The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure: an Estonian and Finnish Perspectives

Aleksei Kelli
Kaarli 3
10119 Tallinn
Estonia
aleksei.kelli@ut.ee

Kadri Vider
J. Liivi 2
50409 Tartu
Estonia
kadri.vider@ut.ee

Krister Lindén
Unioninkatu 40
00014 Helsingin yliopisto
Finland
krister.linden@helsinki.fi

Abstract

The article focuses on the regulatory and contractual framework in CLARIN. The discussion is based on the process analysis approach, which allows an evaluation of the functionality and shortcomings of the entire legal framework concerning language resources and technologies. The article reflects the personal knowledge and insights of the authors gained through their work with legal aspects of language resources and technologies in Estonia and Finland. The analysis may be helpful to CLARIN partners facing similar problems.

Keywords: regulatory and contractual framework, CLARIN agreement templates, contractual and exception model.

1 Introduction[1]

The nature of language resources (LR) and language technologies (LT) can be analyzed from several perspectives such as technological, linguistic, ethical and legal. The authors focus on the legal challenges relating to the development and dissemination of language resources and technologies. From this point of view, the regulatory and contractual framework (legal framework) constitutes one of core infrastructures of CLARIN.

The discussion is based on the process analysis approach, which allows the evaluation of the functionality and shortcomings of the entire legal framework concerning LR and LT. The process starts with the development and ends with the dissemination of language resources and technologies. Different process phases are not addressed separately since they affect each other. Legal issues are defined and analyzed within each phase.

The authors use traditional social science methods and draw on previous legal research conducted by the authors on LR and LT. The analysis relies on the Estonian and Finnish experience. The article also reflects the personal knowledge and insights of the authors gained through work with the legal aspects of LR and LT in Estonia and Finland. The authors make suggestions for improving the existing legal framework. The analysis could also be helpful to other CLARIN partners facing similar problems.

The paper is organized into two main sections. The first section focuses on the establishment of institutional control over the developed language resources and technologies. The second addresses the issue of the development of language resources and deals with the dissemination and potential subsequent utilization of LR and LT. We also study the case of providing public access to fragments of resources in a concordance service versus distributing resources in full for research purposes in light of a research exception in the copyright regulation and the CLARIN contractual framework.

2 Establishment of the institutional control over language resources and technologies

The dissemination and utilization of language resources and technologies depends on several conditions. Institutions distributing resources must have sufficient technological capabilities. Additionally, they must be legally entitled to do so. In practical terms, this means that institutions managing resources must have the capacity to enter into valid transactions and have concluded all relevant contracts regarding LR and LT. In order to avoid too abstract an analysis, the authors use Estonia as an example in addressing these issues.

Estonia has set up the Center of Estonian Language Resources (CELR) as a consortium of 3 institutions at the national level on December 2, 2011. The consortium consists of the University of Tartu (UT) (as leading partner in CELR), the Institute of Cybernetics at Tallinn University of Technology, and the Institute of the Estonian Language. The consortium constitutes an organizational framework for the coordination and implementation of the obligations of Estonia as a member in CLARIN ERIC.

The national consortium is expected to perform obligations, which may bind the whole consortium. The problem, however, is that the national consortium is not a legal person in private or public law (legal entity). The consortium is an agreement. Technically speaking the consortium partners could represent each other but this could create legal uncertainty. In the Estonian consortium, each partner works with certain types of resources. Therefore, the partners have agreed that each one concludes agreements for his respective field of activity.

The Estonian consortium agreement regulates issues relating to the partners' background and foreground IP. However, it does not provide a clear framework concerning the LR and LT developed and owned by persons outside the consortium. To acquire these language resources and technologies, the consortium partners have to conclude agreements with these persons. Since the aim of CLARIN ERIC is to standardize and unify its members' activities, CLARIN has developed standard agreement templates (Licenses, Agreements, Legal Terms). CLARIN also has standard deposition agreement templates. CLARIN deposition license agreements are divided into three categories:

1) CLARIN-DELA-PUB-v1.0 (for public use resources);
2) CLARIN-DELA-ACA-v1.0 (for academic use resources);
3) CLARIN-DELA-RES-v1.0 (for restricted use resources).

In principle, the authors support the ideology of CLARIN having three categories of resources and integrating the approach into a contractual framework (deposition license [DELAs] and end-user license agreements [EULAs]). However, when we analyze specific agreements, we can identify ways of improving them. The process of translating the CLARIN contractual framework (DELAs and Terms of Services) into Estonian and making them compatible with the Estonian law provided a good opportunity to scrutinize once again the existing contracts. A very preliminary meeting was held in Helsinki on 21 May 2015 to discuss the possible amendments to the CLARIN agreement templates and develop them further.[2] The results and observations are discussed below.

The first observation concerns the structure of the DELAs. All DELAs have almost identical provisions. The main difference comes from the provisions concerning intellectual property rights and access rights (Section 7).[3] Therefore, it would be practical to divide the DELA into two parts: a general part for all persons depositing resources and a separate part for selecting a specific category (PUB, ACA, RES). It should be easily achievable in an e-environment.

According to the second observation, the provisions on the warranties and indemnity are among the most important clauses of the DELAs.[4] In the current version of the DELAs, Section 10 regulates liability and indemnity. The provision should be revised to make the regulation clearer. In the following table the current and amended Section 10 is presented:

The current provisions	The amended provisions
10. Legal Obligations	**10. Warranties and indemnity**

[2] The meeting was held in Helsinki on 21 May 2015. Participants: Krister Linden, Kadri Vider and Aleksei Kelli.
[3] There are some differences in annexes too but it should be easy to unify the approach.
[4] The other important part is the license which owners of resources and technologies grant to repositories.

10.1 The Copyright holder shall be responsible for holding a copyright or a sufficient license and/or other rights based on intellectual property law to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright or any other rights based on intellectual property law or other incorporeal right.	10.1 The Depositor warrants and represents that (i) it possesses all proprietary rights, title and interest in the Resource and has full authority to enter into this Agreement. The Depositor shall be responsible for holding copyright, related rights and other rights or a sufficient license and/or other rights to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright, related rights or any other rights.
10.2 The Copyright holder is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1.	10.2 The Depositor undertakes to indemnify and hold harmless the Repository of any liability, directly or indirectly, resulting from the use and distribution of the Resources, including but not limited to claims from third parties. The Depositor is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1.
10.3 Should a third party present a justified claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service.	10.3 Should a third party present a claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service.

In the current version, Section 10 is called "Legal Obligations". This is not the best choice of words since all obligations arising from a contract are legal. Therefore Section 10 should be called "Warranties and indemnity". The amended version also reflects a new terminological approach. The DELA terms identifying the parties to the agreement are replaced as follows: the Copyright curator (CLARIN Centre receiving LR and LT) is replaced with "repository" and the Copyright holder (person licensing LR and LT) with "depositor". Subsection 10.1 and 10.2 are elaborated further to increase clarity. Subsection 10.3 was amended to provide sufficient grounds for removal of resources if a third party presents a claim that her rights are violated. The repository does not have to prove that the claim was justified. In addition, Subsection 10.3 must be compatible with the CLARIN Notice and Take Down Policy.

An additional issue regarding the deposition of resources concerns the question whether institutions should accept LR and LT on an as-is basis without any representations and warranties. In case a depositor does not have title to the resources, it would be out of the question. If resources are developed based on the copyright exception and/or include personal data, DELAs for the categories ACA or RES are suitable.

3 Development and dissemination of language resources

Language resources have two tiers of rights: 1) the rights of the persons who developed the resources and 2) the rights of the persons whose copyright-protected content (sometimes also content with related rights) was used when creating the resources. In the previous section, we addressed the first tier and here we focus on the second tier.

From a legal perspective, language resources constitute copyright protected databases (see Kelli et al. 2012; Tavast et al. 2013). The creation of language resources often requires the use of copyright protected works. The use, however, can be based on two models: 1) the contract model and 2) the exception model. The contract model means that a person developing language resources acquires permission to use copyrighted works (books, journal articles, etc.). The exception model is based on a copyright exception allowing free use of works for research purposes. Both models have their pros and cons.

The contract model allows negotiating terms for commercial use of resources and making them publicly available. The model is expensive even if copyright holders do not ask for remuneration. Administrational costs arise during the negotiations and the management of contracts. There is no guarantee that the right-holders use identical contracts. This could lead to incompatibility between different contracts and restrict the development and dissemination of resources. Another problem is *de facto* orphan

works (anonymous web posts, blogs etc.) since there is no one identifiable who can give permission for their use.

The advantage of the exception model is that there is no need to ask for permission from the right-holders. It is possible both to use works of identified authors and works of unidentifiable authors (*de facto* orphan works). There is no administrative burden to negotiate licenses. The main disadvantage is that it is not possible to use the developed resources for commercial purposes or make them available in the PUB category. Dissemination is possible only in the categories ACA and RES.

The Estonian Copyright Act has a general research exception allowing development of language resources (Autoriõiguse seadus § 19). The draft Copyright and Related Rights Act introduces a specific exception for data mining and text analysis worded as follows: "reproduction and processing of an object of rights for the purpose of text analysis and data mining, on the condition of attributing the name of the author of the used work, the name of the work and the source of publication, except if such attribution is impossible, and on the condition that such use is not carried out for commercial purposes". It was added for the sake of legal clarity.

Finland relies on the contract model. FIN-CLARIN has refrained from paying for resources but has contributed a minimal sum towards the collective extended license for the Finnish billion word newspaper corpus which has been scanned and OCRed by the National Library of Finland comprising newspapers from 1792 to the present. FIN-CLARIN provides access to the full corpus for non-commercial research purposes and access to anyone for small excerpts based on search results.

Similarly the billion word blog Suomi24 maintained and distributed by the commercial company AllerMedia is available in full for non-commercial research purposes via FIN-CLARIN but excerpts can be used by anyone. The motivation for this by AllerMedia is that it welcomes ideas provided by the research community by facilitating access and hopes to provide access to the same data to commercial companies against a fee.

Language resources and technologies are made available within CLARIN through a specific contractual framework. Firstly, a person interested in using LR and/or LT has to accept the Terms of Services (TOS) (Licenses, Agreements, Legal Terms). The DELAs and TOS are two sides of the same coin. When DELA shifts all liability regarding language resources and technologies to the depositors, the TOS disclaims and limits CLARIN's liability regarding resources to the maximum extent allowed by law. Drawing on public licenses such as EUPL, GPL and Creative Commons, we suggest amending Section 5 of TOS so that it is absolutely clear that the resources are provided on an as-is and as-available basis and no liability is assumed.

In addition to the TOS, the prospective user also has to accept the EULA attached to the language resources and technologies.

When it comes to the dissemination of language resources, it is useful to remember the maxim of Roman law saying "*Nemo plus iuris ad alium transferre potest, quam ipse haberet*" (Dig. 50.17.54). This means you cannot give others more rights to something than you have yourself (see Zimmermann, 1996). In other words, resources developed based on a research exception cannot be licensed in the PUB category. In view of this, we study how to provide public access to fragments of resources *versus* distributing resources in full for research purposes using the CLARIN contractual framework. A set of excerpts (*i.e.* search results) may be considered derived works, which are subject to the same conditions as the original work unless otherwise agreed. We may therefore still need a DELA to acquire the right to distribute the search results publicly.

In most cases, the right-holders are willing to make excerpts publicly available while the full corpus is only distributed for academic or restricted purposes. In case there is no research exception, this can still be agreed using one DELA (as the PUB/ACA/RES templates are quite similar). In both cases, the resource needs two metadata records: one for pointing to the PUB excerpts and one for pointing to the original ACA/RES resource, *i.e.* we have data with two different uses provided by two licenses in one agreement.

4 Conclusion

The regulatory and contractual framework constitutes an integral component of the CLARIN infrastructure. Similar to technical standards CLARIN also needs unified legal standards. It is in the interest of CLARIN to lobby for an EU-wide mandatory text and data mining exception.

CLARIN agreement templates have to be integrated and evaluated as one functional system. There are two ways for language resources and technologies to enter CLARIN. Firstly, the employees' rights regarding LR and LT are transferred to the employer (i.e. CLARIN national consortium member). Secondly, authors who are not employed by a CLARIN national consortium member have to conclude a deposition agreement. The aim of the DELA is to shift the liability for the resources to its depositors. The wording of the relevant provisions needs to be amended accordingly so that it is very explicit that a depositor is responsible for the resource.

Users can access CLARIN resources after they have agreed to the Terms of Services. The TOS objective is *inter alia* to limit CLARIN's liability towards users. Resources and technologies are offered on an as-is and as-available basis. The wording of the provisions regulating liability in TOS should be amended to limit CLARIN liability to the maximum extent.

Reference

[Autoriõiguse seadus § 19] Autoriõiguse seadus [Copyright Act] (as entering into force on 12.12.1992). RT I 1992, 49, 615; RT I, 29.10.2014, 2 (in Estonian). Unofficial translation available via https://www.riigiteataja.ee/en/eli/531102014005/consolide (accessed on 12 July 2015);

[Kelli et al. 2012] Aleksei Kelli, Arvi Tavast, Heiki Pisuke (2012). Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. – Juridica International (19), 40-48;

[Dig. 50.17.54]. Available at http://www.thelatinlibrary.com/justinian/digest50.shtml (13.7.2015);

[Licenses, Agreements, Legal Terms]. Available at http://clarin.eu/content/licenses-agreements-legal-terms (13.7.2015);

[Tavast et al. 2013] Arvi Tavast, Heiki Pisuke, Aleksei Kelli (2013). Õiguslikud väljakutsed ja võimalikud lahendused keeleressursside arendamisel (Legal challenges and possible solutions in developing language resources). – Eesti Rakenduslingvistika Ühingu Aastaraamat (9), 317-332;

[The draft Copyright and Related Rights Act] Autoriõiguse ja autoriõigusega kaasnevate õiguste seaduse eelnõu. Versioon: 21.7.2014 [The Estonian draft Copyright and Related Rights Act. Version: 19.7.2014]. (in Estonian), https://ajaveeb.just.ee/intellektuaalneomand/wp-content/uploads/2014/08/Aut%C3%95S-EN-19-7-2014.pdf, (accessed on 5 May 2015);

[Zimmermann, 1996] Reinhard Zimmermann. The Law of Obligations Roman Foundations of the Civilian Tradition. – Oxford University Press, 1996.