Building shared language research environments inside European Union – how to optimize the system based on experiences from real life.

Ville Oksanen
Aalto University
School of Science
ville.oksanen@aalto.fi

Krister Linden
University of Helsinki
Language Technology
krister.linden@helsinki.fi

Abstract

Building a shared research environment in Internet is a challenging task. The licenses for services and datasets are typically tied to certain universities or geographical locations – if the licenses are available at all.

In this article we first describe the general environment, in which the language researchers work e.g. what kind of tools and datasets are being used. Special consideration is given to data driven language research because it obviously relies on both. A brief summary of the relevant EU-regulation is also included in this section.

In the second section we document two case examples. The first one is the EU-CLARIN project, whose goal was to establish an EU-wide network of service providers for language research tools and datasets. In that project IPR-licensing was one of the key themes, and considerable resources were used to seek solutions which would have enabled maximum sharing of the content between all participants. The project also prepared some empirical research on licensing practices pertaining to the datasets. The key results of this research are summarized in this section. The chosen three-tiered licensing model (publicly available – academic content – restricted content) and its justifications are also analyzed in detail.

The second case example is the META-SHARE project, whose aim is to build "an open, integrated, secure and interoperable sharing and exchange facility" for language resources. The project also aims to offer the content to any research domain in which language plays a critical role. In META-SHARE the solution is to use tailored "walled garden" versions of the Creative Commons licenses to facilitate the distribution among the members and also to offer special commercial licenses for commercial datasets and tools. We describe in detail how the licensing is supposed to work and what drawbacks and benefits it has. In the end of the section, we compare the projects and their outcomes and show how their approaches are actually complementary to each other.

The last section of this paper starts with a description of how the licensing and general management of IPRs should be done in an ideal world based on the experiences from the aforementioned projects. We then proceed to consider to what extent the ideal model could be implemented in the real world taking into consideration the limited resources and political realities of IPR regulation. We end the section with an analysis of certain concrete proposals e.g. collective licensing for research databases and an EU-level general exception to copyright for research purposes.

Keywords: Copyright, Datasets, Language technology

1. INTRODUCTION

Modern language research, especially data driven language research, needs as large collections of written and spoken data as possible. The collections are used for testing the language models, or for statistically estimating new language models. Therefore one might think that especially current Internet, which produces humongous amounts of data every day, would be a treasure trove for the researchers. Unfortunately, this is not the case in reality.

There is one major reason. The current legislative environment makes it very hard to collect and share large datasets of written and spoken material. Conserning copyright, the relevant InfoSoc Directive[1] provides only a very narrow option for a research-exception:

> (n) use by communication or making available, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of establishments referred to in paragraph 2(c) of works and other subject-matter not subject to purchase or licensing terms which are contained in their collections;

To make the situation worse, the exception is not mandatory for the member states and thus there is indeed considerable variation between the EU-members states if and how the exception has been transposed to the national laws. To make the situation even worse, a recent judgment from the European Court of Justice[2] set the threshold for a copyrighted material extremely low:

> *"…The act of printing out an extract of 11 words, during a data capture process consisting in scanning of newspaper articles followed by conversion into a text file, electronic processing of the reproduction, storage of part of that reproduction and printing out, does not fulfil the condition of being transient in nature as required by Article 5(1) of Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society and, therefore, that process cannot be carried out without the consent of the relevant rightholders."*

As a consequence, if a research project wants to be on the safe side of the law, it has to make sure that it first of all has a permission from all the authors of the collected material to make a copy of their works. In certain categories of works this is already possible if the works include machine readable metadata which describes the required conditions for free usage. For example, European Commission has decided to publish part of its documentation with a Creative Commons license, which makes it possible to collect and store these documents in an automated process without having legal worries.[3]

[1] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society

[2] Infopaq International A/S v. Danske Dagblades Forening, Case C-5/08, http://curia.europa.eu/juris/document/document.jsf?docid=74239&doclang=EN&mode=&part=1

[3] See COMMISSION DECISION of 12 December 2011 on the reuse of Commission documents

However, the legal problems do not end here. The European Union has at the moment a very strict protection for any kind of personal data. The definition of personal information is very wide:

> (a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity; [4]

Similarly the definition of processing is so extensive that it is impossible to avoid:

> (b) 'processing of personal data' ('processing') shall mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction;

This would mean that in addition to asking permission from all the possible copyright holders of the data, the research project should have secured similar permission from anyone who might be possible to identify in the data. Luckily there is a safety valve in the Directive:

> *2. Subject to adequate legal safeguards, in particular that the data are not used for taking measures or decisions regarding any particular individual, Member States may, where there is clearly no risk of breaching the privacy of the data subject, restrict* by a legislative measure *the rights provided for in Article 12 when data are processed solely for purposes of scientific research or are kept in personal form for a period which does not exceed the period necessary for the sole purpose of creating statistics.*

Unfortunately, the aforementioned requirement is only optional ("may", not "shall") for the member states and therefore there is currently no harmonization how the research exception has been implemented in the member states. The easiest solution would be therefore to use one of the countries with a strong research exception as the legal home as the hosting organization. Unfortunately this approach is available only in those limited cases, in which the choice can be made.

1.1. Methodology

This article is mostly descriptive. Our aim is to document the experiences we have derived from our own efforts to create a European wide distribution system for

(2011/833/EU) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF

[4] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data

language research datasets and tools. The last concluding section also includes a “De Lege Ferenda”-section

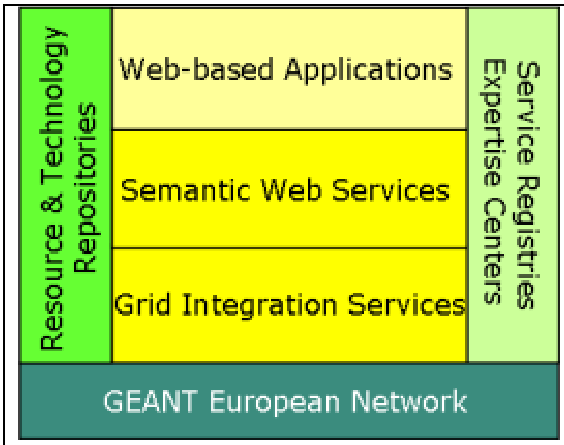
2. CASES - EU-CLARIN AND META-SHARE

In this second section we document two case examples. The first one is the EU-CLARIN project[5], whose goal was to establish an EU-wide network of service providers for language research tools and datasets. In that project IPR-licensing was one of the key themes and considerable resources were used to seek solutions, which would have enabled maximum sharing of the content among all participants.

The project also prepared some empirical research on licensing practices pertaining to the datasets. The key results of this research are summarized in this section. The chosen three-tiered licensing model (publicly available – academic content – restricted content) and its justifications are also analysed in detail.

The second case example is the META-SHARE-project[6], whose aim is to build “an open, integrated, secure and interoperable sharing and exchange facility” for language resources. The project also aims to offer the content to any research domain in which language plays a critical role. In META-SHARE the key part of the solution is to use tailored “walled garden” versions of Creative Commons licenses to facilitate the distribution among the members and also to offer special commercial licenses for commercial datasets and tools. We describe in detail how the licensing is intended to work and what drawbacks and benefits it has. Towards the end of this section, we compare the projects and their outcomes and show how their approaches are complementary to each other.

2.1. EU-CLARIN


Resource & Technology Repositories
Web-based Applications
Semantic Web Services
Grid Integration Services
Service Registries Expertise Centers
GEANT European Network

Picture 1. The Architechture of CLARIN, http://www.clarin.eu/external/img/architecture.png

[5] More info: http://www.clarin.eu/external/index.php?page=about-clarin&sub=1
[6] More info: http://www.meta-net.eu/meta-share/

One of the main points of CLARIN was to make existing datasets available to as large a part of the users as possible while taking into consideration the nature of the content and the ways to identify the users and their status as academic researchers. The distribution of the content takes place in national repositories (Picture 1.). Building on these boundaries, it was evident from early on that not all material could be made available with e.g. Creative Commons licensing and that the system should have good support for legacy licenses.

Unfortunately even the first necessary step of actually finding out what kind of rights there is for the content is often difficult because there is a large group of material, for which there are no written license agreements. Typically, to make things worse, the individuals familiar with the details are no longer easily reachable. Another problem from the CLARIN perspective is the variation in the existing license agreements, which makes it hard to offer a centralized service. To tackle these problems, we created a three-tiered classification system and a pack of (draft) licenses to use for clearing the ambiguities regarding the rights.[7]

Regarding the license variation, we first carried out an extensive survey and found out that it was possible to categorize the licenses into three different groups based on the potential users of the content in question:

- Publicly Available
- Academic Use
- Restricted Use

Picture 2. Symbols for the main distribution classes.

We then followed the model used by Creative Commons and created simple icons ("laundry symbols") that make it easier for the end user to instantly see how the resource could be used, (see Picture 2). In addition, a deed describes the rights in "human readable" textual form and the definition for machine readable legal metadata is part of the larger CLARIN metadata schema.

However, even if the basic structure of Creative Commons is sufficient as such for CLARIN, it is not fully applicable because Creative Commons does not support our aforementioned groupings i.e. for distribution restricted to academia or even more limited groups of users, which is essential for many of the older resources to be included in CLARIN.

The detailed definitions of the different classes are:

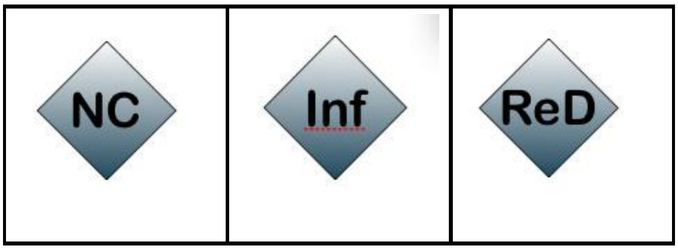
[7] This section is based on earlier article by Oksanen, Lindén and Westerlund (2010)

Publicly Available (PUB) is the category primarily endorsed by CLARIN. To belong to this group, the following requirements have to be met:
- The license should allow distribution of the tools and resources from the CLARIN infrastructure
- There are no limitations (based on status or geographical location etc.) on who can access and use the tools and resources
- There are no limitations on the purpose for which the tools and the resources are used.

In other words, the license should follow the Protocol for Implementing Open Access Data[8] as closely as possible. For the new tools and resources, the preferable license is either the Creative Commons Zero (CC0)[9] or the Open Database License (ODbL). However, for the previously licensed tools and resources, re-licensing is often not possible, and the submitting party has to make a careful assessment of the terms of the existing licensing agreement before labelling the material as PUB.

For **Academic Use** (ACA) the license agreement includes an additional requirement that the use is somehow related to an academic institution. Here the problem is typically the exact definition of academic use. To qualify under this category, the tools and resources:
- Should at least be available for anyone doing research or studying in an academic institution recognized by the user identity management system used by CLARIN;
- Can be used for studying, research and teaching purposes.

The last category, **Restricted Use** (RES) includes the resources that do not fulfil the previous requirements but still could be offered to the users if certain additional requirements are met. The most typical reasons for a resource to fall under the scope of RES are things, which require manual steps before the system can grant the access the content:
- A requirement to submit detailed information (e.g. abstract) about the planned usage;
- Specific ethical or data protection-related additional requirements.



Picture 3. Symbols for additional distribution restriction.

In conjunction with the main license categories, there can also be all or any of three additional requirements. See Picture 3 for the symbols:
- A requirement for strictly non-commercial use (NC)

[8] http://sciencecommons.org/projects/publishing/open-access-data-protocol/
[9] http://creativecommons.org/choose/zero

- A requirement to inform the copyright holder regarding the usage of the tools and/or the resources in published articles (INF)
- Redeposit – modified versions of the tools and resources have to be licensed back (ReD)

All these additional requirements may be combined with the main license categories PUB, ACA and RES.

However, this does not solve all the problems. As described earlier, it is common that there is no license agreement at all, because such an agreement has never been made, or the license is stored in a mailbox of a person no longer working for the organization. Even if there is a license its relatively often somehow problematic (e.g. very low in details), making the categorization uncertain.

To help solving those situations we created the "CLARIN Update Model Agreements" with the purpose to procure the required rights. Here the best (and suggested) option is to re-license the content with the CC0-license.[10] That license is at least well-understood and offers enough rights for all parties in different digital (and non-digital) environments. It is also compatible with most of the other open content licenses.

Unfortunately it is not always possible to use CC0 due to the demands of the copyright holders of the content and Model Agreements for Academic and Non-Commercial Use[11] are available.

In order to test the usability of the classification system and our classification guidelines, we did an initial classification test. We sent out a request to organizations in charge of 116 resources found in the CLARIN LRT inventory. The resources were distributed in Finland (91), Denmark (3), Germany (21) and Greece (1).[12] We received an answer for 40 of the resources, i.e. 34.5%. A response rate above 1/3 should be considered relatively representative.

Distribution type	Number	Percentage
PUB	7	17.5 %
ACA	5	12.5 %
RES	22	55.0 %
No classification applicable	6	15.0 %
Total	40	100.0 %

Table 1: Distribution of resources according to the CLARIN classification.

One of the publicly available resources (PUB) was additionally classified as non-commercial, whereas all of the academically available resources (ACA) were non-

[10] See Berlin Declaration 2003 for more info on the best practices.

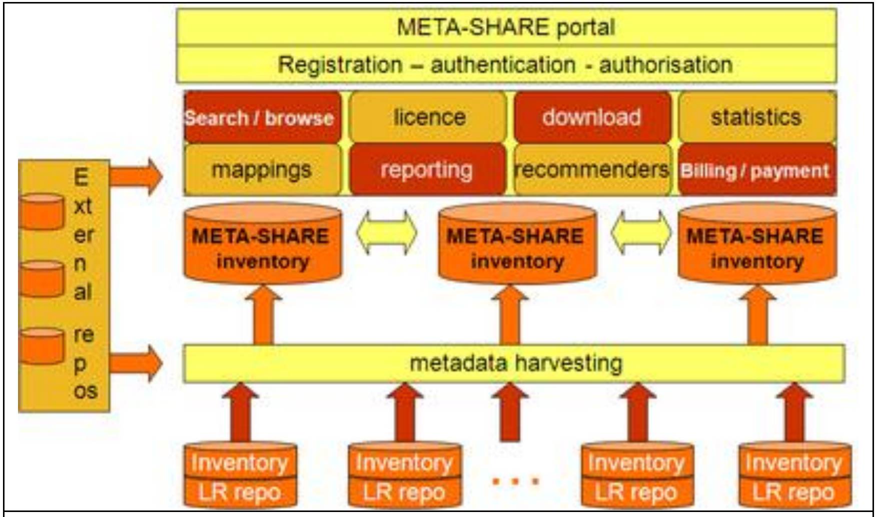
[11] It should be pointed out that the use of NC is strongly discouraged – see Hietanen, Oksanen and Välimäki 2007 for a discussion about the problems related to the term Non-Commercial.

[12] It should be pointed out that Finland is indeed overly represented in the answers. Apparently people answer more readily to the persons they know well already.

commercial. The restricted resources (RES) were roughly equally divided among no additional restrictions (27.3 %), non-commercial (31.8 %) and license from content owner required (40.9 %). The most interesting outcome of the survey was, that academic use is very commonly tied to non-commercial use.

2.2. META-SHARE

The main difference between CLARIN and META-SHARE arise from the architecture of the distribution systems. While CLARIN extensively uses existing national service providers, META-SHARE aims to be based on an open peer-to-peer architecture, in which the content may freely be distributed inside the network.



Picture 4. http://www.meta-net.eu/meta-share/architecture

This difference is naturally reflected in the chosen license strategy. Both of the projects would primarily use standard open content licenses if possible, but due to the realities of the licensing, a second (and third) more restrictive tier is needed. In META-SHARE, the licensing structure is as follows:

- *Creative Commons licences (starting with Creative Commons Zero (CC-0) and all possible combinations along the CC differentiation of rights of use) are the first level of legal machinery applied.*
- *A second layer includes META-SHARE Commons Licences, a fully developed CC-based licensing tool that allows META-SHARE members and Extraneous Depositors to make their resources available to other network members only.*
- *The third legal layer is a set of licenses that allow use and exploitation of the Resources while permitting the LR Owner to have full control over the Resource distribution. These "No Redistribution" licences will effectively help get "closed" resources safely out to the community.*
- *A set of legal document templates (non licences) is offered that is designed to help all stakeholders (resource owners, distributors and end-users) work in a friendly and transparent environment. These include a Depositor's Agreement (DA), a Memorandum of Understanding (MoU) for the Network members and a Service*

> *Level Agreement (SLA). The DA is currently available, while the MoU and the accompanying SLA are under revision.*[13]

Like CLARIN, also META-SHARE offers licenses that can be used to acquire the necessary rights from the right holders. The main distribution licenses are based on CC-licenses but they have additional requirement that the content may be distributed only among the members of META-SHARE.[14]

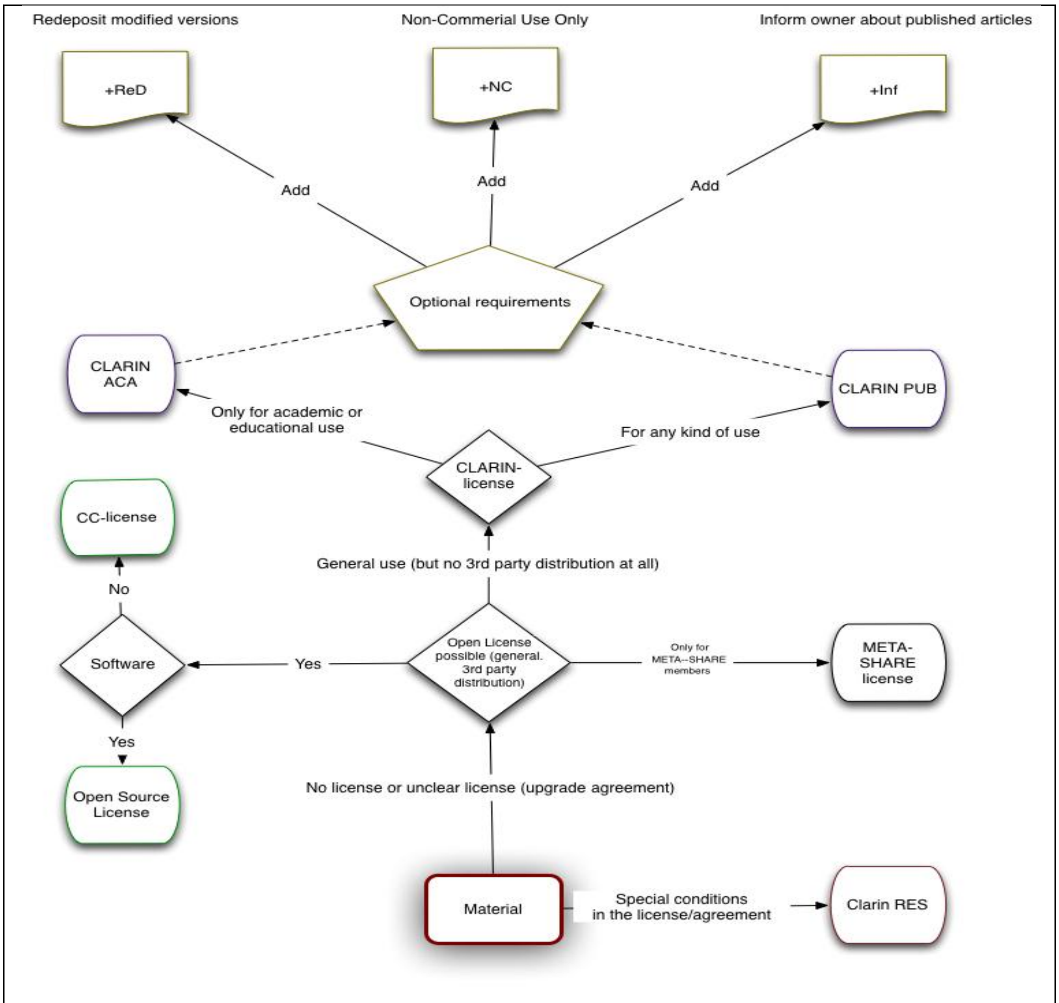
The good side of using CC-licences as a starting point is that they are well-understood and people are already familiar with the terminology. However, the problem is that current CC-licenses do not offer the very features often asked for: specifically research-use licensed content. In addition, the META-SHARE-CC licenses are not compatible with many of the real CC-licenses, which may come as a surprise for the users.

2.3. Choosing the license

The CLARIN network and the META-SHARE project are both trying to accomplish the same task, i.e. offer as much language resources as possible to the users. The natural question is, which one should be used – or is it possible to use both. To answer this question, picture 5 may offer an answer.

[13] http://www.meta-net.eu/meta-share/licenses

[14] The actual licenses has also a bit more streamlined content i.e. clauses, which are not relevant for the purpose, are removed.



Redeposit modified versions
Non-Commerial Use Only
Inform owner about published articles
+ReD
+NC
+Inf
Add
Add
Add
Optional requirements
CLARIN ACA
CLARIN PUB
Only for academic or educational use
For any kind of use
CLARIN-license
CC-license
General use (but no 3rd party distribution at all)
No
Software
Yes
Open License possible (general. 3rd party distribution)
Only for META--SHARE members
META-SHARE license
Yes
Open Source License
No license or unclear license (upgrade agreement)
Material
Special conditions in the license/agreement
Clarin RES

Picture 5. Mapping the License Choices.

The best solution is to use CC0 or a similar license. However, this is often not possible because the distribution of the content is only possible to a limited subgroup. If the subgroup is researchers, the best solution is most likely to use the CLARIN ACA-license. However, if the content is only available for language research, the natural option is to use META-SHARE’s CC-license, which allows the distribution among the network of language researchers. In many cases the only real option is to go for CLARIN-RES, which is aimed to offer the infrastructure for making very detailed decisions to whom the content may be given.

Sometimes it is also worth considering a dual-licensing. For example, CLARIN ACA and META-SHARE licenses are not mutually exclusive, i.e. a resource can be licenses and released in both systems. Of course, if the material can be released under e.g. a real CC-license, it can be used together in CLARIN and META-SHARE.

3. CONCLUSIONS

Even with the re-licensing and limited distribution offered by both systems, a very large set of possible language data remains unreachable. However, this problem does not seem to affect Google and other similar American actors.[15] The reason is that the legal system in the U.S. is much more lenient regarding the copyright exceptions. The concept of *fair use* offers a flexible way to give permission for actions that do not harm the content owners.[16] Unfortunately a similar approach is not currently possible in European Union.

The language scientists have been trying to send the message to Brussels that something has to be done (Oksanen 2009):

> *"... we urge the legislatures at the European level and at the national level to harmonize copyright law in such a way that the free use of copyrighted works for academic purposes, which does not unreasonably prejudice the legitimate interests of the right-holders, is permitted in all member states."*

The formulation was drafted quite carefully. It asks effectively for a very wide exception but limits the demand by adding language from Berne three step test.[17] It is an intentionally vague diplomatic formulation, which can mean practically anything – “The devil is in the details”. The reason for this carefulness is that much of the most valuable content is such that the right holder can prevent the access to it if the rules of release seem too risky, e.g. digital archives of the newspapers.

In practice, the best outcome would be a separate directive, which would add a balanced list of rights for academic users. Unfortunately nothing like this exists currently on the roadmap of the Commission (European Commission 2011). If such a directive is anyway proposed sometime in the future, it would make sense to sort out simultaneously the problem with the protection of the personal data.

4. REFERENCES

Aufderheide, P., Jaszi, P. 2011. Reclaiming fair use: How to put balance back in copyright. Chicago: The University Of Chicago Press Book

Berlin Declaration. 2003. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Berlin: http://www.zim.mpg.de/openaccess-berlin/ berlindeclaration.html.

European Commission. 2011. A Single Market for Intellectual Property Rights - Boosting creativity and innovation to provide economic growth, high quality jobs. Brussels, 24.5.2011 COM(2011) 287 final, http://ec.europa.eu/internal_market/copyright/docs/ipr_strategy/COM_2011_287_en.pdf

Hietanen, H., Oksanen,V., and Välimäki M. 2007. Community Created Content - Law, Business and Policy. Turre Publishing.

[15] It is highly unlikely that this service would ever be made available by any European organization: http://books.google.com/ngrams/

[16] For a recent study on how the scope of *fair use* has been extended lately, see Aufderheide and Jaszi 2011

[17] For a detailed explanation of what the Berne three step test is, see e.g Koelman 2006.

Koelman, Kamiel J. 2006. Fixing the Three-Step Test. European Intellectual Property Review, 2006. Available at SSRN: http://ssrn.com/abstract=924174

Oksanen, V. 2009. NEERI Message - Freedom of use of copyrighted works for academic purposes http://www.csc.fi/english/pages/neeri09/programme/ materials-fri/oksanen2.pdf

Oksanen,V., Linden, K., and Westerlund, H. 2010. Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN, in Proceedings of LREC 2010 : Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management