

Steps towards automatic acquisition and recognition of IPR conditions for parallel publishing.

Preben Hansen, Gunnar Ericsson and Oscar Täckström
[preben, guer, oscar@sics.se]

SICS – Swedish Institute of Computer Science
2009-03-06

This report is part of the PARPUB project (Domain modeling of rights and conditions for parallel publishing of scientific articles) sponsored by the OpenAccess.se.

Content

| | | |
|-----|--|----|
| 1 | Introduction | 2 |
| 1.1 | Limitations | 3 |
| 2 | Data collection and analysis | 3 |
| 2.1 | Procedure | 4 |
| 3 | Results | 5 |
| 3.1 | Models for IPR information | 5 |
| 3.2 | IPR Conditions and information objects | 6 |
| 3.3 | Stored document structure | 8 |
| 3.4 | Additional and extended metadata fields – a proposal | 9 |
| 3.5 | Example of mark-up of IPR conditions (Sage Publishers) | 9 |
| 4 | Conclusion | 10 |
| | Acknowledgement | 11 |
| | Appendix A | 12 |

1. Introduction

Parallel publishing is a rather new term within the area of access to copyrighted content produced by researchers and is sometimes also called post-print and self-archiving¹.

From an author perspective, a reason for investigating parallel publishing² is that it has been claimed that the more accessible (via internet) an article is to anyone (for example from an academic setting), the higher average number of citations it will get. This will of course be of importance when research and universities are going to be evaluated using measures taking citations into account. From a publisher point of view, another reason might be the implication of parallel publishing for designing business models and ensuring proper use of a publisher's copyrighted material.

One of the issues is of focus in this project, is the scenario where a author within a university (or similar environment), is to publish in his article in an academic journal also would like to publish them in their own local environment. This environment could be in a university "Open Access" repository or on a personal homepage. However, this is often not allowed (or restriction), and every publisher has its own set of conditions and rules that need to be considered by the author and the author's organization. Usually these conditions are made available through the contract or copyright agreement.

Closely related to the content of this part of the project, are other resources that might deal with self-archiving or "parallel publishing such as services of Journal Info³ and Romeo/Sherpa⁴.

The database *Journal Info* is developed by Lund University and in this database you search for the name of the journal. If you find the journal in which you have a published paper, you may find information about the copyright policies in general.

The second database is *Romeo/Sherpa*, developed by University of Nottingham. The database contains information on publishers and their policies for parallel publishing – partly or as a whole. The information collected in these databases has been done manually in almost all cases.

The major and overall concern of this project is the issue of automatic recognition and acquisition of IPR conditions or rules for parallel publishing.

In this report, we present the prerequisites for automatic acquisition and recognition of IPR conditions. This report will be linked to the overall project goal of developing a tool for a service of alerting parallel publishing. We investigate how the conditions have been designed and in what form they are represented in among other things.

¹ Self-archiving means that you deposit a digital document in a public and accessible web-based repository. The deposit include both the full-text document in some defined format such as word or .pdf and a metadata document including name, title, date, journal-name etc...).

² Parallel publishing can have different meanings and have different scope. A simple explanation: Out of one single version, create and publish two or more instances of that item.

³ <http://jinfo.lub.lu.se/jinfo?func=home&language=se>

⁴ <http://www.sherpa.ac.uk/romeo/>

The point of departure was then to use a set of publishers not yet registered by the Romeo/Sherpa database (and not previously examined) in order to avoid overlap with the work done in *Romeo/Sherpa*. Furthermore, since the database *Journal Info* in general only deals with information about copyright and not specific about IPR conditions, this would not overlap the content of our project. In order to investigate IPR conditions, it was necessary to examine each single publisher by visiting their online web pages as well as taking contact with them via e-mail requests. The information and data collected were made at two occasions during 2008: April-may and during December.

1.1 Limitations

First, some clarification or limitations:

- We did not have any formal description of the boundaries of what a condition is and what actually constitutes a “condition” as regard to parallel publishing except for the information found on the Romeo/Sherpa project page.
- Furthermore, it should be noted that this area is a moving target and changes will take place as soon as a publisher develop new strategies and business models.

2. Data collection and analysis

In short, this report present

- Procedure of data collection and analysis (2.1)

Furthermore, based on the analysis, we present a set of results (2.2):

- Models and patterns of information related to IPR conditions and rules within the examined set of publishers.
- A set of information object containing conditions
- Stored digital object on IPR conditions.
- Two new fields for the metadata structure in conditions.txt document.
- Sage Publishing - An example of extracting conditions from a copyright agreement will be presented.

We examined 31 different publishers (see list of publishers in Table 3) of different size and contents. The initial goal was to visit a publisher and download the copyright agreement for publishing a journal article.

The assumption was that this single document would contain all the conditions and that a tool then could be trained to extract those conditions. The point of departure was to use a set of publishers not yet registered by the Romeo/Sherpa database and not previously examined.

However, during the project, it was observed that not all the examined publishers had a copyright agreement (or similar) in an online and downloadable form. Furthermore, of those that had their copyright agreements available, it was also observed that not all publishers had IPR conditions for parallel publishing in their copyright agreement, and finally, some of the IPR

conditions was found on other web pages such within sections for authors and author rights.

This situation made us to move into a modified direction in which we needed to make a more detailed examination of what actually was available and recognizable in order to be used for an automatic acquisition of IPR conditions.

2.1 Procedure

We started with a detailed inspection of each publisher and their online resource, e.g. the different publisher's web-pages such as all available copyright documents and author-related pages for each publisher, trying to collect all possible information that could point us to formal conditions for how to behave in a parallel publishing situation. For each publisher, relevant information and data were saved and registered.

Furthermore, every publisher was contacted by e-mail (see App. A). In the e-mail, we asked the publisher for information about the procedures and policies they had regarding self-archiving and to find that information in the publishers web-structure. The answering rate on the email questionnaire was 71%. Finally, we also approached the Romeo/Sherpa initiative and had a discussion about what routines and procedures they used for acquired their data and, in general, about parallel publishing.

All data were then gathered and saved in digital form as desktop folders, one for each publisher. This information was zipped and made as a deliverable.

In a later phase all the available DD:s (Downloadable Documents, cf. Table 2) from the 31 publishers were examined and all passages expressing restrictions or conditions for parallel publishing were marked. Every passage were annotated according to the type of condition or restriction it expressed. This was done in order to identify textual patterns that may serve as seeds for an automatic procedure to harvest conditions from documents of this type. See section 3.5 for an example of the result of the annotation work.

As described, 31 different publishers were approached during the project. These publishers were, at the time of the project start, not yet recorded within the Romeo/Sherpa database and it was decided by the project management to investigate these⁵.

The list of publishers ranged from large to very small publishers, and they contained journals from academic publishers as well as magazines from more popular scientific publishers. The number of journals within each publisher examined also varied. Furthermore some of the publishers actually were imprints of larger publishing houses. The list of publishers examined can be seen in Table 3 (next page).

⁵ During the project, at least 1 of these publishers was recorded within the Romeo/Sherpa database.

Table 1: Size of the examined set of publishers.

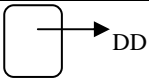
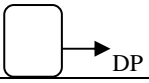
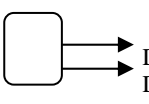
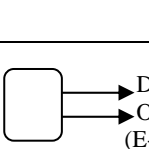
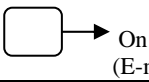
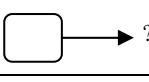
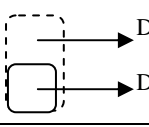
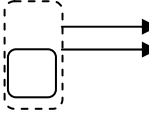
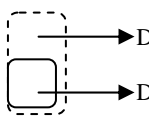
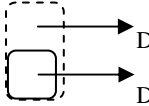
| | |
|-----------------------|----|
| Large | 3 |
| Imprints within large | 10 |
| Small-medium sized | 13 |
| Magazines | 4 |
| Not classified | 1 |
| Total | 31 |

3. Results

3.1 Models for IPR information

Based on a manual examination of all the publishers and the publisher survey the following different detailed models can be identified based on the examined set of publishers.

Table 2: Models of available condition for parallel publishing

| | | Publisher | Size of publisher |
|------|---|---|----------------------------------|
| A |  | Heldref Kinetics | M-S M-S |
| B |  | Economist Business Journal WGL Firenze | Mag M-S M-S |
| C |  | Emerald IEEE IEE Greyter Hogrefe IOP | L L L M-S M-S M-S |
| D |  | Indiana Toronto IOS | M-S M-S M-S |
| E |  | Tieme Internet Scientific Stewart | M-S Mag M-S |
| F |  | Primedia Reed Business VNU | Mag Mag |
| ImpA |  | Hodder Consultant Bureau | Imp Imp |
| ImpB |  | Churchill Swets | Imp Imp |
| ImpC |  | ...Medicales... | Imp |
| ImpD |  | Higher education Press MAIK Mosby Saunders | Imp Imp Imp Imp |

Legend: Imp=imprint; L=large; M-S=medium and small (subjective definition based on examined publishers) DD=Downloadable document (e.g. copyright agreement) and DP = Document Pointer (web-page).

It must be noted that this information may change with short notice. The information on which the table rests on, have been extracted manually between May and December 2008.

3.2 IPR Conditions and Information Objects

In the next step of the examination, we investigated more closely the web pages for each publisher. We searched different levels of their web-structure in order to find documents or pointers to information about IPR conditions. The assumption was that conditions could be found on the copyright agreement form. This showed not to be the case.

Conditions were found both in whole or partially in copyright agreements. Furthermore, other document and pointers to contained conditions.

This means that with an automatic acquisition procedure in mind, this situation is more complex than first anticipated. There were even some cases in which the copyright and conditions for parallel publishing only could be acquired from the publisher (via e-mail) at a case-by-case or on-demand situation.

During this examination of the publishers, we identified single downloadable digital copyright agreements in Word or .pdf formats and also pointers (web-links) to web pages that include rules and conditions for self-archiving or parallel publishing.

Analysis: We identified the following different information object as “container” of condition statements.

DD = Downloadable (digital) Documents.

These objects were usually in the form of a Word-document or a pdf-document to be downloaded, signed and returned to the publisher.

DP = Document Pointer.

This type of document is residing inside the publisher’s own web-pages. These pointers were sometimes hard to find and they had different logical placements and showed some variations. Of course these document can be downloadable. However, the distinction here is made on the fact the DD are made for downloading and it is used for signing and then stored in some way acting as a legal agreement.

E = E-mail

In some cases the author needed to contact the publisher in order to gain access to the copyright issues. There were variations of this alternative. Sometimes some information were found on the web but the author needed to state that they wanted to publish a paper before the publisher actually did launch a copyright agreement form.

The list below shows identified types of information objects found for each publisher.

Table 3: List of publishers and identified objects containing “conditions”.

| Publisher | DD | DP | Email |
|---|-----------|-----------|--------------|
| Arnold_Hodder_Headline (Sage Imprint) | x | | |
| Churchill_Livingstone_Elsevier | x | x | |
| Consultants_Bureau_(Springer Imprint) | | x | |
| Editions Scientifiques Medicales (Elsevier) | | x | |
| Elsevier_Masson | x | x | |
| Emerald | x | x | |
| Georg_Thieme_Verlag | | | x |
| Heldref_Publications | x | | |
| Henry_Stewart_Publications | | | x |
| Higher_Education_Press | | x | |
| Hogrefe & Huber Publishers GmbH | x | x | |
| Human_Kinetics | x | | |
| IEEE Computer Society | x | x | |
| Indiana University Press | | x | x |
| Internet Scientific Publications | | | x |
| IOP Publishing | x | x | |
| Koninklijke Swets Zeitlinger (taylor & Francis) | x | x | |
| MAIK Nauka - Interperiodica | x | | |
| Mosby Elsevier | x | | |
| North Holland Elsevier | X | x | |
| Primedia Inc | - | - | x |
| Reed Business Information | - | - | - |
| Saunders Elsevier | x | | |
| Institution of Engineering and Technology (IEE) | x | x | x |
| Walter de Gruyter Publishing | | x | x |
| University of Toronto Press | | x | x |
| VNU_Business_Publications | | | |
| Economist_Intelligence_Unit | | x | |
| Firenze Univerity Press | - | - | - |
| IOS Press | | x | |
| Warren, Gorham & Lamont | | x | |

In order to automatically find both downloadable document and document pointers for extracting conditions, we found that these were named with a great variation. The following is a list of recognizable names for each of the defined document types DD and DP.

Documents containing copyright or license agreements that may contain IPR conditions were assigned different link-names. From a web-crawler or a link-storage perspective, the naming conventions may be an issue to consider:

Naming conventions used for Downloadable Documents (DD):

/authorrights.pdf
/copyright_form.doc
/copyright_form.pdf
other.doc
/jpa_example.pfd
/author_use.doc
/jarform.pdf
/licence to publish.doc
/journal-policy.pdf
/cpyrigh-rj.pdf
/copy_and_perms.pd

Naming conventions (examples of) used for Document pointer (DP):

/nih⁶
/terms & conditions
/author_chapter
/conditions
/policies
/permissions
/authorguide
/public_repositories
/author faq
/author rights

The identification of different types of documents objects that may be the container of conditions will have implications for the development of tools that will recognize IPR conditions.

3.3 Stored document structure

Based on the existing structure of the knowledge representation format for archiving information (a modification of the structure used in the Leonardo system) about the 31 publishers, we stored both separate documents in full-text and metadata (data.txt) as well as additional information related to parallel publishing.

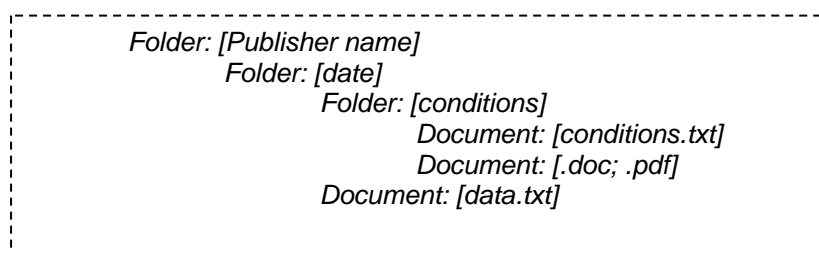
Data collected and used in this report have been saved and stored in a folder that can be reused. However, since the examination showed a more complex situation, we have made some additions to the original document structure within the folders.

Since we downloaded both digital objects containing IPR conditions and identified pointers to digital objects, we suggest the creation of one new folder and one new document and 2 new fields within the new document (or it may be placed elsewhere). We make the following suggestion:

- A folder called “*conditions*”. Within that folder we stored:
 - o Within the folder conditions, a text document called “*conditions.txt*” is created containing the following fields:
 - contract-url and
 - conditions-url.
 - o Within the folder conditions, all downloadable documents found containing IPR conditions are stored.

⁶ NIH Public Access Policy Details (<http://publicaccess.nih.gov/policy.htm>)

Consequently, the file structure looks as follows:



All this information will be saved as a zipped file and made available as a deliverable from the project.

3.4 Additional and extended metadata fields – a proposal

In the original file *data.txt* that resides in each publisher folder there is the following field structure:

```
publisher-url:
contract-url:
contract2-url:
contract-date:
publisher-name:
commentpage-url:
comment:
```

Based on the experience and knowledge gained through the examination of the publishers described above, we may suggest two new additional fields that would be quite useful:

conditions-url:

description: a pointer to a web page (web-address) which contains identified conditions for parallel publishing. It could be additional conditions or conditions for a specific journal.

imprint-url:

description: a pointer to a formal and separate web page for an imprint publisher and its journals that may reside within or outside a larger publisher. Here one may find additional or diverging information regarding the imprint journals copyright status etc.

In the final document structure of the examined publishers, only the first two have been recorded in each single publisher-folder.

3.5 Example of mark-up of IPR conditions (Sage Publishers)

The following is an example from a copyright agreement document with markup of identified conditions for self-archiving or parallel publishing. The

text marked with *italics* marked areas are related to conditions for parallel publishing. The following example is taken from a personal communication with a responsible person and the *Sage Publishers*.

Without further permission, you may:

- At any time, distribute on a not-for-profit basis photocopies of the published article for your own teaching needs or to supply on an individual basis to research colleagues.
- *At any time, circulate or post on any repository or website the version of the article that you submitted to the journal (i.e. the version before peer-review).*
- *At least 12 months after publication, post on any non-commercial* repository or website* the version of your article that was accepted for publication.*
- At least 12 months after publication re-publish the whole or any part of the Contribution in a printed work written, edited or compiled by you provided references is made to first publication by SAGE/SOCIETY.

For any use not detailed above, please contact SAGE at permissions@sagepub.com. Please forward to SAGE all inquiries and requests by third parties for permissions, reprint rights, subsidiary rights licenses, and all other use and licensing of the Contribution.

- *The SAGE-created PDF of the published Contribution may not be posted at any time.*
- In each instance of use of the Contribution, or any part of it, must include the copyright notice that appears on the issue of the Journal in which the Contribution is first published and a full bibliographic citation to the Journal as published by SAGE;
- Copies of the Contribution, or any part of it, shall not be sold, distributed, or reproduced for commercial purposes (i.e., for monetary gain on Contributor's own account or on that of a third party, or for indirect financial gain by a commercial entity);
- The Contribution, or any part of it, shall not be used for any systematic external distribution by a third party (e.g., a listserv or database connected to a public access server).

4. Conclusions

Extracting IPR conditions only from copyright agreements proved to be a more complex task than expected, and the results does not satisfy the initial goals of the this part of the project.

The results presented shows that in order to be able to automatically extract IPR conditions from a publisher's a) downloadable documents such as a copyright agreement or b) from a publishers web-based documents on the publishers online web-pages, the tools(s) developed need to take the observed facts into account: recognize downloadable document naming conventions

(these might differ from publisher to publisher), recognize onsite web-based document naming conventions since they are even more diversified.

An even larger problem in this context is that the information about specific IPR conditions for one publisher can be distributed over several different documents. We suggest at least two new possible metadata fields that can be used for specifying links and pointers to documents.

Finally, we presented a list of different models of structures pointing to document containing IPR conditions. These needs considering when designing and developing tools for automatic acquisition of IPR conditions for parallel publishing.

Acknowledgement

The present memorandum is a progress report from one activity within the project “*Domänmodellering av rättigheter och bivillkor vid parallellpublicering av vetenskapliga artiklar (PARPUB)*” (Domain modeling of rights and conditions for parallel publishing of scientific articles) which is funded by the OpenAccess.se programme at Kungl. Biblioteket (the National Library of Sweden). Their support is gratefully acknowledged. Erik Sandewall is the project leader for this project; he has specified the goals of the present activity and given useful advise during the work and in the preparation of the report.

Appendix A: Letter sent out to all 31 publishers.

Dear Sir/Madam,

I would like to ask you for information regarding the issue of parallel publishing. Where on the web and in what documents can I find information about what my rights are as an author are in the following case:

Let say that I have a paper published in one of your journals, then, what are my rights to publish this paper at

- my own web-page at my university department, or
- that the paper is stored and accessible from a repository at my university or institute

Please, could you point me in the directions were I can read the conditions that might apply for this question.

Regards

Preben Hansen

SICS – Swedish Institute of Computer Science