Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes?

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

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

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

Penny Labropoulou
R.C. Athena/ILSP
Epidavrou & Artemidos
151 25 Maroussi
Greece
penny@ilsp.gr

Erik Ketzan
Birkbeck, University of London
43 Gordon Square
WC1H 0PD London
United Kingdom
Eketza01@mail.bbk.ac.uk

Abstract

The article introduces a preliminary investigation on the compatibility of the current CLARIN license categorization scheme vis-à-vis the open science paradigm. To this end, the first section presents the main definitions and requirements attached to "openness" in related science, research and distribution/access frameworks. The next section focuses on the current tripartite licensing scheme, which consists in the distinction of resources into PUB (public), ACA (academic) and RES (restricted), and the rationale and requirements associated with these categories. The question as to whether it is fit to serve open policies is brought forward. Finally, alternative categorization schemes are suggested, and critically evaluated with arguments in favour and against them. The article intends to open up the discussion for a reformed scheme inviting further work in the area.

Keywords: open science, CLARIN license categories

1 Introduction[1]

The aim of this paper is to explore how, and if, the existing CLARIN license categories (PUB, public; ACA, academic; and RES, restricted) should be kept, modified, or replaced to further the goals of CLARIN's open science policy. This paper will be used as a starting point in evaluating how compatible open science requirements are with the way CLARIN manages language resources[2]. In addition, this paper could support the development of the CLARIN open science policy itself.

In the first section, the authors outline various definitions of open science. In the second section, we explore the compatibility of the CLARIN framework for management of language resources with the requirements of open science. In the third section, a new categorization model for language resources is tentatively introduced and discussed.

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

[2] CLARIN deposition license agreements (https://www.clarin.eu/content/licenses-agreements-legal-terms) define language resources as "material owned by the Copyright holder as defined in this Agreement, including software, applications and/or databases". In this article the terms 'CLARIN language resources', 'CLARIN resources', 'language resources' and 'resources' are used as synonyms.

2 Open science definitions

CLARIN has expressed its commitment to **open science** (see, CLARIN Value Proposition 2016). This commitment, however, requires additional clarification. To evaluate whether CLARIN follows open science requirements while managing language resources, it is necessary to define open science.

The European Commission (2016) and OECD (2015) provide rather general definitions of open science. We therefore turn to more operational definitions.

There are several initiatives which provide criteria for open science. The Berlin Declaration on **Open Access** (2003) requires that 1) open access should cover research results, raw data and metadata, source materials, digital pictorial and graphical materials, etc.; 2) right holders grant to all users a license to use, distribute, and to make and distribute derivative works; 3) a complete version of the work and all supplemental materials in an appropriate standard electronic format is deposited.

The policy document titled "Ten years on from the Budapest Open Access Initiative: setting the default to open" (BOAI 2012) has some specific requirements for licensing and reuse (e.g. a recommendation to use CC-BY or an equivalent license).

The Open Knowledge International[3] sets the following key features of **openness**: 1) availability and access; 2) reuse and redistribution; 3) universal participation.

Open Knowledge International has also adopted the **Open Definition** (version 2.1), which has detailed conditions for the determination of open works and open licenses. It essentially allows conditions such as attribution, integrity, and share-alike. If there are additional restrictions on re-use of the data (e.g. non-commercial use, no derivatives), then the content is not open.

The director of OpenScience project (which is dedicated to writing and releasing free and open source scientific software) defines[4] open science through four fundamental goals: 1) transparency in experimental methodology, observation, and collection of data; 2) public availability and reusability of scientific data; 3) public accessibility and transparency of scientific communication; 4) using web-based tools to facilitate scientific collaboration.

3 Re-thinking the CLARIN framework for management of language resources

The management of CLARIN language resources is based on a tripartite division of resources: PUB (public), ACA (academic), RES (restricted).[5] Currently, 364,448 language resources have been tagged with one of these categories[6] within the Virtual Language Observatory catalogue.[7] Although the categorization scheme has been incrementally improved (see Kelli et al. 2015), the conceptual framework remains the same.

This division does not aim to replace licenses but to group them together in a way that supports end-users in restricting the search area for resources they can deploy for their purposes. License categories give them a first indication of the licensing terms of the resources.

The researchers who created the license categories of CLARIN resources (Oksanen et. al. 2010) provided arguments to explain their choices. First, the categorization was based on an extensive survey. Secondly, it is argued that the licensing categorization must take into account licensing terms, such as limiting the distribution to academia or to even more limited groups of users, that are not covered bystandard licenses such as Creative Commons[8] but which are commonly used for language resources.

The three categories are defined through specific requirements:

[3] Open Knowledge International is a global non-profit organisation focused on realising open data's value to society. Information available at https://okfn.org/ (15.4.2017).

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

[5] The tripartite division is not unique. For instance, ORCID also has three levels for access to data: 1) everyone; 2) trusted parties; 3) only me. Additional information available at http://support.orcid.org/knowledgebase/articles/124518-orcid-privacy-settings (17.4.2017).

[6] 364,448 records out of 893,368 have been labelled Public (167,900), Academic (138,905) and Restricted for individual use (57,643).

[7] CLARIN Virtual Language Observatory. Available at: https://vlo.clarin.eu/ (3.7.2017).

[8] https://creativecommons.org/

- PUB resources should have no use limitations (e.g. based on geographic location, purpose of use, etc.). Recommended licenses are the Creative Commons Zero (CC0) or the Open Database License[9] (ODbL).
- ACA resources must be available for studying, research and teaching purposes.
- The availability of RES resources is even more limited. Their use requires following specific ethical or personal data protection requirements (Oksanen et. al. 2010).

PUB, ACA, RES categories may also be subject to additional conditions such as non-commercial use (NC), non-derivative use (ND) and to redeposit modified resources with CLARIN (RED) (Oksanen et al. 2010). The question is whether the division of resources into PUB, ACA and RES category scheme can be improved in light of an open science policy.

3.1 Public vs. Open

"Public" and "open" are to some extent competing concepts. It could be argued that the concepts "open" and "openness" are clearer and more widespread than the concept of "public". However, there is no official and universal definition for the term "open", as demonstrated above. Openness is used in different policy documents (Berlin declaration, BOAI 2012, etc.), by different institutions (OECD, EU), and organizations (Open Knowledge International, Open Source Initiative). The names of some standard license also include the term "open" (e.g. Open Database License). This, however, is not the only practice for naming licenses.

The concept of "public" is used in a similar context as well. For instance, "public domain" refers to material which is no longer protected by copyright. Copyright legislation uses the term "public". For instance, the InfoSoc Directive (2001) regulates communication to the public (art. 3).[10] Several well-known standard licenses such as the European Union Public License[11] (EUPL), GNU General Public License[12] (GPL), Eclipse Public License[13] (EPL) and Mozilla Public License[14] (MPL) use the term "public" in their title. There is also the term "free" used in the title of several standard licenses (e.g., Academic Free License[15], Free Public License[16]).

There is no consistent use of the terms "open" and "public". Replacing one term with the other probably does not make the situation better. Although the concept of "public" can be limited, not necessarily covering the general public (everyone), replacing it with the concept of "open" would not solve these definitional problems.

One outcome of replacing "public" with "open" is that many resources would technically be classified as "restricted", in contrast to "open". This is a more general sense of restricted than what is intended by CLARIN RES, which is defined in contrast to public or academic.

3.2 Academic use

In CLARIN, academic (ACA) and restricted (RES) resources are both restricted for copyright or personal data protection reasons. 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 could cause

[9] For additional information, see Open Data Commons Open Database License (ODbL). Available at https://opendatacommons.org/licenses/odbl/1.0/ (3.7.2017).

[10] The Estonian Copyright Act defines the public as "means an unspecified set of persons outside the family and immediate circle of acquaintances" (§ 8). This approach is more or less similar across Europe.

[11] Additional information on EUPL is available at https://joinup.ec.europa.eu/community/eupl/og_page/european-union-public-licence-eupl-v11 (17.4.2017).

[12] Additional information on GPL is available at https://www.gnu.org/licenses/gpl.html (17.4.2017).

[13] Additional information on EPL is available at https://www.eclipse.org/legal/epl-v10.html (17.4.2017).

[14] Additional information on MPL is available at https://www.mozilla.org/en-US/MPL/ (17.4.2017).

[15] Additional information on Academic Free License is available at https://opensource.org/licenses/AFL-3.0 (17.4.2017).

[16] Additional information on Free Public License is available at https://opensource.org/licenses/FPL-1.0.0 (17.4.2017).

[17] Additional information on Academic Free License is available at https://opensource.org/licenses/AFL-3.0 (17.4.2017).

confusion. The primary question 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). Additionally, commercial research should under certain conditions be allowed under academic use, e.g. in an industry-sponsored academic setting.

3.3 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[18]) rather than create new bespoke licenses to replace them. The problem is that the use of language resources, due to the many unique situations we have described, cannot easily be based only on well-known standard licenses. 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**. This category scheme fits better, conceptually, with the open science doctrine, which is becoming increasingly supported and emphasized across the EU and globally. However, renaming PUB resources to Open would not improve legal clarity. The transformation from PUB to Open would require moving some resources to a restricted category (when the license is not broad enough). The current system has largely worked well and has been in place for many years. A new license categorization scheme could cause confusion within the CLARIN community.

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, academic, restricted. This article 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. Future work is needed to refine such a proposal and argue, quantitatively and qualitatively, for its rejection or adoption.

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 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);

[18] Additional information on Creative Commons is available at https://creativecommons.org/about/ (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 (15.4.2017);

[InfoSoc Directive 2001] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. Official Journal L 167 , 22/06/2001 P. 0010 – 0019. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32001L0029&qid=1492406578166&from=en (17.4.2017);

[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

[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, 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 (14.4.2017);

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