadata for machines. Furthermore, it is rapidly gaining in popularity among major players such as the Open Archives Initiative and the Dublin Core Metadata Initiative. However, problems with CC were, first, that it allows works to be incorporated in collective works that could be reproduced without the author having any control, and, second, that it allows third parties to sell the work. For example the non-commercial sale of a work cannot be prevented under this license [2].

The RoMEO project therefore decided to adopt a mixed solution: the basis for the RoMEO approach would be CC with an new RDF/XML schema. RoMEO has written ODRL versions of CC licenses that would conform to the ODRL XML schema. [2]. A similar approach could be adopted by the CoMa project and its successors to ensure compatibility with related initiatives and acceptance by key actors in the field.

4. Rights Disclosure and Harvesting

Rights are only useful if they can be disclosed and harvested. The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) provides a standard for doing this [21]. Its main

mechanism is to specify a set of http-based requests in the form of a list of key-value pairs that are sent to data providers. Upon receipt of these requests, metadata records are returned by the providers.

All items in an OAI-compliant repository need a Dublin Core metadata record for harvesting. One of the fields, <dc:rights>, is to contain a human-readable statement of rights. Project RoMEO, however, went one step further, and proposed that this field is machine readable. They have now started a joint initiative, called OAI-Rights [22]. The goal of this project is to investigate and develop means of expressing rights about metadata and resources in the OAI framework. In a successor project to CoMa, such rights disclosure and harvesting issues will need to be investigated in greater detail.

5. Services

Services should have components of license and digital rights management, coupled to some form of rights disclosure and harvesting. In a follow-up project, the Tilburg University Library (and its national partners) will have to examine related work in-depth in order to find their specific niche in the service arena. The following questions should then at least be answered:

- What comparable services have been developed?
- What are their rights and license management approaches
- What functionality do they offer that make use of these management approaches?
- Where functionality gaps are not yet addressed by current initiatives?
- What functionality modules should Tilburg develop to address these gaps?
- How can these new modules be aligned and (partially) integrated with other services?

6. Conclusions

In this paper, we have given a brief overview of the context in which CoMa is being developed. Much work has been achieved in other projects that CoMa can build on. Basic standards are emerging to describe license management and digital rights management issues. Our advice would be to adopt these standards where possible.

Although much work has already been done in projects like RoMEO and ARNO, many issues are still unaddressed. License management interfaces and digital rights management applications are still primitive and not tailored to the workflows of their academic users. In the field of rights disclosure and harvesting, and especially the services to be provided, many opportunities for new research and development should be possible. In a successor to CoMa, these issues all need to be addressed.

There are many interesting research issues that could be addressed in future projects. The focus in copyright management is still very much on individual authors. However, new forms of publishing are emerging. For example, dynamic publishing lets the consumers play an active role in the continuous (re)creation of works by allowing them to give feedback and derive new works from ideas expressed online [23]. Furthermore, authoring becomes ever more a group process, in which publications consists of webs of topics written by different experts [23]. Copyright management will require new structures and processes to deal with such issues. Another problem is the classification of metadata. Licenses now take very limited categories of concepts into account, such as work, author, and a few rights. However,

the types, creation, validation, and even rights on metadata themselves create a whole range of further complexities to be addressed in future more sophisticated services [24].

References

1. Project RoMEO homepage:
<http://www.lboro.ac.uk/departments/ls/disresearch/romeo/>
2. Gadd, E., C. Oppenheim, and S. Proberts, *RoMEO Studies 6: Rights Metadata for Open Archiving*. 2003, Loughborough University: Loughborough, UK.
<http://www.lboro.ac.uk/departments/ls/disresearch/romeo/Romeo%20Studies%206.pdf>
3. Consortium of University Research Libraries, *SHERPA: Securing a Hybrid Environment for Research Preservation and Access, version 2*. 2002.
<http://www.sherpa.ac.uk/documents/proposal.pdf>
4. SHERPA homepage: <http://www.sherpa.ac.uk/>
5. SHERPA Publisher copyright policies & self-archiving:
<http://www.sherpa.ac.uk/romeo.php>
6. ARNO Project Site: <http://www.uba.uva.nl/arno>
7. DARE homepage: <http://www.surf.nl/themas/index2.php?oid=18>
8. Tilburg University, *CoMa: Copyright Management, July 6*. 2003.
9. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities:
<http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>
10. Budapest Open Access Initiative: <http://www.soros.org/openaccess/>
11. Mossink, W., *Auteursrechten op Wetenschappelijke Publicaties*. 1999, Open Universiteit Nederland. <http://www.surf.nl/download/iwiboekje2.pdf>
12. Zwolle Group: Copyright Management for Scholarship:
<http://www.surf.nl/copyright/zwollegroup.php>
13. Zwolle Group: Principles - Balancing Stakeholder Interests in Scholarship Friendly Copyright Practices:
<http://www.surf.nl/copyright/keyissues/scholarlycommunication/principles.php>
14. Stix, G., *Some Rights Reserved*. Scientific American, 2003(February 10).
15. Mommers, L., *Het Recht in de Computer: Auteursrecht op Internet*. IT Monitor, 2003(3): p. 6-7.
16. Creative Commons: <http://creativecommons.org>
17. Liu, Q., R. Safavi-Naini, and N.P. Sheppard. *Digital Rights Management for Content Distribution*. in *Australasian Information Security Workshop 2003 (AISW2003)*, Adelaide, Australia. 2003.
18. XRML homepage: <http://www.xrml.org>
19. ODRL homepage: <http://odrl.net>
20. ODRL Initiative, *Open Digital Rights Language: A Rights Expression Language for Digital Asset Management and E-Commerce*. 2003. <http://odrl.net/docs/ODRL-brochure.pdf>
21. The Open Archives Initiative Protocol for Metadata Harvesting:
<http://www.openarchives.org/OAI/openarchivesprotocol.html>
22. Open Archives Initiative and Project RoMEO Initiate OAI-rights:
<http://www.openarchives.org/news/oairightspress030929.html>
23. Henry, G., *On-Line Publishing in the 21st Century*. D-Lib Magazine, 2003(October 2003).
24. Kircz, J. and A. de Waard. *Metadata in Science Publishing*. in *Informatiewetenschap 2003*. 2003. Eindhoven. <http://www.wis.win.tue.nl/infwet03/proceedings/8/>