Originally published in: Fišer, Darja/Witt, Andreas (Eds.): CLARIN. The infrastructure for language resources.
Berlin/Boston: de Gruyter, 2022. Pp. – 457-479. (Digital Linguistics 1)
DOI: *https://doi.org/10.1515/9783110767377*

Paweł Kamocki, Aleksei Kelli, and Krister Lindén

The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN Infrastructure

Ville Oksanen, in Memoriam (26 December 1976–23 November 2014)

Abstract: The normative layer of CLARIN is, alongside the organizational and technical layers, an essential part of the infrastructure. It consists of the regulatory framework (statutory law, case law, authoritative guidelines, etc.), the contractual framework (licenses, terms of service, etc.), and ethical norms. Navigating the normative layer requires expertise, experience, and qualified effort. In order to advise the Board of Directors, a standing committee dedicated to legal and ethical issues, the CLIC, was created. Since its establishment in 2012, the CLIC has made considerable efforts to provide not only the BoD but also the general public with information and guidance. It has published many articles (both in proceedings of CLARIN conferences and in its own White Paper Series) and developed several LegalTech tools. It also runs a Legal Information Platform, where accessible information on various issues affecting language resources can be found.

Keywords: CLIC, copyright, license, ethics

1 Introduction

CLARIN, just like any research infrastructure (and most infrastructures for that matter), is a shared (or *common*) resource: instead of being privately owned, it is managed by a community for the benefit of all its members, and sometimes even of the general public. Experience teaches us that in the case of shared resources, the ideal of peaceful and sustainable exploitation is particularly diffi-

Paweł Kamocki, Leibniz-Institut für Deutsche Sprache, Mannheim, Germany, e-mail: kamocki@ids-mannheim.de
Aleksei Kelli, University of Tartu, Tartu, Estonia, e-mail: aleksei.kelli@ut.ee
Krister Lindén, University of Helsinki, Helsinki, Finland, e-mail: krister.linden@helsinki.fi

https://doi.org/10.1515/9783110767377-018

cult to achieve in practice. Hardin (1968) argued that such resources are doomed to fail. To illustrate the inevitable "tragedy of the commons" (overpopulation and the resulting depletion of natural resources), Hardin quoted the example from a little-known 19th-century pamphlet (Lloyd 1980) of an overgrazed common pasture: by making an economically justifiable decision to increase the number of their cattle, individual herders cause damage to the community. This theory was criticized (if not disproven) by a 2009 Nobel prize laureate, Elinor Ostrom, who formulated a list of design principles for sustainable self-organized governance of common resources. All these principles circle around clear, effective, and enforceable community rules regulating the use of the resource (e.g., mechanisms for exclusion of non-participants, sanctions for violations, and mechanisms of conflict resolution).

Both Hardin's and Ostrom focused on natural resources, many of which are indeed in imminent danger of over-exploitation; some could argue that there is no such danger for digital and intellectual resources, which do not depreciate with use, and which can be identically reproduced at little to no cost. However, deterioration is also a menace for shared digital resources. Wikipedia, probably the most successful digital commons so far, can be quoted as an example. Wikipedia owes its success partly to its large community respecting some fundamental principles (e.g., neutrality, freedom, mutual respect) summarized in the Five Pillars, and partly to the use of "copyleft" licenses, mostly CC BY-SA, which require that any modifications be released under the same license – a condition that can, if necessary, be enforced in court. It is easy to imagine that without either of these elements, Wikipedia could quickly fork into a privately-owned commercial product, or become useless due to low quality of its content (caused by e.g., overuse of editing rights), or simply be deserted by its users and gradually forgotten. The same can happen to a digital infrastructure like CLARIN.

In addition to this, CLARIN also faces another challenge which relates to the restricted use of language resources affected by third-party rights (intellectual property and personal data protection); we can call it "the tragedy of the anticommons". In the literature, the tragedy of the anticommons is explained as follows: "The anticommons thesis is that simple: when too many people own pieces of one thing, nobody can use it. Usually, private ownership creates wealth. But too much ownership has the opposite effect – it leads to resource underuse in an anticommons" (Heller 2013: 7). It has been suggested that "Fixing anticommons tragedy is a key challenge for our time" (Heller 2013: 6).

Fortunately, efforts are continuously being made to guarantee the sustainability of CLARIN and to protect it from the tragedies of both the commons and the anticommons. These efforts take many forms: updating the technical side of the infrastructure to keep it state-of-the-art, carefully defining long-term strategic

goals, or building a strong community. But no such effort would be sufficient without the legal and normative frameworks of the infrastructure, which are needed to address the use restrictions of language data and support the dissemination of language resources.

It has been argued that law and legal norms on their own are never sufficient; despite burglary being a criminal offence punishable by imprisonment, people still lock their homes, and many are willing to invest in state-of-the-art locks and alarm systems. But the reverse is also true: no lock and no alarm system would be useful if burglary was not punished by law, and burglars could spend hours in broad daylight trying to work around them. The legal (or ethical) rule comes first, and the technical and organizational solutions that safeguard it or put it into action come second. It is therefore fair to say that without the Normative Layer, neither the technical nor the organizational structure of CLARIN would exist.

This chapter is structured as follows: Section 2 introduces and defines the notion of the Normative Layer of CLARIN, which consists of three components: the Regulatory Framework, the Contractual Framework, and Ethical Norms. Section 3 presents the CLARIN Committee for Legal and Ethical Issues (CLIC), its history, structure, missions, and tasks. Section 4 then discusses the CLIC's actions related to the Regulatory Framework, and Section 5 those related to the Contractual Framework.

2 The normative layer of CLARIN

Legal norms (i.e., to put it simply, state-enforceable rules meant to regulate people's behaviour) that regulate the functioning of CLARIN can stem from external (objective) or internal (subjective) sources. We will refer to the norms stemming from the former as the Regulatory Framework, and to the latter as the Contractual Framework. These two frameworks form what we jointly refer to as the Normative Layer.

2.1 The regulatory framework

The most important part of the Regulatory Framework is statutory law, that is, acts passed by legislative bodies such as national parliaments (for national law) or by EU institutions (for EU law). From the perspective of CLARIN, the main legal challenges can be divided into two groups: intellectual property (chiefly copyright and related rights) and personal data protection. Indeed, language data

potentially contains copyright-protected content (written or oral works), subject matter of related rights (performances, parts of phonograms, and databases), or personal data. These fields of law are harmonized at EU level. Therefore, the most important legal acts for CLARIN include, for example, several copyright directives, the Database Directive, the Open Data Directive, or the General Data Protection Regulation (GDPR), as well as a plethora of associated national laws in every CLARIN member country. In addition to this, the Regulation 723/2009 on the community legal framework for a European Research Infrastructure Consortium (ERIC) is fundamental for the functioning of CLARIN ERIC; this Regulation was the basis for the European Commission's Decision of 29 February 2012, setting up the CLARIN ERIC consortium.

Another objective source of legal norms (i.e., another part of the Regulatory Framework) are court decisions, especially those emanating from the highest courts, which often – *de facto* or *de jure*, depending on the legal system – also apply beyond the facts of a specific case and provide a binding interpretation of statutes or even fill some grey areas in statutory law. For CLARIN as an entity, the most important court decisions are undeniably those emanating from the Court of Justice of the European Union (CJEU).

Finally, some highly authoritative guidelines, although not binding *de jure*, can also be regarded as an objective source of legal norms, and therefore part of the Regulatory Framework. This is especially the case of guidelines emanating from the European Data Protection Board (EDPB) and its direct predecessor, the Article 29 Data Protection Working Party. The EDPB is an independent advisory body made up of representatives of Data Protection Authorities from every Member State of the European Economic Area. Its guidelines are very likely to be followed by national courts and administrative bodies and provide a *de facto* binding interpretation of the GDPR.

The Regulatory Framework form a "legal exoskeleton", independent of CLARIN ERIC's will – that is, it cannot be altered by the sole decision of CLARIN bodies and they cannot opt out of it. Instead, it needs to be integrated and navigated in the decision-making process, as well as in the day-to-day functioning of the infrastructure. However, a powerful actor like CLARIN ERIC should not adopt a completely passive attitude toward the regulatory framework; it can also try to actively influence its future shape by participating in public consultations or by lobbying efforts.

The practical impact of the Regulatory Framework (in this case, the GDPR) on the compilation of language data is amply illustrated by Lindén et al. (2022).

2.2 The contractual framework

Unlike the Regulatory Framework, the Contractual Framework is internal to CLARIN, and CLARIN bodies can exercise proactive (i.e., not retroactive) control over it.

A part of this "legal endoskeleton" discussed in this chapter consists of contracts related to the everyday functioning of the CLARIN infrastructure, that is, contracts between CLARIN centres and providers of resources or tools (DELAs, Deposition License Agreements), and between CLARIN centres and end-users (EULAs, End-User License Agreements (for every user of a given resource) and ToS, Terms of Service (for all users of a repository).

In the spirit of Open Science and according to the FAIR principles, providers are encouraged to make their resources and tools open (i.e., available to anyone and for any purpose). This is best accomplished using standardized public licenses such as Creative Commons Attribution (CC BY) 4.0 (for datasets), or the General Public License (GPL) 3.0 (for software). Such licenses can be analysed as offers from the rights holder to the general public (hence the name "public license"); the actual contract is formed when a member of the public accepts the offer simply by using the content. In this model, there is no middleman (such as a distributor). To respond to the growing need of the community, public licenses are also incorporated in the CLARIN Contractual Framework, as parts of DELAs and EULAs.

Since these internal rules need to comply with the Regulatory Framework (or, to continue with the skeleton metaphor, the endoskeleton cannot extend beyond the exoskeleton), they need to be regularly revised and updated to adapt to the changes in the latter.

The practical impact of the Contractual Framework on the CLARIN infrastructure is illustrated, for example, by Hajič et al. (2022).

2.3 The role of ethical norms

Ethical norms are also part of the Normative Framework. Although they are not as such enforceable by the State, they indirectly shape both the Regulatory and the Contractual frameworks.

There seems to be no fixed content in ethics, as it changes very significantly, even over short periods of time. To an extent, the scientific community has developed its own ethical codes (Merton 1942).

3 The CLARIN Committee for Legal and Ethical Issues

As highlighted in the previous section, the Normative Layer of the CLARIN infrastructure is quite complex. Navigating thousands of pages of legal acts, court decisions, guidelines, and standard contracts requires considerable expertise. In order to advise the Board of Directors, the CLARIN Committee for Legal and Ethical Issues (hereinafter: the CLIC) was created.

3.1 History

The CLIC was formally established in 2012, during the first CLARIN Annual Conference in Sofia, Bulgaria (25–28 October 2012). However, even before that the CLIC existed informally (as the CLARIN Legal Issues Committee, hence the acronym). In its early days, Ville Oksanen and Krister Lindén from the University of Helsinki played the key role in the development of the CLIC.

Ville Oksanen (26 December 1976–23 November 2014) was a Finnish civic activist and lawyer known as a defender of civil rights and the freedom of expression, as well as a social debater. He defended his doctoral dissertation on digital copyright in 2008. In 2014, he was employed as a researcher at the Department of Computer Science at Aalto University. He was one of the founders of the Electronic Frontier Finland association and served as its president from 2004 to 2005, and was its vice-president at the time of his death. He wrote blogs for several computer magazines, and from a political point of view, Oksanen was a member of the Liberal Coalition Party. He was the Vice-Chairman of the Coalition Youth and was a deputy member of the Coalition Party Board from 1999 to 2000.

Research Director Krister Lindén, PhD in Language Technology and national coordinator of FIN-CLARIN, has a background in business, where he received hands-on training in legal issues as the first CEO of Lingsoft while negotiating with WordPerfect, Xerox, and Microsoft on including spellchecking and grammar-checking technology for all the Scandinavian languages and German in their products. He was the first chair of CLIC (2012–2015).

Erik Ketzan, a legal scholar from the Institute for the German Language in Mannheim, was the first co-chair. Since the very beginning it has been a tradition that the chair and co-chair are a language researcher and a legal scholar.

In 2016, Aleksei Kelli, professor of law from the University of Tartu in Estonia, took over as chair of the CLIC. Penny Labropoulou from Athena/ILSP served as co-chair.

In 2021, Paweł Kamocki, legal expert from the Institute for the German Language in Mannheim, holding both a PhD in law and a master's degree in linguistics, became chair of the CLIC. It is worth noting that he was among the original members of the CLIC when it was established in 2012, underpinning the CLIC's activities. Vanessa Hannesschläger, a digital humanities researcher from the Austrian Academy of Sciences, became co-chair.

3.2 Structure

The CLIC members are experts appointed by national consortia for two years, with a possible prolongation (Article 2 of the CLIC Bylaws). Every consortium is invited to appoint an expert, but there is no obligation to do so; as a result, several consortia are regretfully not represented in the CLIC. There is also a possibility for a consortium to appoint more than one expert – for example, Germany has always appointed two or three experts – however, only one expert per consortium (a "core member") has the right to vote. There is no formal requirement for appointed experts to be affiliated with a CLARIN centre. CLARIN observers ("emerging consortia") can appoint experts upon invitation from the Board of Directors. The Board of Directors can also appoint additional experts, or invite related initiatives to appoint representatives, but none of these powers have been used so far. The CLIC Chair can invite external experts to participate in CLIC meetings (without granting them membership). Members of the Board of Directors can also attend meetings of the CLIC. Traditionally, one of the Directors is delegated to serve as liaison between the Board and the Committee.

There are no formal requirements as to the level of expertise of CLIC experts, or as to their training. Despite this, the absence of candidates with legal training seems to be one possible reason why some CLARIN consortia are not represented in the CLIC. It should be emphasized that the CLIC is not a traditional legal "department" providing legal assistance but it is rather a legal competency centre interdisciplinarily integrating domains of legal and language research. Therefore, researchers whose background is not in law are also needed and valuable members. In practice, most CLIC members have long-standing experience in handling legal issues, acquired through management of language resources and tools. Some members of the CLIC are trained lawyers with experience in academia or in administration. The number of trained lawyers seems to have slowly but steadily increased since the establishment of the Committee, which is a very positive tendency, given the nature of the CLIC's missions. The CLIC is not intended to become a group composed exclusively of members with legal training, as this could move it far from the reality of language research and technology.

The CLIC Chair and Vice-Chair are appointed by the Board of Directors after consultation with the Committee, for two years. In practice, the Committee recommends the candidates in a vote, and the Board follows the recommendation. The same Chair and Vice-Chair can be appointed for more than one consecutive term; the bylaws do not limit the number of consecutive terms, but the practice seems to have limited it to two.

The CLIC meets at least once a year (often during the CLARIN Annual Conference). In practice, the Committee meets at least on a quarterly basis, and most of the meetings are virtual. Meetings are called by the Chair; at least three members of the CLIC may ask the Chair to call a meeting. Traditionally, CLIC meetings are open to non-members – physical meetings during the CLARIN Annual Conference can be attended by anyone, and the virtual meetings are announced on the CLARIN legal mailing list.

So far, the CLIC members have always been able to reach consensus on all debated matters, and there has been no need to vote. However, according to the by-laws, where consensus cannot be reached, the Committee should make decisions by simple majority vote, with casting vote held by the Chair.

Besides formal meetings, members of the CLIC are in regular contact via e-mail, usually in smaller groups, working on articles, updates of the CLIC materials or various other tasks. Apart from such informal subgroups, there is also a possibility to create formal subcommittees of at least three members, with the obligation to report on their activities to the Chair every year. Due to the relatively small and manageable size of the CLIC, and the fact that the Chair or Vice-Chair participate in every activity of the Committee, so far there has been no need to establish a formal subcommittee.

Neither membership nor chairmanship of the CLIC are remunerated, but they can be listed as contributions from national consortia to the CLARIN ERIC.

3.3 Mission and tasks

The main mission of the CLIC, as per Article 1 of its by-laws, is to "advise the Board of Directors on all issues related to [Intellectual Property], privacy and data protection and ethical matters, as well as legal issues related to access and dissemination policies and their implementation". This mission is particularly important if one takes into consideration the fact that the legal portfolio (unlike some other portfolios, like User Involvement, which is attributed to a member of the Board of Directors and a thematic committee) has never been expressly attributed to any member of the Board of Directors, therefore it can be assumed that it falls within the many competences of the Executive Director, who is likely

to need advice on legal matters. The scope of this mission is limited (for example, legal issues related to the establishment of new national consortia, or contractual relations between consortium partners, are not included), yet still very broad.

Firstly, it covers advice on all questions (also those unrelated to language resources) related to any form of intellectual property including, but not limited to, copyright, the *sui generis* database right, but also trade secrets, patents, and trademarks – areas that today remain underexplored by the CLIC and the language community in general, but which potentially may become important for the whole infrastructure.

Secondly, all questions related to privacy and data protection are also within the scope of the CLIC's mission. The distinction between privacy and data protection is fully justified, as privacy laws are not limited to data protection (privacy claims can be based on, e.g., tort law, criminal law, or specific grounds, such as Article 9 of the Napoleonic Code and the numerous legal norms that it inspired), and conversely data protection (i.e., a legal framework stemming mostly from the GDPR and the ePrivacy Directive) does not only apply to the private sphere of individuals' life.

Thirdly, the CLIC also advises the Board on all legal issues "related to access and dissemination policies and their implementation". Therefore, when it comes to policies concerning access to and dissemination of language resources and tools, the CLIC is also competent to advise the BoD on issues that go beyond intellectual property, privacy and data protection, and include for example, contract law, administrative law (such as reuse of public sector information) and even criminal law (e.g., hate speech in language resources).

Fourthly, advice on "ethical matters" of all sorts is also within the scope of the CLIC's mission. This should be interpreted as analysing compliance with commonly recognized norms of general and scientific ethics. Purely ethical issues are rarely debated within the CLIC, as it seems that they are better handled at the national or even institutional level.

In order to fulfil its mission, CLIC by-laws attribute the following tasks to the Committee:

- to prepare and publish analyses;
- to organize and participate in competency-building events;
- to collect, consolidate and prepare for publication in a single place various documents ("findings and recommendations") related to CLARIN activities;
- to maintain and adapt a set of licenses;
- to develop and implement procedures for the discussion and adoption of new recommendations for dealing with legal and ethical issues;
- to liaise closely with the Standing Committee for CLARIN Technical Centres;

- to ensure harmonization of legal and ethical policies between CLARIN ERIC and related initiatives;
- to publish and promote legal and ethical policies adopted by CLARIN;
- to follow the ongoing debates on legal and ethical issues at EU and national level, and to report on this to the BoD;
- to make an annual workplan; and
- to advise the Board of Directors in all legal and ethical issues [within the scope of its Mission].

3.4 Communication channels

The CLIC's tasks require regular communication within the Committee, as well as with other CLARIN bodies and the general public.

The most important communication channel within the CLIC are meetings – as stated above, there is one face-to-face meeting per year (unless, of course, travelling is made impossible, as it was during the Covid-19 pandemic), which is collocated with the CLARIN Annual Conference. The remaining meetings are virtual, using online communication technology. In between meetings, CLIC members usually work in smaller, task-oriented groups which communicate via e-mail or, occasionally, by videoconference.

The CLIC also communicates with the Board of Directors and the National Coordinators' Forum. The Board appoints a liaison who participates in CLIC's meetings. Once a year, the CLIC chair reports to the National Coordinator's Forum during one of their meetings. Communication with other CLARIN standing committees is less formalized; it normally takes place by e-mail (usually between chairs and/or co-chairs), but exceptionally also during face-to-face meetings.

The CLIC's primary channel for communicating with the general public is its dedicated webpage on the CLARIN ERIC's website,[1] administrated by the CLARIN Office. It contains an up-to-date list of CLIC members, a description of its mission and tasks, links to the description of the CLARIN licensing framework, the Legal Information Platform, the latest CLIC White Paper, as well as the address of the legal mailing list.

The Legal Information Platform is part of the CLARIN website administered directly by the CLIC. It contains detailed explanations on copyright and related rights, licensing and data protection, written by Paweł Kamocki and Erik Ketzan

1 https://www.clarin.eu/governance/legal-issues-committee

specifically with the CLARIN audience in mind. The platform also features a legal bibliography and links to useful online resources.

The mailing list legal@lists.clarin.eu is accessible to anyone, and is also used to communicate with all stakeholders within the CLARIN community and beyond. The CLIC also communicates its analyses to the general public via its White Paper Series, openly available online and licensed under the CC BY 4.0 license. The series was launched in 2014 on the initiative of Erik Ketzan, then the CLIC's vice-chair. The idea behind it was to present complex legal issues of fundamental importance for language science and language resources in a comprehensive yet concise form, approachable by language scientists with no legal training. The White Papers are not intended to reflect the official position of CLARIN ERIC, hence they do not use the CLARIN ERIC name or logo. In the future, the Board of Directors may intend to publish official statements on certain legal issues (such as future changes in EU law), in which case it will be the CLIC's role to provide the Board with advice.

On occasion, the CLIC, with financial and organizational assistance from the CLARIN Office, also organizes workshops and other events for other members of the CLARIN community or for the general public.

4 CLIC and the regulatory framework

The EU regulatory framework concerning access to and reuse of language resources, especially for research purposes, has undergone substantial modifications since the establishment of CLARIN ERIC.

In 2012, when the CLIC was officially established, copyright exceptions for research in EU Member States were based on Article 5.3(a) of the Directive 2001/29/CE on Copyright in the Information Society (InfoSoc). This exception, albeit quite broad, covers only non-commercial scientific research and teaching. Moreover, due to its non-mandatory character the exception was implemented very narrowly in most Member States, which made it quite incompatible with modern research practice, especially compared to the relative freedom offered by the fair use doctrine in the United States.

Shortly after the establishment of CLARIN ERIC, the first European countries adopted specific exceptions for text and data mining, still based on the same provision of the InfoSoc Directive, and therefore limited only to non-commercial scientific research. This was the case, for example, in the UK (in 2014), in France (in 2016), and in Germany (in 2017). In 2019, a mandatory exception for text and data mining for scientific research purposes was included in the Directive 2019/970 on

Copyright in the Digital Single Market (DSM). The new exception seems satisfyingly broad in enabling research institutions to copy copyright-protected material for scientific research purposes; however, it does not in itself allow any sharing of the copies (although it can be combined, as in German national law, with the "general" research exception in the InfoSoc Directive, which allows limited sharing), and requires for the copies to be stored with appropriate level of security, which increases the role of specialized research data archives.

The 2019 DSM Directive also introduces other changes that are relevant for language resources, such as, for example, extended collective licensing, or facilitated access to out-of-print works. It had to be implemented in all EU Member States by 7 June 2021.

The adoption of the DSM Directive was not the only important development in the field of copyright law in recent years. For the language community, another noteworthy document is the 2012 Orphan Works Directive (2012/28/EU), allowing for certain uses of some copyright-protected works whose rightsholders cannot be identified or located despite diligent search.

Another branch of EU law that affects language research and that has been thoroughly reformed since the establishment of CLARIN ERIC is data protection law. The General Data Protection Regulation (EU Regulation 2016/679; GDPR), which replaced the 1995 Data Protection Directive (95/26/EC), was adopted in 2016 and entered into application in 2018. Although the GDPR is typically described as a product of an evolution rather than a revolution (indeed, most of the fundamental concepts from the 1995 Directive remain unchanged), it does introduce some important changes for research, emphasizes the importance of accountability and self-assessment (e.g., through records of data processing activities, or through Data Protection Impact Assessments), and substantially increases fines for non-compliance. GDPR-related best practices in research are still crystallizing today, as awareness of data protection issues is growing not only in the research community, but also among research funding organizations.

Besides copyright and data protection, other branches of law are becoming increasingly important for access to and reuse of language data for research purposes. This is the case, for example, with EU rules on the reuse of Public Sector Information (PSI); the 2003 PSI Directive (2003/98/EC) was first amended in 2013 (by the Directive 2013/37/EU), and then replaced by the 2019 Open Data Directive (2019/1024), which also covers research data resulting from public funding. The extended scope of PSI/Open Data regulation created new sources of freely (and legally) reusable language data.

In recent months, the European Commission has been very actively publishing new drafts (e.g., for the Data Governance Act, or the Artificial Intelligence Act) which, when adopted, may have considerable impact on the CLARIN infrastruc-

ture. Therefore, the European Union is not at the "end of history" when it comes to legal developments, and there will be no shortage of work for the CLIC in the years to come.

4.1 CLIC's direct participation in the lawmaking process

CLARIN ERIC, a representative of the language research community in the EU, is an important stakeholder in many EU law reforms. As such, it becomes actively involved in stakeholder consultations which are necessary part of a democratic lawmaking process. This involvement requires not only time and qualified effort, but also substantial funds, and therefore it must remain limited. It is also a particularly delicate task, given that CLARIN consortia are financed by national governments which may have conflicting interests and take different positions in negotiations over various law reforms.

One of the earliest and perhaps most prominent examples of such involvement is the participation of Erik Ketzan, then vice-chair of CLARIN ERIC, in the thematic group "Text and Data Mining for Scientific Research Purposes" within Stakeholders Dialogue "Licenses for Europe", organized by the European Commission in 2013.[2] The goal of a long series of meetings (from February to December) was to find rapid, industry-led solutions to facilitate access to online content. Unfortunately, it was only very moderately successful, as most organizations representing the scientific research and open access publishing sectors withdrew from the process due to concerns about its scope, composition, and transparency.[3] CLARIN ERIC was one of the few representatives of the scientific research sector who did not leave the negotiation table until the end. The outcome of the TDM thematic group was a joint statement of commitment by scientific publishers to a roadmap to enable TDM for non-commercial scientific research in the European Union.[4] The roadmap envisioned by publishers was largely license- and subscription-based, and as such it was not widely acclaimed in the scientific research community. The apparent failure of the Stakeholders' Dialogue prompted the EU legislator to adopt a statutory exception for text and data mining for research purposes, which is currently part of the 2019 DSM Directive (Article 3).

Another example of CLARIN ERIC's (and the CLIC's) direct involvement in the lawmaking process is its admission, on the initiative of Ville Oksanen, as a

2 https://digital-strategy.ec.europa.eu/en/library/licences-europe-stakeholder-dialogue

3 https://libereurope.eu/article/stakeholders-representing-the-research-sector-smes-and-open-access-publishers-withdraw-from-licences-for-europe-2/

4 http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=49203

Category C observer (*Other Regional Intergovernmental Organization*, which is a "high" category, given precedence over national governments) at the World Intellectual Property Organization (WIPO). Established in 1967, WIPO is a specialized agency of the United Nations created to protect and promote intellectual property.

Participation in such endeavours is above all an opportunity to establish CLARIN as the representative of the language community in Europe, to accentuate its awareness of legal issues, and to liaise with other stakeholders with similar interests.

4.2 CLIC's research articles and White Papers

Writing research articles is a big part of CLIC's activity. For many years, the CLIC has been submitting a joint article (co-written by several CLIC members on the Chair's initiative) for the CLARIN Annual Conference, which often embraces a comparative approach to various legal issues affecting language science. Most of these articles deal primarily with the Regulatory Framework.

In 2015, an article by Aleksei Kelli, Kadri Vider, and Krister Lindén entitled "The Regulatory and Contractual Framework as an integral part of the CLARIN Infrastructure" (Kelli, Vider, and Lindén 2015) was accepted for the CLARIN conference in Wrocław. General in nature, the article provided a comprehensive overview of the legal framework applicable to language resources and introduced the concept of regulatory and contractual frameworks, which also serve as the backbone of this very chapter.

In 2018, a CLIC paper accepted for the CLARIN Annual Conference in Pisa (Kelli et al. 2019) examined the possibility of processing personal data without consent of the data subject for the development and use of language resources. The paper studied the implementation of research exceptions in various CLARIN countries, as well as the possibility to use alternative grounds (i.e., other than consent) for the processing of personal data for research purposes (such as public interest or legitimate interest).

In 2019, a CLIC paper accepted for the CLARIN Annual Conference in Leipzig (Kelli et al. 2020) explored the degree of legal control that copyright holders and data subjects can exercise over language models derived from "their" data.

CLIC White Paper #3, published in 2018 (Kamocki, Ketzan, and Wildgans 2018), was also devoted to a part of the Regulatory Framework, namely the GDPR. Over 25 pages long, the White Paper contains three parts. The first presents basic terminology and main principles of the Regulation, discusses the rights of data subjects and obligations of data controllers, and summarizes the principles related to cross-border transfers of personal data. The second part analyses

research exemptions in the GDPR, while the third presents new opportunities for bottom-up standardization, namely Codes of Conduct and Data Protection Seals. As of November 2021, the CLIC works on an extensive set of around 30 handouts on various GDPR-related issues, which are intended to replace the White Paper.

The idea of a GDPR Code of Conduct, first launched by Kamocki et al. (2018) and likewise discussed in the above-mentioned White Paper, was heavily debated within the CLIC. GDPR Codes of Conduct, regulated in Article 40 of the GDPR, are sets of sector-specific rules intended to contribute to proper application of the GDPR; if a Code of Conduct is approved by a competent supervisory authority, and if an accredited body monitors compliance with the Code, the Code can effectively "supplant" the GDPR for every organization who adheres to it. One could imagine a GDPR Code of Conduct for processing personal data in language resources, monitored by CLARIN, as a mechanism to unify GDPR-related practices in the community, and significantly facilitate cross-border endeavours such as building a pan-European infrastructure for language resources.

However, it emerged from the discussions within the CLIC that such far-reaching measures may not be desirable, given that certain CLARIN centres have already adopted specific data processing policies. It became apparent that many GDPR-related issues remain divisive in Europe: for example, researchers in countries like Germany or Austria are very attached to consent as the legal basis for personal data processing, whereas researchers in other countries, like Finland, rely more on alternative grounds such as public interest. In this context, the CLIC's ambition should be to adopt guidelines in several specific areas of GDPR compliance, such as data anonymization or rights of data subject, rather than opting for an instrument aimed at full unification.

4.3 CLIC's events

In recent years, the CLIC has also organized several events dedicated specifically to informing the language community about the relevant regulatory framework, and discussing legal challenges that language researchers have to face.

A workshop entitled "Hacking the GDPR to Conduct Research with Language Resources in Digital Humanities and Social Sciences", organized by the CLIC, hosted by CLARIN-LT, and supported by the CLARIN ERIC, took place in Vilnius on 7 December 2018.[5] The event, which brought together around 25 participants, featured presentations by several members of the CLIC, as well as invited guests.

5 https://www.clarin.eu/tags/clic

The use cases discussed during the event were focused on such GDPR-related concepts as suitable legal grounds for processing, data anonymization, pseudonymization, storage limitation, and appropriate safeguards.

A CLARIN café on the rights of data subjects in language resources, organized by the CLIC and supported by the TRIPLE project, took place on 30 March 2021.[6] The two-hour event was attended by 50 participants from both the CLARIN community and the private sector. The presentations given by CLIC members discussed not only the content of the rights of data subjects, but also practical aspects of their exercise in the specific context of language resources.

Another CLARIN café organized by the CLIC, this time on Text and Data Mining exceptions in the Directive on Copyright in the Digital Single Market, took place on 28 October 2021. The event attracted 25 participants willing to learn; it will hopefully mark the beginning of a community-wide debate on the impact of the recent EU copyright reform on language resources and language technology.

4.4 LegalTech co-developed by the CLIC and ELDAH

In 2020, three CLIC members (Vanessa Hannesschläger, Paweł Kamocki, and Walter Scholger), who are also members of the DARIAH ELDAH (Ethics and Legality in Digital Arts and Humanities) Working Group, teamed up to create the DARIAH ELDAH Consent Form Wizard.[7] This online tool enables researchers to quickly generate a GDPR-compliant consent form for collecting personal data for research purposes, but which can also be used, for example, for creating mailing lists or organizing academic events. Currently the tool is available in English, German, Italian and Croatian, although there are plans to have it translated to other languages. The launch of the Consent Wizard was an opportunity for the CLIC to liaise more closely with ELDAH, and organize the first joint meeting of both groups in June 2021.

5 CLIC and the contractual framework

The main role of the CLIC with regards to the contractual framework is to host and update the CLARIN license suite. It also prepared guidelines on the use of another popular license suite, Creative Commons 4.0, and two LegalTech tools intended to

6 https://www.clarin.eu/blog/recap-clarin-cafe-rights-data-subjects-language-resources
7 https://consent.dariah.eu/

assist researchers in the choice of an appropriate license for their tools and data. Recently, the CLIC has also started analysing other standard form contracts which govern access to and reuse of large quantities of language data, such as Terms of Service of popular social media services.

5.1 The development of the CLARIN Licensing Framework

At a meeting in Berlin in 2006, a handful of representatives of potential CLARIN members convened to prepare an EC-funded project application for the preparatory phase (PP) of CLARIN ERIC. Ideas were collected on what work package should be included and language technology and language resources were given favourites that every partner candidate was bidding for. However, Prof. Kimmo Koskenniemi from Finland insisted that legal and contractual issues also needed a work package, WP7. As no one else was eager to take on this task, it fell on Finland to carry out this part of the CLARIN PP project. The formation of the CLARIN ERIC statues had a separate work package WP8 and was carried out by Denmark under the direction of Prof. Bente Maegaard.

During the CLARIN preparatory phase project, two significant legal frameworks for the CLARIN operations were drawn up in WP7. The first framework was the CLARIN Service Provider Federation (SPF) which implemented the Single-Sign-On (SSO) principle on a large scale between the CLARIN consortium partners. It was a precursor to EduGain, although EduGain and EduRoam were already being developed at the time. The fact that one could use SSO – i.e., one's own university account and credentials – to sign in to various services was a key driving factor for the design of the second framework designed in WP7 (i.e., the licensing framework).

It was clear that open-source licensing and public licensing like Creative Commons (CC) were to be endorsed by CLARIN and named the PUB licensing category, but at the time CC was still in the process of establishing itself and most data was proprietary or had no clear license, having been painstakingly collected by individual researchers as, for example, manually transcribed and annotated letters, newspaper clips, or interviews, or just individual sentences from such sources. As many researchers had spent years, if not decades, of their life just collecting data, they were reluctant to give up such material to others, but some were willing to share on an individual basis. As mentioned, the datasets collected by researchers were also often personal data based on interviews, so the varying data protection legislations in the EU countries were an additional challenge. A restricted license category named RES was needed for such datasets. The idea

was that such resources should only be made available upon individual request and, if containing personal data, only for a limited time.

For political and practical reasons, the CLARIN PP Project Board thought that there should be some benefit to being an accredited researcher who was part of the CLARIN SPF, that is, the fact that a person already was an established researcher with credentials at a university should be recognized when accessing resources provided by CLARIN partners. For this reason, an academic license category was established called ACA. The nature of this category was intended as a public license to a limited public consisting of researchers, but therein lay a conundrum. Who was a legitimate researcher? Should only people from academia count or were people from industry acceptable as well (i.e., should the affiliation or the purpose limit access)? CLARIN opted for the practical solution: organizational status could be checked based on the login credentials. This provided a technological solution to the philosophical problem of restricting the category to academic researchers. Later, this seems to have caught on and will now be enshrined in the text and data mining exception in the DSM Directive, which grants a special status to research organizations, in particular by enabling them to store the copied data for verification purposes.

Both the RES and ACA categories were frowned upon by people from the hard-core Open Access Community, to whom only fully publicly available data was real data. They often had a technological background, where real data is measurement data produced by the research infrastructures themselves through measuring devices, and therefore having no copyright except for a potential *sui generis* database right on the data collection. In social sciences and humanities, interviews and questionnaires may be primary data, for which a license can be determined by the collector, but most data is secondary data, that is, it is used in research for some other purpose than it was originally created for by a human being imbuing it with either copyright or personal data rights.

Initially, the CLARIN categories were intended as metadata, that is, a way to inform the end user about the license to expect when accessing a resource. The idea was that licenses in a particular category had to provide at least a minimum number of rights to the end user, and at most some restrictions to qualify for the category. Originally this was intended only as a checklist to determine the category of the license. This is still visible in the CLARIN License Category Calculator. As Ville Oksanen had also been part of the origins of the Creative Commons (CC) movement, he adapted the CC categories. So as not to infringe on the CC look, it was decided that CLARIN would use a "+" (plus sign) as a connector between the subcategories where CC uses a "-" (hyphen) or a " " (space).

Based on an analysis of the manifold licensing conditions for resources in the Language Bank of Finland, Ville Oksanen devised a few more subcategories

in addition to those in CC. With this initial analysis as a basis, the categories and subcategories were tested on a set of more than 800 existing licenses throughout Europe by the CLARIN partners in EU (Oksanen, Lindén, and Westerlund 2010). Based on the response and the clarifying requests, the leading questions currently visible in the category calculator were designed. However, researchers wanted practical advice on how to make new or unpublished resources available, so what were originally only intended as categories and example clauses for classifying existing agreements evolved into ready-made sample contracts called CLARIN license templates. The final stage of the CLARIN licensing category adoption was to include the categories in the VLO as originally intended, using the laundry symbols to offer visual guidance on the openness of a resource. For an example of the use of CLARIN license templates in a repository, see Andersen and Gammeltoft (2022).

The work with the Contractual Framework continues in the CLIC. The majority of CLIC members are involved (see Kelli et al. 2018), and currently there is a preliminary plan to restructure the contractual framework in view of the GDPR.

5.2 CLIC's research articles and guidelines

The Contractual Framework was the subject of several joint CLIC articles. Apart from being discussed in the foundational paper (mentioned in Section 3.3) by Aleksei Kelli, Kadri Vider, and Krister Lindén (2015), it was also thoroughly and critically analysed in an article by Kelli and others (2018), first presented at the 2017 CLARIN Annual Conference in Budapest. The paper, entitled "Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes?", discusses the utility of CLARIN ACA and RES categories, and the possibility of replacing them with other requirements.

Kelli et al. (2021) also discussed some aspects of the Contractual Framework in their paper accepted for the 2020 CLARIN Annual Conference, entitled "CLARIN Contractual Framework for sharing language data: The perspective of personal data protection". The paper provided a preliminary analytical background for redesigning the CLARIN contracts to bring them up to speed with the GDPR.

Finally, a paper by Kamocki et al. (2021) presented at the 2021 CLARIN Annual Conference analysed another aspect of the Contractual Framework affecting language resources, namely the terms and policies of Twitter, an important source of language data. There are plans to extend the scope of this analysis to include other popular social media services in the next CLIC White Paper.

The very first CLIC White Paper (Kamocki and Ketzan 2014) was also dedicated to the contractual framework, namely the Creative Commons 4.0 license

suite; it was published in 2014, shortly after the launch of the latest version of Creative Commons licenses. Its intended purpose is to present the CC licenses and their building blocks to language researchers and discuss their utility for licensing language resources.

5.3 LegalTech developed by the CLIC

The first LegalTech tools created by the CLIC were developed to address the complexity of the contractual framework. The CLARIN License Category Calculator categorizes any resource license and aims at extracting an overview of the key licensing conditions, while the Public License Selector specializes in public licenses, guiding the user in choosing the best one for a particular purpose.

The CLARIN License Category Calculator[8] guides a resource depositor when choosing a license category for a language resource. The CLARIN classification system for licenses has been devised for more efficient and transparent management of language resources by providing an at-a-glance overview of the main usage conditions of a language resource. Based on their licenses, language resources compatible with the CLARIN infrastructure can be divided into three main categories: CLARIN PUB, CLARIN ACA, or CLARIN RES. In addition, there are several subcategories based on the most common conditions of use associated with the distribution of language resources. The CLARIN License Category Calculator guides depositors of language resources in selecting the most fitting license category for their resource based on a series of choices they make relating to its contents and intended use(s). CLARIN deposition license agreements (made between resource providers and the CLARIN centres) are available for curating a minimal set of access conditions to include a resource in the CLARIN PUB, ACA, or RES categories. The minimal deposition licenses can be used as checklists if a CLARIN Centre wishes to use its own set of deposition licenses to agree on additional usage conditions with the resource provider.

The Public License Selector[9] was created in 2015 by two CLIC members who also worked together on the EUDAT project: Paweł Kamocki and Pavel Straňák, assisted by software developer Michal Sedlák (Kamocki, Straňák, and Sedlák 2016; see also Hajič et al., 2022). A CLARIN mobility grant was allocated for Kamocki to travel to the Prague CLARIN Centre and finalize the tool. The Public License Selector is intended to assist a researcher in selecting a public license for his or

8 https://www.clarin.eu/content/clarin-license-category-calculator
9 https://github.com/ufal/public-license-selector

her datasets and tools. It covers popular data licenses (such as Creative Commons or Open Data Commons), as well as Free/Open Source Software licenses (GPL, BSD, MIT, Apache, etc.). The user is asked a series of simple, usually "yes/no" questions, with the answers serving to narrow down the available set of compatible licenses. The Public License Selector itself is released under an open license and has been widely reused both within and outside of the CLARIN community.

6 Conclusion

The Normative Layer of CLARIN is, alongside the organizational and the technical layers, an essential part of the infrastructure. It consists of the Regulatory Framework (statutory law, case law, authoritative guidelines, etc.) and the Contractual Framework (licenses, terms of service, etc.), and ethical norms. Navigating the normative layer requires expertise, experience, and qualified effort. In order to advise the Board of Directors, a standing committee dedicated to legal and ethical issues, the CLIC, was created.

Since its establishment in 2012, the CLIC has made considerable efforts to provide not only the BoD but also the general public with information and guidance. It has published many papers (both in proceedings of CLARIN conferences and in its own White Paper Series) and developed several LegalTech tools. It also runs a Legal Information Platform, where accessible information on various issues affecting language resources can be found.

Today, as CLARIN transitions from the development phase to the phase of sustainable functioning, the Normative Layer is changing dynamically, and continuous efforts from the CLIC are still needed to provide the Board of Directors and the whole community with information and guidance.

Bibliography

Andersen, Gisle & Peder Gammeltoft. 2022. The role of CLARIN in advancing work in terminology: The case of Termportalen – the national terminology portal for Norway. In Darja Fišer and Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Hajič, Jan, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko & Pavel Straňák. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer and Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Hardin, Garrett. 1968. The tragedy of the commons. *Science* 162 (3859). 1243–1248.

Heller, Michael. 2013. The tragedy of the anticommons: A concise introduction and lexicon. *The Modern Law Review* 76 (1). 6–25. https://scholarship.law.columbia.edu/cgi/viewcontent.cgi?article=2779&context=faculty_scholarship (accessed 17 May 2021).

Kamocki, Paweł, Vanessa Hannesschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén, Andrius Puksas. 2021. Legal Issues Related to the Use of Twitter Data in Language Research. *CLARIN Annual Conference 2021, Proceedings. 27–29 September 2021, Virtual Edition.* Utrecht: CLARIN. 150–153. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/10718/file/Kamocki_Hannesschlaeger_Legal_issues_2021.pdf (accessed 17 May 2022).

Kamocki, Paweł & Erik Ketzan. 2014. *Creative commons and language resources: General issues and what's new in CC 4.0* (CLIC White Paper Series 1). https://www.clarin-d.de/images/legal/CLIC_white_paper_1.pdf (accessed 17 May 2021).

Kamocki, Paweł, Erik Ketzan & Julia Wildgans. 2018. *Language resources and research under the General Data Protection Regulation* (CLIC White Paper Series 3). https://www.clarin.eu/sites/default/files/CLIC_White_Paper_3.pdf (accessed 21 May 2021).

Kamocki, Paweł, Erik Ketzan, Julia Wildgans & Andreas Witt. 2018. Toward a CLARIN data protection code of conduct. *CLARIN Annual Conference 2018, Proceedings. 8–10 October 2018, Pisa, Italy*. Utrecht: CLARIN. 49–52. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8189/file/Kamocki_Ketzan_Wildgans_Witt_Toward_a_CLARIN_Data_Protection_Code_2018.pdf (accessed 21 May 2021).

Kamocki, Paweł, Pavel Straňák & Michal Sedlák. 2016. The public license selector: Making open licensing easier. *International Conference on Language Resources and Evaluation (LREC)* 10. 2533–2538.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Paweł Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavrilidou & Pavel Straňák. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In Inguna Skadin & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2018: Pisa, 8–10 October 2018* (Linköping Electronic Conference Proceedings 159), 72–82. Linköping: Linköping University Electronic Press.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Paweł Kamocki, Arvi Tavast, Ramūnas Birštonas, Gaabriel Tavits, Mari Keskküla & Penny Labropoulou. 2021. CLARIN contractual framework for sharing language data: The perspective of personal data protection. In Costanza Navarretta & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2020: 5–7 October 2020*, 171–177. Utrecht: CLARIN.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Paweł Kamocki & Pavel Stranák. 2018. Implementation of an Open Science Policy in the context of management of CLARIN language resources: A need for changes? In *Selected papers from the CLARIN Annual Conference 2017: Budapest, 18–20 September 2017* (Linköping Electronic Conference Proceedings 147), 102–111. Linköping: Linköping University Electronic Press.

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Stranák & Jan Hajic. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. In Kiril Simov & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2019* (Linköping Electronic Conference Proceedings 172), 53–65. Linköping: Linköping University Electronic Press. https://ep.liu.se/ecp/159/008/ecp18159008.pdf (accessed 21 May 2021).

Kelli, Aleksei, Kadri Vider & Krister Lindén. 2015. The regulatory and contractual framework as an integral part of the CLARIN infrastructure. In Koenraad De Smedt (ed.), *Selected papers from the CLARIN Annual Conference 2015: October 14–16, 2015, Wroclaw, Poland* (Linköping Electronic Conference Proceedings 123), 13–24. Linköping: Linköping University Electronic Press. https://ep.liu.se/ecp/article.asp?issue=123&article=002 (accessed 17 May 2021).

Lindén, Krister, Tommi Jauhiainen, Mietta Lennes, Mikko Kurimo, Aleksi Rossi, Tommi Kurki & Olli Pitkänen. 2022. Donate Speech: Collecting and sharing a large-scale speech database for Social Sciences, Humanities and Artificial Intelligence research and innovation. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Lloyd, W. F. 1980. W. F. Lloyd on the checks to population. *Population and Development Review* 6 (3). 473–496. https://doi.org/10.2307/1972412.

Merton, Robert K. 1942. The normative structure of science. In Robert K. Merton (ed.), *The sociology of science: Theoretical and empirical investigations*, 267–278. Chicago: University of Chicago Press.

Oksanen, Ville, Krister Lindén & Hanna Westerlund. 2010. Laundry symbols and license management: Practical considerations for the distribution of LRs based on experiences from CLARIN. LREC 2010, Workshop on Language Resources: From storyboard to sustainability and LR lifecycle management. Valletta, Malta, 23 May 2010. https://helda.helsinki.fi//bitstream/handle/10138/29359/LREC2010.pdf?sequence=2 (accessed 17 May 2021).