As an example for the changes this replacement would have, we listed in Table 1 some of the CLARIN PUB licenses and their present and proposed classes; for the "open" category, we follow the recommendations of the Open definition.[15]

Academic use

In CLARIN, academic (ACA) and restricted (RES) resources are both restricted for copyright or personal data protection reasons[16]. Note that licenses for "academic use" are not unique to CLARIN (see, e.g., Academic Free License[17]).

The concept "academic use" is admittedly vague and can cause confusion. The first question that arises is whether commercial research is covered or not. If not, then one option could be to replace the academic category with non-commercial (NC). This solution is problematic as well, however, as there is community-wide confusion regarding what types of use are "non-commercial" (Kamocki and Ketzan 2014). This argument is further supported by findings from the VLO, where the condition of "non-commercial use" is found across all three license categories (PUB, ACA and RES) - cf. Section 3.4.

Another feature of the ACA category is that it poses a requirement on affiliation of the user to a recognised higher educational or research institution (i.e. AFFIL=EDU). This is a crucial issue since it requires private firms and non-profit organizations to acquire a "home-for-the-homeless-researcher" status for their researchers so that they can access data in the ACA category. The affiliation condition can be upheld using the Eduroam network[18] , which is the secure, world-wide roaming access service developed for the international research and education community, and can thus cover both educational and research institutes. However, not all institutions from all European countries are yet connected to Eduroam. If a user is not part of the Eduroam network, they can apply for researcher status, in which case they do not need to apply for access to the ACA-labeled resources separately. In CLARIN there is already a technical solution called "home-for-the-homeless" by providing an ID for those that need to acquire individual access rights to RES-labeled resources, but are not yet securely identified. A similar technical solution can be provided as a "home-for-the-homeless-researcher", which in addition should require some documentation that a person is engaged in academic research. RES-labeled resources still require individual permission, e.g. due to personal data legislation.

This dichotomy of ACA meaning Academic Use vs. Academic User is reflected in the CLARIN license templates: the CLARIN ACA EULA mentions "educational, teaching or research", while the CLARIN ACA Deposition License Agreement (DELA) specifies two additional conditions: "ID: A user needs to be authenticated or identified.", and "EDU: A user needs to be affiliated with the community of academic researchers through a university".[19] Either way, should CLARIN decide to keep the ACA category, this EULA and DELA should be either retired or made compatible. In the current state the depositor says they ask for the additional restrictions, but end-users are then not presented with those restrictions. This should of course be remedied in the EULA, although in practice end-users do not gain access unless they have been identified as having acknowledged researcher status.

It should also be pointed out that all ACA licenses are NORED, i.e. no redistribution. As researchers can anyway easily get access to the original point of distribution, there is no need to share ACA resources directly.

CLARIN is an exception, rather than the rule, in the use of an ACA license category. The license scheme by META, the Multilingual Europe Technology Alliance, contains no ACA category, but broadly distinguishes between commons (i.e. only for META-SHARE members) and restricted (with license categories for commercial/noncommercial, for-a-fee/not-for-a-fee, etc.).[20]The META-SHARE

[15] For additional information, see Guide to Open Licensing. Available at http://opendefinition.org/guide/ (7.3.2018).
[16] ACA can be seen as a type of restriction which in CLARIN was considered important enough to be "upgraded" into a category of its own (just like educational use for the rights statements or embargoed access in the COAR vocabulary).
[17] Additional information on the Academic Free License is available at https://opensource.org/licenses/AFL-3.0 (17.4.2017).
[18] Additional information available at https://www.eduroam.org/ (7.3.2018).
[19] CLARIN ACA EULA and CLARIN ACA DELA template are available at https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEULA and https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinSA (7.3.2018).
[20] Additional information available at http://www.meta-net.eu/meta-share/licenses (9.3.2018).

catalogues use both the unrestricted/restricted dichotomy and the conditions of use for faceted browsing. The research licensing portal from our colleagues in DARIAH, contains public licenses.[21]

There are also differences in the regulations for research and/or educational purposes included in European and national legislations. It should be clear that the ACA category is meant as an interpretation of the license accompanying a resource; it does not say anything about the legislation covering research in a given country. A resource collected based on some national research exception and distributed with an ACA label, puts the burden on a user in another country to check their own legislation to see if their intended plans are conformant with it.

3.3 Overview of categorization schemes and use of the CLARIN scheme in the VLO

The use of classification systems (e.g. types of resource, domain, provenance information) in order to organize resources contained in digital catalogues is a common practice that contributes to efficient search and retrieval. Facets created on the basis of these systems allow users to browse through the catalogues and restrict their search space using multiple filters.

Facets related to access and usage are found in most digital catalogues and are among the ones most frequently applied by users, demonstrating the importance of users knowing if and how they can access a resource and under what conditions they can use it for their purposes. Although there are currently various efforts for standardization of metadata, there is not yet a single, widely accepted classification system for access and usage. Still, most of cases fall under the following options (not necessarily excluding one another):

- classification based solely on the license of the resource (e.g. CC-BY, AGPL, etc.);
- grouping of the licenses into categories (e.g. "open access", "free for educational purposes", etc.) and organization of the resources on the basis of their licences; these categories are mutually exclusive, i.e. a resource can only be assigned to one category based on its license;
- analysis of the licenses based on the conditions of use they regulate (e.g. "attribution", "non-commercial use", "fee required" etc.) and linking of the resources with the conditions of use of their license; in this case, a resource can be linked to one or more conditions of use.

The latter two options are not meant to replace licenses, but to support users in their search through the appropriate deployment of formal metadata elements and values.

The choice of the values used for these two options largely depends on the intended audience. They are usually selected and formulated in a way that users can have a general understanding of what they can do with the resources and understand in a user-friendly way some basic notions of the legal text.

For illustration purposes, we will briefly present here two classification schemes that are relevant to our purposes and that could help us in our discussion:

- the COAR controlled vocabulary of access rights: COAR is the Confederation of Open Access Repositories and one of its activities relates to the development of controlled vocabularies for bibliographic metadata to ensure interoperability between the various repositories[22]; the access rights vocabulary[23] declares the degree of "openness" of a resource and has four values: *open access*, *restricted access*, *embargoed access* and *metadata access*. The last two values can be regarded as two types of specific restrictions (temporal restriction and content blockage), which are considered important enough to be promoted into values of their own.
- the rights statements that have emerged from a joint initiative of Europeana and the Digital Public Library of America[24]: these include 12 rights statements ('"high level summaries of the underlying rights status"') mainly intended for use by cultural heritage institutions. Two main features are used in the creation of these statements: the copyright status and the declaration of permission or prohibition of selected uses, mainly use for educational purposes and use in commercial applications, which are the ones most frequently associated with cultural objects

[21] Additional information available at http://forschungslizenzen.de/ (9.3.2018).
[22] For additional information, see COAR Vocabularies. Available at https://www.coar-repositories.org/activities/repository-interoperability/coar-vocabularies/ (7.3.2018).
[23] For additional information, see Controlled Vocabulary for Access Rights (Draft V1). Available at http://vocabularies.coar-repositories.org/documentation/access_rights/ (7.3.2018).
[24] For additional information, see RightsStatements.org. Available at http://rightsstatements.org/en/ (7.3.2018).

distributed via these institutions. What is also noteworthy is that there are specific statements for resources with unknown or doubtful copyright status, taking into account whether this has been investigated or not.

The CLARIN licensing categorization scheme is currently used for the facet "Availability" in the Virtual Language Observatory catalogue (VLO).[25] The VLO harvests metadata descriptions of language resources from CLARIN centres but also from any other source that uses the OAI-PMH harvesting protocol and has agreed to be harvested by CLARIN, such as the OLAC catalogue of language resources[26] () and EUROPEANA[27].

Given the fact that resources come from multiple sources that do not use the same metadata schema for describing them, there has been a mapping procedure to the CLARIN license categories from the original elements and values.[28] For resources whose metadata records included a metadata element for the license, the mapping was easy and straightforward. However, a large number of resource descriptions contained no element at all for licensing information or included a free text statement, such as "available for research", "please ask provider", "academic research only", etc.; where possible, mapping of these values to the license categories was decided. As a result, 509,971 resources have been tagged with one of the CLARIN labels, which amounts to around 31.5% of the total resources in the VLO.

Below are some statistics on how the categories have been used for different resources in the VLO as of January 7, 2018:

Categories of All Resources	Number of Resources	Category Distribution	Categorized
Public	269,673	52.3 %	31.5 %
Academic	138,740	27.2 %	
Restricted for individual	101,558	19.9 %	
Unspecified	1,109,949	–	68.5 %

Table 2. Overall distribution of license categories.

Comparing this with the distribution of resources containing Finnish:

Categories of Finnish Resources	Number of Resources	Category Distribution	Categorized
Public	24,437	96.4 %	98.9 %
Academic	51	0.2 %	
Restricted for individual	850	3.4 %	
Unspecified	275	–	1.1 %

Table 3. Distribution of license categories among Finnish resources.

Having examined the unspecified resources for Finnish, it seems that they have been mainly harvested from other sources than the Finnish CLARIN Centre and their metadata records have had no clearly specified license information.

For those who wish to analyse the license categories and subcategories, VLO already provides more advanced search options using keywords. Users can use “NC” as a search condition in VLO and get a list of “non-commercial” resources from all categories. Here are the figures as of January 7, 2018:

- Public (10,511)

[25] CLARIN Virtual Language Observatory. Available at https://vlo.clarin.eu/ (7.1.2018).
[26] For additional information, see Open Language Archives Community. Available at http://www.language-archives.org/ (7.3.2018).
[27] See https://www.clarin.eu/faq-page/275#t275n3923 for a list of providers to the CLARIN VLO.
[28] The mapping has been the output of co-ordinated work between the VLO development team, the metadata curation experts from the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences and the CLARIN Legal Issues Committee.

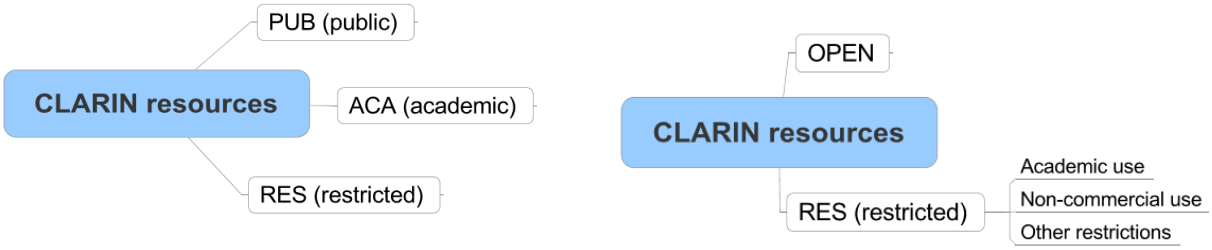
- Academic (70)
- Restricted for individual (959)
- Unspecified (580)

This serves as an example of why NC is not sufficient to describe Academic Use, as argued in Section 3.3, given that NC can be used in any category. Together, the figures also show that more than 100,000 Academic resources are available for academic activities also in a commercial setting, as it does not matter who or what entity produces the academic results. The key point is that when someone wishes to exploit the results commercially, they need to acquire the necessary rights. Note that currently ACA still comes with a need for the researcher accessing the data to be affiliated with an academic institution.

3.4 Alternative categorization

The question remains whether it would be reasonable to change the current categorization of resources. Considering the problems caused by license proliferation (e.g., the existence of conflicting clauses), it would, in theory, be preferable to rely on existing standard licenses (e.g. Creative Commons[29]) rather than create new bespoke licenses to replace them. The problem is that the use of language resources cannot easily be based on well-known standard licences due to the many unique situations we have described as well as to the existence of legacy resources with licences that cannot be replaced. Additional permission and restrictions are fundamentally required. An alternative categorization scheme should therefore be considered.

One option is to divide resources into two main categories: **open** and **restricted** as proposed by P. Kamocki at the CLARIN annual meeting in Wroclaw, 2015. This category scheme fits better, conceptually, with the open science doctrine, which is becoming increasingly supported and emphasized across the EU and globally. The transformation from PUB to Open would require moving some resources to a restricted category (when the license is not broad enough). The following diagrams exemplify the current and alternative categorization.



The existing categorization **The alternative categorization**

Figure 1. Existing and alternative categorization of licensing

Another alternative would be basically the same, but avoid using the Restricted label altogether. Reason for this would be that in a simple dichotomy of Open and the rest does not bring more information to the user, while labelling data with a not very positive, potentially even scary, label. In this proposed scenario, we would use one main category of Open, which is well defined: represented by established labels Open Access, Open Data and Open Source Software.

In addition to the Open category, we would also use a set of established labels for common license conditions and map them to existing licenses for quick user orientation. However, care must be taken to inform users to read the actual license[30], because "by" in CC-BY and for instance a copyright notice in MIT license are not exactly the same thing, just as the "SA" requirement as described in CC-BY-SA and in GPL-v2 differ. Still, for basic orientation it seems helpful to use these broad classification labels. We therefore propose to use established and easy to understand BY, NC, ND, SA labels from CC licenses, and possibly add some other that are used often in CLARIN[31]. If by reviewing current use of licenses in CLARIN we find that "research only" is a commonly used condition, we should add that label with some simple visualisation, as well.

[29] Additional information on Creative Commons is available at https://creativecommons.org/about/ (17.4.2017).
[30] Users must be warned to read the licence text in all cases; as mentioned earlier, licensing categories and indication of conditions of use serve only as hints for the end-user and in no way replace the licences.
[31] Work on gathering conditions of use linked to language resources, taking into account mainly CLARIN, META-SHARE and ELRA licenses has been initiated in the framework of the W3C Linked Data for Language Resource Community Group

4 Conclusion

As the open science doctrine becomes increasingly prevalent at national, regional and international levels, CLARIN's goals and policies should adapt to reflect this as it continues its mission of disseminating language resources as widely as possible.

Under its existing license category scheme, CLARIN resources are divided into three categories: public (PUB), academic (ACA), restricted (RES). This article analysed the existing categories and explored whether an alternative scheme, focusing on a division between "open" and "restricted", would be more compatible with open science and be more useful for the CLARIN community.

Due to the cooperation of several authors with divergent suggestions, we have not yet reached final conclusions. The common understanding is that we need to continue our analysis among the CLARIN Legal Issues Committee. The article also serves as an indication of legal discussions relevant to CLARIN to the larger CLARIN community, so that additional voices may contribute.

Before making any final choices, we recommend a user survey to investigate the CLARIN community satisfaction with the current license category scheme, how accustomed they already are to using it and their interaction with other classification systems from other repositories. The current implementation of the classification system in CLARIN centres should also be taken into account.

Reference

[Berlin Declaration on Open Access 2003] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities of 22 October 2003. Available at https://openaccess.mpg.de/Berlin-Declaration (16.4.2017);

[BOAI 10] Ten years on from the Budapest Open Access Initiative: setting the default to open (2012). Available at http://www.budapestopenaccessinitiative.org/boai-10-recommendations (15.4.2017);

[CLARIN] CLARIN. Licenses, Agreements, Legal Terms. Available at https://www.clarin.eu/content/licenses-agreements-legal-terms (15.4.2017);

[CLARIN license classification system] CLARIN. Licenses and the CLARIN license classification system. Available at https://www.clarin.eu/content/licenses-and-clarin-license-classification-system (7.3.2018);

[CLARIN Value Proposition 2016] CLARIN Value Proposition (2016). Available at https://office.clarin.eu/v/CE-2016-0847-CLARINPLUS-D5_4.pdf (12.4.2017);

[Estonian Copyright Act] Autoriõiguse seadus (valid since 12.12.1992). RT I 1992, 49, 615; RT I, 31.12.2016, 2 (in Estonian). Translation available at https://www.riigiteataja.ee/en/eli/524012017001/consolide (17.4.2017);

[European Commission] European Commission (2016). Open innovation, open science, open to the world – a vision for Europe. Available at https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world-vision-europe (12.1.2018);

[FSF] Free Software Foundation. What is free software? Available at https://www.gnu.org/philosophy/free-sw.html#mission-statement (13.1.2018);

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (12.1.2018);

[Gezelter 2009] Dan Gezelter (2009) "What, exactly, is Open Science?", The Open Science Project. Available at: http://openscience.org/what-exactly-is-open-science/ (7.3.2018).

[Kamocki and Ketzan 2014] Paweł Kamocki and Erik Ketzan (2014) Creative Commons and Language Resources: General Issues and What's New in CC 4.0 (May 2014). Available at https://www.clarin.eu/content/legal-information-platform (12.1.2018);

[Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press,

(cf. http://www.cosasbuenas.es/static/ms-rights/ for the MS-rights ontology, which can be seen as an extension of the ODRL model). (Rodriguez and Labropoulou 2015).

Linköpings universitet, 13−24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (14.4.2017);

[OECD 2015] OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. Available at http://dx.doi.org/10.1787/5jrs2f963zs1-en (12.1.2018);

[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 (13.4.2017);

[Open Knowledge International] Open Knowledge International. What is open? Available at https://okfn.org/opendata/ (15.4.2017);

[Open Definition 2.1] Open Knowledge International. Open Definition 2.1 Available at http://opendefinition.org/od/2.1/en/ (15.4.2017);

[Open Source Initiative 2007] Open Source Initiative. The Open Source Definition. Available at https://opensource.org/osd (13.1.2018);

[Rodriguez and Labropoulou 2015] Victor Rodriguez-Doncel & Penny Labropoulou (2015). Digital Representation of Rights for Language Resources. In Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), ACL-IJCNLP 2015, pages 49 - 58