UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

# The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure

**Kelli, Aleksei; Vider, Kadri; Lindén, Krister**

**DeSmedt, Koenraad**

**2016-04-11**

http://hdl.handle.net/10138/174339

Kelli, A, Vider, K & Lindén, K 2016, The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. in K DeSmedt (ed.), Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland : NEALT Proceedings Series. vol. 123, 002, Linköping Electronic Conference Proceedings, vol. 123, Linköping University Electronic Press, Linköping, pp. 13-24, CLARIN Annual Conference, Wroclaw, Poland, 15/10/2015. < http://www.ep.liu.se/ecp/123/002 /ecp15123002.pdf >

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

# The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure

**Aleksei Kelli**
Department of Private Law
University of Tartu, Estonia
aleksei.kelli@ut.ee

**Kadri Vider**
Centre of Estonian Language Resources
University of Tartu, Estonia
kadri.vider@ut.ee

**Krister Lindén**
Department of Modern Languages
University of Helsinki, Finland
krister.linden@helsinki.fi

**Abstract**

The article focuses on the regulatory and contractual frameworks in CLARIN. A process analysis approach has been adopted to allow an evaluation of the functionality and shortcomings of the entire legal framework applicable to language resources and technologies. The article discusses and provides background information to amendments of key provisions of CLARIN license templates. The authors also address issues relating to the research exception allowing for the development of language resources without the copyright holder's consent. The article introduces some practical information on a new version of the license category calculator. The article reflects the authors' personal understanding and insights gained by examining the legal aspects of language resources and technologies in Estonia and Finland.

## 1 Introduction

The nature of language resources (hereinafter resources) is defined[1] as software, applications and/or databases[2]. This definition can be analysed from a wide range of perspectives such as technological, linguistic, ethical and legal. We focus on the legal challenges relating to the development and distribution[3] of language resources and technologies. In view of this, the regulatory and contractual framework (hereinafter legal framework) constitutes one of the core infrastructures of CLARIN.

We base the discussion on our analysis of the process for developing and distributing language resources. In this paper we employ the analysis to facilitate our evaluation of the functionality and shortcomings of the entire legal framework concerning language resources. The process commences with the development and results in the distribution of language resources. However, we do not address different process phases separately since they are clearly intertwined. Instead, we identify and analyze legal issues across individual phases.

We resort to traditional methods in social sciences and draw on the previous legal research conducted by the authors. The analysis incorporates the Estonian and Finnish experience. Both countries are CLARIN ERIC members. The article reflects the personal understanding and insights of the authors gained while studying the legal aspects of language resources in Estonia and Finland. Subject to our insights, we have amended the CLARIN license agreement templates and terms of service. We provide background information to the amendments by offering suggestions how to improve the existing legal framework. Since the article outlines the practical legal issues pertaining to the management of resources, it can be of use to other CLARIN members as well.

We construe the Terms of Service (TOS) to mean the general conditions for using a CLARIN service, the End-User License Agreement (EULA) to mean the conditions designed for an end-user to use

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

[1] See Article 3 of a CLARIN Deposition License Agreement (e.g. CLARIN-DELA-PUB-v1.0).

[2] Note that from a legal perspective, in a research infrastructure like CLARIN, researchers primarily use works as databases from which they extract facts, in contrast to libraries from which they primarily borrow copies of works.

[3] Distribution means, inter alia, making available to the public, communication to the public and distribution to the public by wire or wireless means.

a language resource, and the Deposition License Agreement (DELA) to mean the conditions on which a CLARIN service can distribute a language resource to end-users. For an overview of the relationship between these concepts, see Figure 1. We designate main license category to mean one of the three broad categories of usage rights and restrictions conferred on the end-user by TOS, DELA and EULA, see Section 2. In addition, several subcategories entailing specific rights and restrictions have been defined, see Section 4. A laundry symbol[4] is an icon attributed to a license category.

The paper is organized into three main sections. The first section focuses on the establishment of institutional control over the existing language resources. The second addresses the issue of the development of language resources and deals with their distribution and potential subsequent utilization. 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 in the CLARIN contractual framework. The third section explains the concept of a license category calculator.

## 2 Establishment of the institutional control over language resources and technologies

The distribution and utilization of language resources and technologies depends on several conditions such as technological capabilities, the existence of resources, etc. The institutions and organizational units managing resources must also have the legal capacity to enter into valid transactions and obtain sufficient rights to distribute language resources. In order to avoid a purely abstract discussion of the legal framework characteristic of national CLARIN consortiums, we rely on Estonia as an example when addressing these issues.

Estonia 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 the 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 Estonian consortium agreement regulates issues relating to the background and foreground intellectual property[5] (IP) of the Estonian partners. However, it does not provide a clear framework for the resources developed and owned by persons outside the consortium. To acquire their language resources, the consortium partners have to conclude individual agreements with them. However, this poses a serious problem as the Estonian national consortium is not deemed a legal person in private or public law (also called legal entity). The consortium is an agreement between independent partners to cooperate on specific issues.[6] In theory, the right of representation could be derived from the consortium agreement and the partners could represent each other. However, this may create legal uncertainties as to the scope of the right of representation. Therefore, a possible way forward is to develop an understanding where each consortium partner concludes agreements governing certain types of language resources with external partners. This is compatible with the current situation where each partner is responsible for certain types of resources. Another option would be to establish a legal entity (e.g., a non-profit association, a private limited company, etc.).

The acquisition and distribution of resources within the CLARIN framework by its members has to be aligned and standardized as much as possible. Standard agreements constitute a key infrastructure of CLARIN, and CLARIN has developed standard agreement templates (Licenses, Agreements, Legal Terms) which can be used for this purpose.

CLARIN standard agreement templates are based on a conceptual division of all language resources into three main categories, i.e. resources which are publicly or openly available (PUB), those which are available for research or academic use (ACA) and those which are restricted to individual use, e.g.

[4] In the textile industry, a laundry symbol represents the maximum permitted treatment. The CLARIN symbols indicate in a similar vein a "maximum treatment" of language resources that can safely be permitted to an end-user by the CLARIN repository based on its DELAs, TOS and EULAs.

[5] Background IP are typically resources that existed before the consortium and foreground IP are resources created within the consortium.

[6] These include the items mandated for national consortia by the CLARIN ERIC Statutes (http://www.clarin.eu/sites/default/files/OJ-2012-136-EU-Decision.pdf)

due to personal data content, (RES). According to Oksanen & al. (2010) explaining the conceptual background and the evolution of this approach, this categorization is based on an extensive survey indicating that it is possible to group licenses in this manner. The PUB category allows wide distribution. The ACA category is designed to make resources available for research purposes[7] and RES permits limited use with additional requirements relating to data protection, e.g. research plan, etc. In addition to the three main categories, there are several subcategories for a more nuanced picture of the conditions of use as discussed in Section 4 on the License Category Calculator.

The authors support the ideology of CLARIN having a tripartite[8] main division of resources integrating this into a contractual framework for several reasons. Firstly, the current division contains more or less all language resources. Its appropriateness has been proven in practice[9] and no major problems have been identified. Secondly, if adopting a different main division, CLARIN would have to address the legal status of language resources and technologies having already entered CLARIN. Subsequently, a need would arise for the reclassification of already deposited and distributed resources. Thirdly, PUB, ACA and RES categories are already integrated into the CLARIN infrastructure. They are accepted by CLARIN stakeholders and thereby socially embedded. The potential adoption of a new main division would create unnecessary confusion. Nevertheless, the authors are not asserting that CLARIN should not develop the current categorization further. The key idea is that potential changes have to be made within a clear conceptual framework, be mature enough, absolutely necessary and outweigh potential negative aspects. An evolutionary approach, where the categorization is changed incrementally and gradually (i.e., the existing categories are specified, subcategories adopted, etc.), should be preferred over a radical approach (i.e., a totally new categorization). Stability as a value is a valid consideration.

A starting point for the analysis of the CLARIN license templates is the deposition license agreements (DELA) which resource right-holders can form when depositing resources in national CLARIN repositories. They are divided into three categories:

1) CLARIN-DELA-PUB-v1.0 (for publicly or openly available resources);
2) CLARIN-DELA-ACA-v1.0 (for resources for research or academic purposes);
3) CLARIN-DELA-RES-v1.0 (for resources restricted to individual use).

The acceptance of a general conceptual framework dividing language resources and agreements into PUB, ACA and RES does not preclude amendments to the existing templates. The process of translating the CLARIN agreement templates (DELAs and Terms of Service) into Estonian and bringing them into conformity with the Estonian legislation offered a good opportunity to scrutinize once again the existing agreement templates. The results and observations are discussed below.[10] An outline of the relationship between the CLARIN agreement templates is provided in Figure 1.

[7] In its fundamental form, ACA covers both commercial and non-commercial research.
[8] CLARIN is not the only community to have a tripartite division of resources, e.g. the ORCID community (http://orcid.org/) has a similar division with the ORCID Privacy Settings defining access for all, for a trusted community, or for the owner (http://support.orcid.org/knowledgebase/articles/124518-orcid-privacy-settings). While ORCID takes the perspective of the data producer, CLARIN has an end-user perspective on data access.
[9] The categories are currently in use as license metadata in the Virtual Language Observatory by CLARIN (https://www.clarin.eu/content/virtual-language-observatory)
[10] The new templates: http://www.helsinki.fi/finclarin/calculator/ClarinLicenseCategory.html (24.11.2015).

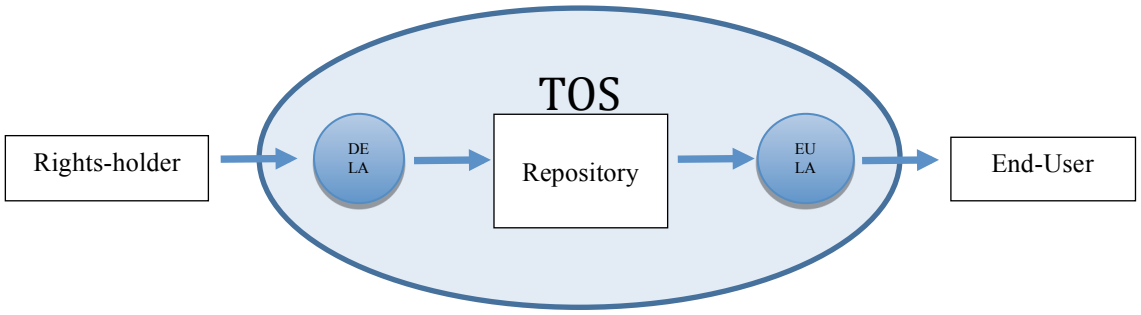


Figure 1. An outline of how the CLARIN agreement templates confer access rights

The first observation concerns the structure of DELAs. All DELAs have almost identical provisions. The main difference emanates from the provisions concerning intellectual property rights and access rights (Section 7 of DELA).[11] Therefore, it would be practical to have one standard deposition agreement and three different annexes regulating intellectual property (IP) matters. IP provisions mainly[12] determine the categorization of resources into PUB, ACA or RES.

According to the second observation, the provisions on the warranties and indemnity are deemed essential clauses[13] of DELAs and therefore require special scrutiny. In the previous versions of DELAs, Section 10 regulates liability and indemnity. The provision was revised to simplify the regulation. In the following table, the previous and amended Section 10 are presented in parallel:

| The previous provisions | The amended provisions |
|---|---|
| 10. Legal Obligations | 10. Warranties and indemnity |
| 10.1 The Copyright holder shall be responsible for holding 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 for 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. |

[11] There are also some differences in the annexes, which are easily brought into conformity.
[12] The use of personal data also has an impact on distribution of resources and their categorization.
[13] Another important part is the license granted to repositories by the owners of resources and technologies.

In the previous version, Section 10 is named "Legal Obligations". This is not the most appropriate wording since all obligations arising from a contract are legal. Therefore, Section 10 should be called "Warranties and indemnity". Subsection 10.1 and 10.2 are elaborated further to increase clarity. Subsection 10.3 was amended to provide sufficient grounds for the removal of resources if a third party files a claim due to the infringement of her rights, and the repository is under no obligation to prove that the claim was justified. In addition, Subsection 10.3 must be compatible with the CLARIN Notice and Take Down Policy.

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".

DELAs use the term "distribution" in a broad sense. This could cause misinterpretations since international conventions, the EU directives and national laws usually confine "distribution" to the context of tangible goods. For instance, Article 6 (1) of the WIPO Copyright Treaty (WCT) defines "distribution" as making works or their copies available to the public through the sale or other transfer of ownership. According to the agreed statements concerning Articles 6 and 7 added to WCT, the right of distribution under the said Articles refers exclusively to fixed copies that can be put into circulation as tangible objects (WCT 1996). Article 4 of the Directive on the harmonization of certain aspects of copyright and related rights in the information society (Information Society Directive 2001) has a similar approach. Language resources, however, are not distributed in a tangible form but are made available on-line. Acknowledging this discrepancy, there are two options: 1) replace the term "distribution" with "communication to the public" which is an umbrella term encompassing any communication to the public, by wire or wireless means, including making available to the public in such a way that members of the public may access them from a place and at a time individually chosen by them (Article 3 (1) of the Information Society Directive) or 2) define the term "distribution" widely so that it is compatible with the actual practice. Opting for the first could have created new issues. Some resources have already clearly been deposited using DELAs containing the term "distribution". Should we now decide to modify DELAs and replace "distribution" with "communication to the public" then this could give rise to a question whether resources deposited prior to the change only provide for the distribution of resources in a tangible form (e.g., on CD, print-outs, etc.). Any chances of a misunderstanding ought to be avoided at the outset. Therefore, the second option was preferred. The current version of DELA (Section 3) defines "distribution". According to the definition "*Distribution* means, inter alia, making available to the public, communication to the public and distribution to the public by wire or wireless means."

## 3 Development and distribution of language resources

Two tiers of rights are applicable to language resources: 1) the rights of the persons who put their intellectual effort into developing the resources (within employment transferred to the employer) and 2) the rights of the persons whose copyright-protected content (sometimes also content with related rights) was used for creating the resources. For an illustration of the tiers, see Figure 2.

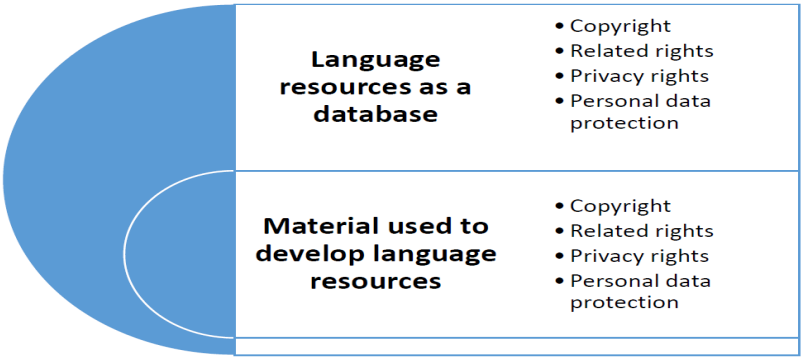


Figure 2. The two tiers of rights covering language resources.

From a legal perspective, language resources constitute copyright protected databases (Kelli et al. 2012; Tavast et al. 2013). The creation of language resources often requires the use of copyright protected works. The use of copyrighted material can be based on two models: 1) the contract model and 2) the exception model. Both models have their strengths and weaknesses.

### 3.1 The contract model

The contract model means that a person developing language resources acquires a permission (a license) to use copyrighted works (books, journal articles, etc.). The contract model allows negotiation of suitable terms for commercial use of copyrighted material to develop resources. This model contains two major problems. Firstly, the contract model involves high administrative costs relating to negotiation of contractual terms and management of contracts (especially in the absence of standard agreements). Secondly, incompatibilities between different contracts could restrict the development and distribution of resources. It should also be borne in mind that there are *de facto* orphan works (anonymous web posts, blogs etc.). Their authors are not reasonably identifiable and there is no one who can grant permissions for their use. Therefore, the contract model is regarded as expensive and non-functional.

Finland has opted for the contract model. FIN-CLARIN has refrained from paying for resources but has contributed a minimal sum towards the collective extended license[14] for the Finnish billion-word newspaper corpus which has been scanned and OCRed by the National Library of Finland comprising newspapers from 1792 to date. In December 2015, the FIN-CLARIN collective license has been extended to cover the copyright of all printed works in Finland subject to the consent of the editor and the publisher. However, most *de facto* orphan works are still not included but require separate agreements. FIN-CLARIN provides access to the full corpora 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 through a separate agreement in full for non-commercial research purposes via FIN-CLARIN and as excerpts for anyone. The motivation for this by AllerMedia is that the company welcomes and encourages ideas developed by the research community by facilitating access to the data, and looks forward to providing access to the data for commercial companies against payment of a fee.

### 3.2 The exception model

The exception model is based on a copyright exception allowing the free use of works for research purposes. For instance, Section 19 of the Estonian Copyright Act provides a general research exception allowing, inter alia, the use of copyright protected works for the development of language resources. The development of language resources in Estonia takes place within the framework of the research exception. It should be noted, however, that there is no case law regarding the exact scope of the exception. For the sake of legal clarity 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" (for further discussion on copyright reform in Estonia, see Kelli 2015).

The exception model has several advantages. First of all, there is no need to ask for permissions from the right-holders to use copyrighted content. There is no administrative burden to negotiate licenses. It is equally possible to use works of identified authors and works of unidentifiable authors (*de facto* orphan works). The main disadvantage is that it is not possible to use the developed resources for

[14] The National Library is authorised by law to make the works available within its premises. Works whose copyright has expired can be displayed on the National Library website, but anything after 1911 is still regarded as potentially containing copyright and therefore needs a license by the collected extended license provided by the collecting society Kopiosto (http://www.kopiosto.fi/kopiosto/).

commercial purposes[15] or make the entire resource available in the PUB category. Distribution of the entire resource is possible only in the ACA and RES categories.

In this context, it is necessary to consider policy issues as well. Data originally created by researchers for research purposes should have public or open licenses. Data created for other purposes can enter the research domain by licensing or by a statutory research exception. The license template for ACA states that the teaching, education and research-purpose is the underlying criterion for ACA resources. The question is how willing the right-holders are to provide resources for research without any further control over the usage and distribution. This is also the fundamental concept behind a statutory research exception.

Typically, a statutory research exception makes data available only for non-commercial purposes. In addition, statutory research exceptions do not usually recognize intermediaries such as research infrastructures hosting the data on behalf of right-holders. Statutory exceptions often only allow data to be used by (but not distributed by) individuals, so a research infrastructure like CLARIN may need to license the data to distribute it even in the case of such a statutory exception for research in the EU copyright legislation.

### 3.3 Right-holders concerns and an interim solution

A relevant question is what additional conditions right-holders could impose to let CLARIN distribute their data for them? To answer this question we should examine the level of verification of the teaching, education and research-purpose the right-holders themselves apply to someone seeking their data. Is it sufficient for them that the user has a researcher status, an IP address at a research institution, a research plan or a self-declared research-purpose?

Currently an IP address at a research institution seems to be enough for many publishers to recognize a vague identity in order to give access to individual works. An identified user with researcher status would provide much stricter identification, and it would have the advantage of not tying the researcher to a location. However, if the researcher status or research purpose cannot be reliably verified, a proxy such as an IP address at a research institution also seems to suit the publishers.

If a mere IP address is acceptable, then maybe it is rather secondary that the user attributes (provided by an identity federation like EduGain) only roughly match a researcher identity. It probably suffices to distinguish between persons affiliated with non-commercial research institutions as opposed to guest users in order to convince the right-holders that a research infrastructure like CLARIN can safely distribute data for non-commercial research purposes. Distributing to a limited, albeit a substantial, number of users so as not to destroy the remaining market in the process is the main concern of right-holders regarding a broad research exception.

While waiting for the adoption of a broad statutory research exception in the EU-wide copyright legislation that will include infrastructure facilities like CLARIN ERIC, we propose the ACA license category to serve the same purpose. Until such a statutory research exception has been implemented, the resources in the ACA category require explicit agreements between CLARIN Centers and the right-holders.

### 3.4 Distribution of resources

Language resources are made available within CLARIN through a specific contractual framework. Firstly, a person interested in using the resources agrees to accept the Terms of Service (TOS) further specified in DELAs and EULAs. DELA is a resource specific set of usage permissions and restrictions while TOS is the general framework. When DELA shifts all liability regarding language resources to depositors, TOS disclaims and limits CLARIN's liability regarding resources to the maximum extent allowed by law. The users can access resources on as-is and as-available basis. CLARIN should not be held liable for potential errors and "bugs" in the resources. Resources are considered work in progress. Drawing on public licenses such as the European Union Public Licence (EUPL), the GNU General Public License (GPL) and Creative Commons (CC), we amend Section 5 of the TOS so that it is abso-

[15] The issue whether commercial research is allowed under the exception model remains somewhat controversial. Since the author does not get any remuneration when her works are used, then it is more likely that the research for the commercial purposes is not allowed.

lutely clear that the resources are provided on an as-is and as-available basis and no liability is assumed. In addition to TOS, the prospective user also has to consent to EULA annexed to the language resources and technologies.

As to the distribution of language resources, it is useful to remember the maxim of Roman law stating "*Nemo plus iuris ad alium transferre potest, quam ipse haberet*" (Dig. 50.17.54). This means you cannot grant others more rights to something than you have yourself (see Zimmermann, 1996). In other words, resources developed based on the research exception cannot be licensed in the PUB category.

In this context there are three issues which need to be addressed: 1) the acceptance of resources on as-is basis 2) the distribution of fragments of resources and 3) the distribution of virtual resources.

A question that CLARIN Centers could face is whether they should accept resources on an as-is basis. In case the depositor does not have rights to a resource as a database (the first tier of rights), it is ultimately out of the question. If the depositor has rights to a resource as a database but its development was based on the copyright exception and/or it includes personal data, the resource can probably be accepted. It can most likely be made available as an ACA or RES resource.

Another frequently raised relevant issue is the difference between licensing whole works as opposed to fragments of works. Copyrighted fragments are derived works of the original and therefore might hold the same license as the original unless otherwise agreed. In some cases, the original right-holders may be willing to permit distribution of derived fragments, e.g. sentences or paragraphs through a concordance tool such as https://korp.csc.fi, while being reluctant to provide a license to distribute the full work. The question is how to indicate this in the Deposition License Agreement (DELA) and in the metadata description of the resource without giving a misleading representation of the end-user's rights.

If a resource cannot be distributed under any circumstances, it has a proprietary license that does not even fall into the CLARIN RES category. In practice, the original may not be available to anyone besides a search tool provider. However, search results consisting of fragments such as sentences or paragraphs may still be available to anyone in the CLARIN PUB category. Describing the resource as being accessible in the CLARIN PUB category would be a misrepresentation, because it gives a false impression that the whole work is available for download.

One solution is to explicitly state in DELA that derived fragments can be distributed through a search interface and specify the license applicable to the search results. Even if the original resource cannot be distributed, the search results, i.e. the derived works, can be distributed in e.g. the CLARIN PUB category based on DELA.

The remaining problem is technical because such a virtual derived work cannot necessarily be given a persistent identifier (PID) in advance, even if the underlying resource has a PID. A new derived work will arise every time someone makes a search, i.e. a search produces a virtual corpus. One solution is to provide the search parameters with a PID on the fly. The search parameters are a combination of the query, the PID of the search engine and the PID of the underlying corpus. The search parameters are comparable with a small program which operates a search engine on a corpus, and can be regarded as a work on its own. The search program can be efficiently stored and persistently identified to be shared with others reproducing the virtual corpus as a search result. Similar solutions are needed for distributing virtual corpus collections in a federated search environment.

## 4 License category calculator

The license category calculator is a tool for assigning metadata to a license and now also helps the depositor determine the right DELA when depositing a resource in a CLARIN repository. The old license category calculator only provided the license metadata. The new calculator also proposes a DELA that can be signed by the depositor and the repository.

To fully grasp what an end-user can do with a resource, CLARIN provides a resource with license metadata, also known as "laundry tags" or license categories. The goal is to have icons for each laundry tag to make the permissions and restrictions visually recognizable for the end-user. The idea for license subcategories was adopted from the Creative Commons (CC) initiative. Based on a survey and a trial labeling of licenses of several hundred resources from various European countries (Oksanen et

al. 2010), subcategories were developed based on frequently occurring access conditions for language resources. The subcategories are detailed in the list of questions in Figure 3.

Result: **CLARIN PUB**

[TOS] [EULA] [DELA] [see also Open License Selector]

| | | |
|---|---|---|
| **Identification and Access conditions** | | |
| | Does the user need to be authenticated, i.e. identified? | ○ Yes ◉ No |
| | Does the user need to be affiliated with some specific community, e.g. through a university or research institution (EDU) or a community of language resource and technology researchers more generally (META)? | ○ EDU<br>○ META<br>◉ No |
| | Can the user only be given permission to use the resource on a case-by-case basis, e.g., based on a mandatory fee or a research plan? | ○ Yes ◉ No |
| **FF** | Is a fee required to get access to the resource? | ○ Yes ◉ No |
| **PLAN** | Does the right holder require a research plan for granting access? | ○ Yes ◉ No |
| **General use conditions** | | |
| **BY** | Is attribution, i.e. acknowledgement of authorship, required? | ○ Yes ○ No |
| **NC** | Is the content available only for non-commercial purposes? | ○ Yes ○ No |
| **INF** | Is informing the rights owner about the use of the resource required? | ○ Yes ○ No |
| **LOC** | Is the content available only at a single location, center, or site? | ○ Yes ○ No |
| **LRT** | Is the content available only for language research and technology development? | ○ Yes ○ No |
| **PRIV** | Are there personal data in the resource? | ○ Yes ○ No |
| **Distribution conditions** | | |
| **NORED** | Can the user distribute the original resource to third parties? | ○ Yes ○ No |
| **ND** | Can the user distribute derived works, i.e. works containing copyrighted parts of the original? | ○ Yes ○ No |
| **SA** | If the user can distribute derived works, should the same license be used, i.e. is the license reciprocal? | ○ Yes ○ No |
| **DEP** | If the user cannot distribute derived works, is the user still allowed to distribute modified versions via CLARIN? | ○ Yes ○ No |
| **Other conditions** | | |
| * | Are there other non-standard conditions in the license that the user should pay attention to? | ○ Yes ○ No |

Figure 3. License Category Calculator

To assist CLARIN Centers with the labeling and classification of licenses when the depositor offers a language resource for distribution via CLARIN, a license category calculator has been developed: https://www.clarin.eu/content/clarin-license-category-calculator. A new version of the license calculator also provides CLARIN Centers with a set of license templates outlined in the previous sections including the Terms of Service (TOS), the End-User License Agreement (EULA) and the Deposition License Agreement (DELA). The variable content of the templates corresponds to the conditions that can be identified in an existing resource license or in the wishes of the depositor of a new language resource. The conditions include commonly occurring restrictions and permissions with regard to user identification and access as well as resource usage and distribution. By answering the yes-no questions in the license category calculator, the laundry tags on the top line of the calculator are interactively updated. When the relevant questions have been answered, the depositor can click on the EULA and

DELA buttons to view the EULA and to print and sign the corresponding category-specific DELA before submitting the resource to a CLARIN repository[16]. For an illustration, see Figure 3.

Other license templates may be used by a CLARIN center if such templates effectively comply with e.g. national legislation. The templates provided by CLARIN ERIC are intended to serve as a first-aid kit for CLARIN Centers that have not yet developed an electronic resource submission workflow. In an electronic submission workflow, many of the boilerplate provisions in the deposition agreement template (DELA) can be incorporated into the Terms of Service (TOS) and the end-user conditions can be verified by an interface like the license category calculator so as to leave virtually only the end-user license conditions (EULA) for approval by a depositor.

## 5 Conclusion

The regulatory and contractual frameworks constitute an integral part of the CLARIN infrastructure. The regulatory framework is adopted by the CLARIN member states of the European Union. CLARIN can influence its development by issuing policy recommendations and using other means of lobbying. However, the contractual framework is adopted by CLARIN itself and enacted by CLARIN members, and therefore, CLARIN has direct control over its contractual framework.

The CLARIN contractual framework is based on a conceptual division of language resources together with their corresponding agreements into three main categories: PUB (publicly and openly available), ACA (research and academic use) and RES (restricted to individual use). The authors support this tripartite categorization since it comprises more or less all language resources, it has proven its practical value and is well-integrated and socially embedded in the CLARIN infrastructure.

A review of CLARIN agreement templates has to be conducted within the context of development and distribution of language resources and take into the account the legal nature of resources. Resources are usually governed by two tiers of rights. The first tier of rights covers the resources used as a database. The second tier encompasses the intellectual effort used for creating the database.

The creation of resources usually requires extensive use of copyrighted material. It can be grounded in two models: 1) the contractual model (a license for use of copyrighted material is acquired); and 2) the exception model (the use of copyrighted material is based on a statutory exception). The selected model affects how resources can be used further. Estonia has opted for a research exception covering the development of language resources. For the sake of clarity, the Estonian draft Copyright and Related Rights Act introduces a specific exception for data mining and text analysis. Finland relies on the contractual model. It is in the interest of CLARIN to lobby for an EU-wide mandatory statutory exception for research purposes.

Since the CLARIN license agreement templates have to be integrated and evaluated as a uniform functional system, we proceeded from the three deposition agreements. They are almost identical. The main difference resides in the intellectual property and access provisions. Therefore, it is more expedient to consolidate these provisions into three annexes. This way we have only one deposition agreement and the depositor can choose from among the three annexes, which determine whether the deposited resource is PUB, ACA or RES.

The deposition agreement purports to shift the liability for the resources to their depositors. The wording of the relevant provisions is amended accordingly so that it becomes very explicit that the depositor is liable for the resources and indemnifies CLARIN from any damage claims.

A key issue is 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 subject to the same conditions as the original work unless otherwise agreed. We may, therefore, still need a deposition agreement to acquire the right to distrib-

[16] Note that also publicly available repositories like GitHub or Sourceforge have elaborate terms of service, e.g. https://help.github.com/articles/github-terms-of-service/ or https://slashdotmedia.com/terms-of-use/, that require depositors of resources to adhere to strict policies even if the deposition processes are streamlined to make the deposition as smooth as possible. GitHub in particular requires the resource to have an open license in order for the resource to be freely and openly distributed by GitHub. For a more limited distribution, GitHub charges a monthly fee. CLARIN distributes public and open resources with, e.g. CC licenses in the PUB category, and resources with more limited licenses within the confines of the ACA and RES categories.

ute the search results publicly. In most cases, 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 on in a deposition agreement with a relevant annex (PUB, ACA, RES). In both, the resource needs two metadata records: one applicable to the PUB excerpts and the other applicable to the original ACA/RES resource, i.e. we have data with two different uses provided by two licenses in one agreement.

The next central agreement in the CLARIN system is the Terms of Services (TOS). Before users can access resources, they have to agree to the Terms of Services. The objective of TOS 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 is amended to limit CLARIN's liability to the maximum extent permitted.

A license category calculator has been developed to assist CLARIN Centers with the labelling and classification of licenses when language resources are offered for distribution. The article provides information on a new version of the calculator automatically generating CLARIN End-User and Deposition License Agreements with reference to the Terms of Service.

## References


[CC] Creative Commons. Available at http://creativecommons.org/licenses/ (24.11.2015);

[DELA]. CLARIN PUB Deposition License Agreement Template 1.1. Available at https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinDelaV11?PUB=1 (24.11.2015);

[Estonian Copyright Act] Autoriõiguse seadus (valid since 12.12.1992). RT I 1992, 49, 615; RT I, 29.10.2014, 2 (in Estonian). Unofficial translation available at https://www.riigiteataja.ee/en/eli/531102014005/consolide (12.07.2015);

[EUPL] European Union Public License. V. 1.1. Available at https://joinup.ec.europa.eu/sites/default/files/eupl1.1.-licence-en_0.pdf (24.11.2015);

[GPL] GNU General Public License. V. 3. Available at http://www.gnu.org/licenses/gpl-3.0.en.html (24.11.2015);

[Information Society Directive 2001] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society. - OJ L 167, 22.6.2001, p. 10-19;

[Kelli 2015] Aleksei Kelli (2015). The Conceptual Bases for Codifying Estonia's IP Law and the Main Legislative Changes: From the Comparative Approach to Embedding Drafted Law into the Socio-Economic Context. – International Comparative Jurisprudence 1 (1), 44-54. Available at http://www.sciencedirect.com/science/article/pii/S2351667415000050 (24.11.2015);

[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);

[Oksanen et al. 2010] Ville Oksanen, Krister Lindén, Hanna Westerlund (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. Available at https://helda.helsinki.fi/handle/10138/29359 (18.11.2015);

[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/AutÕS-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.

[WCT 1996] WIPO Copyright Treaty. Adopted in Geneva on December 20, 1996. Available at http://www.wipo.int/wipolex/en/treaties/text.jsp?file_id=295166 (20.11.2015).