

# Laundry Symbols and License Management - Practical Considerations for the Distribution of LRs based on experiences from CLARIN

Ville Oksanen\*, Krister Lindén†, Hanna Westerlund†

\*Aalto University

P.O. Box 19210, 00760 Aalto, Finland

†University of Helsinki

P.O.Box 24, 00014 University of Helsinki

E-mail: ville.oksanen@tkk.fi, krister.linden@helsinki.fi, hanna.westerlund@helsinki.fi

## Abstract

One of the most challenging tasks in building language resources is the copyright license management. There are several reasons for this. First of all, the current European copyright system is designed to a large extent to satisfy the commercial actors, e.g. publishers, record companies etc. This means that the scope and duration of the rights are very extensive and there are even certain forms of protection that do not exist elsewhere in the world, e.g. database right. On the other hand, the exceptions for research and teaching are typically very narrow.

## 1. Introduction

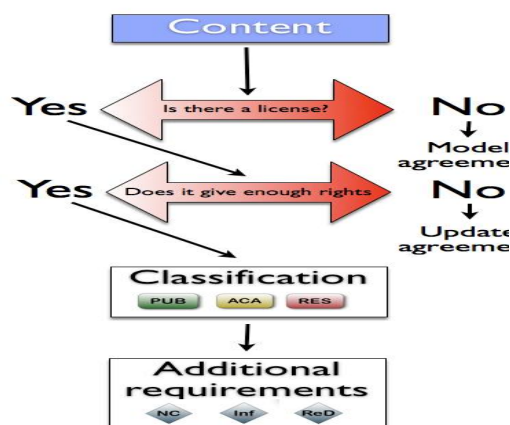
One of the most challenging tasks in building language resources is the copyright license management. For a more general discussion on open access data licensing, see e.g. Klump & al. (2006). There are several reasons for this. First of all, the current European copyright system is designed to a large extent to satisfy the commercial actors, e.g. publishers, record companies etc. This means that the scope and duration of the rights are very extensive and there are even certain forms of protection that do not exist elsewhere in the world, e.g. database right. On the other hand, the exceptions for research and teaching are typically very narrow. To make the situation worse, the possible sanctions for copyright violations are severe, e.g. in Finland the maximum penalty for copyright violation on the Internet is a two-year prison sentence.

This means that there is very little space for errors in the management of copyright licenses - at least in theory. In practice the system mostly “just works” even if the formal agreements are often totally missing or the distribution of material was agreed on in a phone call years ago between persons who no longer work in their respective organizations. The reason for this is that there is typically no commercial interest to start a legal process and high legal fees form an effective preventive factor. However, when building an EU-wide system, one cannot rely on such an informal approach.

In the first part of this article, we describe how we plan to handle the matter in the CLARIN project. In the second part, we describe our early practical experience with the proposed classification. In the last part of the article, we briefly discuss aspects that could be generalized and possible actions for making the use of copyrighted material in research more flexible.

## 2. CLARIN Resource Distribution Types

In CLARIN the typical flow of the content is the following: A copyright holder, e.g. a newspaper, licenses its content to a CLARIN Content Provider that distributes the content to the End Users through a CLARIN Service Provider. This means that the license chain has to follow a similar structure. Unfortunately even the first step is often difficult because there is a group of resources, for which there are no written license agreements and individuals familiar with the details are no longer available. 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, several sets of agreements have to be used. For an outline of the resource classification procedure, see Picture 1.

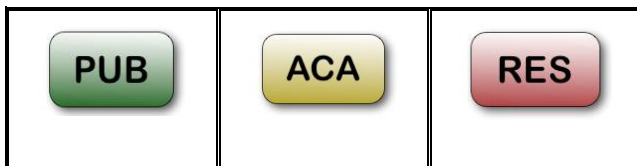


Picture 1. Resource classification task.

Regarding the license variation, we carried out an extensive survey and found that it is possible to categorize the licenses into three different groups:

- Publicly Available Resources
- Resources for Academic Use
- Resources for Restricted Use

We followed the model used by Creative Commons and created simple icons, i.e. *care symbols*, making it easier for the end-user to immediately see under which conditions the resource can be used, see Picture 2. In addition, a deed describes the rights in human readable textual form. Finally, there is also the actual license agreement and the metadata, i.e. the machine readable information. However, Creative Commons is not sufficient as such for CLARIN, because Creative Commons does not allow 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.



Picture 2. Symbols for the main distribution classes.

**Publicly Available (PUB)** is one of the categories 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 must be no limitations, e.g. based on status or geographical location etc., on who can access and use the tools and resources and
- there must be 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<sup>1</sup> as closely as possible. For the new tools and resources, the preferable license is either the Creative Commons Zero (CC0)<sup>2</sup> or the Open Database License (ODbL). However, for the previously licensed tools and resources, re-licensing is often not possible, and the submitting party should make a careful assessment of the terms of the existing licensing agreement.

For **Academic Use (ACA)** the license agreement includes an additional requirement that the use is somehow related to an academic institution. Here the problem may arise from the definition of academic use. To qualify under this category, the tools and resources:

- should be available at least for anyone doing research or studying in an academic institution recognized by the Identity Provider Federation and

- should be available for studying, research and teaching purposes.

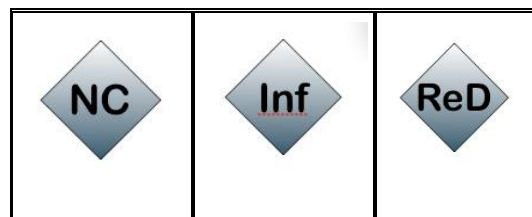
The last category, **Restricted Use (RES)** includes the resources that do not fulfill the previous requirements but still could be offered to the users if certain additional requirements are met. The most typical reasons for a resource to fall under the scope of RES are:

- a requirement to submit detailed information, e.g. an abstract, on the planned usage or
- specific ethical or data protection-related additional requirements.

In conjunction with the main license categories PUB, ACA and RES, there can also be all or any of three additional requirements:

- A requirement for strictly non-commercial use (NC)
- A requirement to inform the copyright holder regarding the usage of the tools and/or the resources in published articles (INF)
- A requirement to redeposit modified versions of the tools and resources with the Service Provider (ReD)

Picture 3 displays the symbols designed for the additional requirements.



Picture 3. Symbols for additional distribution restrictions.

However, this does not solve all the problems. In some cases there either is no license agreement at all, because such an agreement has never been made. It is also quite common that the existing agreement is somehow problematic, e.g. very low in details, making the categorization impossible. For those situations we created the *CLARIN Update Model Agreements* with the purpose to procure the required rights. The best option is to re-license the content with the CC0-license. See the Berlin Declaration (2003) for best scientific licensing practices. It is 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. Thus Update Model Agreements for Academic and Non-Commercial Use are also available.

These agreements presuppose that there are existing agreements but that the rights are not adequate or too unclear. It should be pointed out that both the terms

<sup>1</sup><http://sciencecommons.org/projects/publishing/open-access-data-protocol/>

<sup>2</sup><http://creativecommons.org/choose/zero>

non-commercial and academic are relatively ambiguous and it is a relatively demanding task to write generally accepted definitions. See Hietanen & al. (2007) for a discussion on the problems related to the term Non-Commercial in Creative Commons. Especially the scope of accepted commercial use is something that needs first to be solved on a political level and only after that formulated in legal terms.

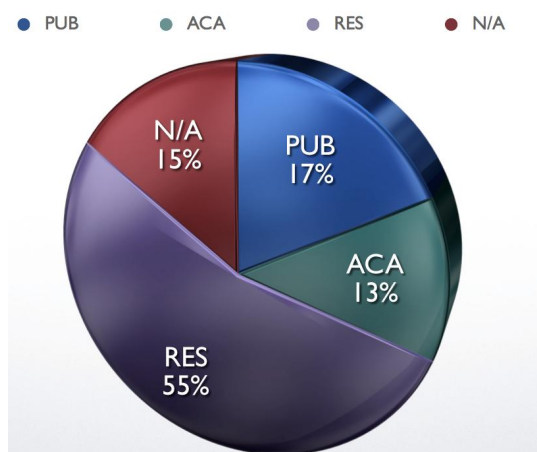
### 3. Practical Experience

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 the custodians of 116 resources found in the CLARIN LRT inventory. The resources and their custodians were located in Finland (91), Denmark (3), Germany (21) and Greece (1). A certain preference was given to our home turf in this initial survey, because we thought that if there were problems in the instructions, it would be easier to correct closer to home. We received an answer for 40 of the resources, i.e. 34.5%. A response rate above 1/3 makes the survey fairly reliable.

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.

In Table 1 and Picture 4, we see that more than half of the resources were classified into the (RES) restricted category. Approximately one third were classified as (PUB) publicly or (ACA) academically available. Finally, one sixth was found to be exceptional.



Picture 4. Main distribution categories.

One of the publicly available resources (PUB) was also classified as non-commercial, whereas all of the academically available resources (ACA) were non-commercial. The restricted resources (RES) were roughly equally divided among no additional restrictions (27.3 %), a non-commercial restriction (31.8 %) and a requirement that the license be personally granted by the content owner (40.9 %).

Only one resource was such that the content provider found no applicable distribution type because there was no formal agreement between the content owner and the content provider. In this case, the content owner had given his consent to the content provider to use the data by email and the data had been further analyzed by a commercial company providing parsing services, but here as well there was no formal agreement regulating the use of the analyzed data. In addition, there were 5 resources for which the research project was still ongoing and the question of distribution would be discussed only after the project had finished. All in all, some kind of classification was received for a total of 40 resources.

A number of feed-back questions concerned the fact that some corpora did not seem to fit a category completely. In this case either an upgrade agreement needs to be concluded with the content owner or the resource will have to be classified into a more restricted category for which it has all the necessary distribution rights. For this reason a number of legacy resources currently fall into the RES category, even if they probably could be brought into the ACA category by procuring some minor additional rights.

An additional question about the classification process was the issue of how to classify commercially available corpora and who should pay for them. Electronic payment is possible and well-regulated within EU so it is more of a political issue than it is a legal issue how the funding should be arranged. Some of the content providers also saw a need for a full blown digital rights management system, and it is technically possible for certain types of resources, so it is also a political decision for a future CLARIN ERIC whether such resources will be included.

### 4. Future Work

Finally, an important future goal would be to add the necessary research exceptions directly into the national copyright laws as permitted by the EC Infosoc Directive (Directive 2001/29/EC). For this purpose we have created a lobbying message (Oksanen, 2009) aimed at the EU Commission together with DARIAH - Digital Research Infrastructure for the Arts and Humanities and Cessda - Council of European Social Science Data Archives. The purpose of this message is to push the Commission to make the research exceptions mandatory in the national legislations. Writing such a document is a delicate

balancing act. It should be broad enough to bring some benefits. On other hand, too wide demands just cause strong opposing reactions from publishers and lead nowhere.

The current formulation of the lobbying message includes two main points:

- the legislation should *allow free use of copyrighted works for academic purposes* and
- the legislation should *not unreasonably prejudice the legitimate interest of the rights holder*,

which follow the language of the three step test of the Berne Convention<sup>3</sup>. By using this approach, the benefits are the same as using standardized license agreements – the main actors know at least what the language *most likely* means even if there are some ambiguities (Hugenholtz and Okediji, 2008).

Unfortunately, it is unlikely that any lobbying in this area will bring quick results. Most of the resources of the European Commission are currently dedicated to the directive which aims at extending certain aspects of the copyright duration. That process is currently in a gridlock because many of the member countries oppose the directive and before a solution is found, no new hard law pertaining to copyright will be introduced (Hugenholtz & al, 2008). One option is that there will be some kind of general *open data* regulation that would resolve the situation. The movement for creating open databases is currently very strong in some of the member states, e.g. in UK and Finland, and it is not totally out of the question that EU would step in to harmonize the field further than the PSI directive (Directive 2003/98/EC), which covers only public sector information.

One aspect, which we do not cover in depth in this paper are the questions pertaining to the privacy regulation concerning data enabling recognition of persons. However, in most cases a clear written consent from the research subjects to reuse the data for research solves the problems. Older material containing personal data that have been collected without written consent to reuse can still benefit from the exceptions for scientific, historical or statistical research. It should be pointed out that due to the nature of these exceptions material containing personal data is typically available only for the ACA or RES categories unless it is anonymized in which case it typically falls into the PUB category. Anonymizing personal data is often feasible for text data but it may become prohibitively costly for audio and video data.

## 5. Conclusion

It would be preferable to have most of the resources in the

---

<sup>3</sup> The Berne Convention for the Protection of Literary and Artistic Works, usually known as the Berne Convention, is an international agreement governing copyright, which was first accepted in Berne, Switzerland in 1886.

public or at least in the academic domain in order to facilitate sharing, but according to our initial classification test, it seems likely that a sizable portion of the resources for various reasons have restricted access and some will even require a fairly intricate authorization protocol with letters of recommendation and an abstract describing the research purpose. This will hopefully change over time, when researchers realize that they can get citations and fame for making their research material available to others. In addition, some research funding agencies have added the requirement that data collected with their grant funding should be made available to subsequent research projects, which makes sense both from a research financing point of view and from a scientific inter-subjectivity point of view, i.e. the funding agency can avoid paying repeatedly for the same data collection effort and the research results become easier to verify by other research teams.

One obvious problem is that opening resources for research, if they have been created even partially with private funding, should not threaten the business interests of the right holders. There is no easy solution for this and in practice there will always be conflicts of interest when opening databases that have dual usage possibilities, i.e. commercial exploitation and scholarly research, e.g. non-historical news article collections.

## 6. Acknowledgements

We thank the CLARIN FP7 project for the financing and the numerous content owners that took time to answer our questions regarding the distribution types for their resources.

## 7. References

- 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>.
- Hietanen, H., Oksanen, V., and Välimäki, M. (2007). *Community created content*. Law, business and policy.
- Hugenholtz, P.B., and Okediji, R. (2008). *Conceiving an international instrument on limitations and exceptions to copyright*. <http://dare.uva.nl/record/301952>
- Hugenholtz, P.B., Helberger, Dufft, N., and van Gompel, S.J. (2008). *Never Forever: Why Extending the Term of Protection for Sound Recordings is a Bad Idea*, E.I.P.R., 2008-5, p. 174-181.
- Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Hock, H., Lautenschlager, M., Schindler, U., Sens, I., and Wachter, J. (2006). *Data publication in the open access initiative*. Data Science Journal 5: 79–83.
- 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>