The Legal and Ethical issues Workshop @LREC2022, pages 22–26
Marseille, 20 June 2022

Categorizing Legal Features in a Metadata-Oriented Task: Defining the Conditions of Use

Mickaël Rigault[1], Victoria Arranz[1], Valérie Mapelli[1], Penny Labropoulou[2], Stelios Piperidis[2]
[1]ELDA/ELRA,
[1]9 rue des Cordelières, 75013 Paris, France, [2]Institute for Language and Speech Processing
[2]Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Athens Greece
mickael@elda.org, arranz@elda.org, mapelli@elda.org, penny@athenarc.gr, spip@athenarc.gr

Abstract

In recent times, more attention has been brought by the Human Language Technology (HLT) community to the legal framework required to render Language Resources (LR) and tools available for later use. Licensing is now an issue that is foreseen in most research projects and that is essential to provide legal certainty for repositories when distributing resources. Some repositories such as Zenodo or Quantum Stat do not offer the possibility to search for resources by licenses which can turn the searching for relevant resources into a very complex task. Other repositories such as Hugging Face propose a search feature by license which may make it difficult to figure out what use can be made of such resources.

During the European Language Grid (ELG) project, we moved a step forward to link metadata with the terms and conditions of use. In this paper, we document the process we undertook to categorize legal features of licenses listed in the SPDX license list[1] and widely used in the HLT community as well as those licenses used within the ELG platform.

Keywords: Copyright, Open-Source Licenses, Licensing, Metadata

1. Introduction

Nowadays, the number of licenses that exist to define the framework of use of tools and Language Resources (LRs) in the field of Human Language Technologies (HLT) is tremendously high. There are several widely known license suites available for research teams to make their content available (Creative Commons[2], MIT[3], ELRA[4], META-NET[5], CLARIN[6], BSD[7]…). Therefore, it is increasingly difficult for researchers and potential users to have clear information on the terms and conditions of use of a particular resource. Therefore, repositories transcribe legal concepts into metadata information to allow for the display of legal information to users and thus allow both a) to know what can be done with a resource at first glance and b) the implementation of search functions within catalogues of resources for popular conditions of reuse. A thorough study was initiated within the Meta-Share project (Piperidis, 2012; Piperidis et al., 2014) to highlight licenses and related concepts that apply to HLT tools and LRs (Choukri et al. 2012,). ELDA also built a License Wizard[8] that enables users to select licenses depending on the legal metadata used as search criteria.

Following upon Meta-Share, the European Language Grid (ELG)[9], a project funded by the European Union, has developed a platform to enable access and use of HLT tools and LRs.

To support the ELG platform (Rehm 2020), the project team developed a metadata schema (Labropoulou et al. 2020) for the description of Language Resources and Technologies (LRTs). For the free text search and faceted view, the ELG platform uses a subset of the metadata elements deemed important for discovery by the users. Findability[10] is a crucial feature in the lifecycle of an LRT.

In this paper we relate the research that we performed to power this search engine with legal metadata features.

For this purpose, we identified general legal concepts and transcribed those into metadata values, we cross-checked a list of licenses through the lens of these general concepts and categorized these licenses according to their conditions of use and the corresponding metadata values.

2. License Framework

The main purpose of this task was to define legal categories and add them to the ELG metadata scheme[11]. This work was done through a thorough investigation of the licenses available on the SPDX license list[12] and those used for LRTs already included in the ELG platform, which provides a list of commonly used licenses in the open-source community. All the different aspects analyzed and addressed are described in the coming sections.

[1] https://spdx.org/licenses/
[2] https://creativecommons.org/about/cclicenses/
[3] https://opensource.org/licenses/MIT
[4] http://www.elra.info/en/services-around-lrs/distribution/licensing/
[5] http://www.meta-net.eu/
[6] https://www.clarin.eu/content/licenses-and-clarin-categories
[7] https://opensource.org/licenses/BSD-3-Clause
[8] http://wizard.elda.org/
[9] https://www.european-language-grid.eu/
[10] Please refer to (Wilkinson et al., 2016) for the FAIR Principles.
[11] https://european-language-grid.readthedocs.io/en/stable/all/A2_Metadata/Metadata.html
[12] https://spdx.org/licenses/

3. Licensed Rights

In the general theory of copyright, the set of rights granted by the law aims to foster innovation and protect creators with respect to their original works. Moreover, the law allows copyright owners to deal freely with the rights they own. Generally, this can be done through "proprietary licenses" where copyright owners or allowed licensees keep control over who has the right to use the set of rights to the underlying original works.

However, in recent years, under the influence of the open-source movement, specific licenses were designed so that creators could allow redistribution and reuse of their works' contents with fewer restrictions

In the following sub-sections we will detail the set of rights that may be granted by those licenses. It should be noted that we tried our best to generalize legal concepts found in licenses that may not be expressed with the same terms in all licenses and/or may have differences in the semantic nuances and presentation in the texts (Rodriguez-Doncel and Labropoulou 2015).

3.1 Right to Reuse

Copyright protection prevents third parties from reusing the intellectual property to create copies of the original work and create derivative works or products based on the original.

During our investigation we found out that some of the licenses that allow open access to their content, widely called "open-source" licenses, provide or imply that the licensor grants licensees the right to reuse the content of the protected works for their own use. This right to reuse will also help us imply some further reuse possibilities down the line when we will deal with the items linked to restrictions and conditions attached to reuse in Section 5.

3.2 Right to Copy

The core of copyright is to allow the creator of the original work to have copies made of its work and to allow for their exploitation. We can see this type of exploitation in several industries such as edition, cinema and many others.

In research, the right to copy is useful towards the training of a language tool or the modification of a software. Therefore, the majority of "open-source" licenses grant licensees the right to copy content from the original work and to reuse this content for subsequent use. One exception of note is the *Community Data License Agreement – Permissive, version 1.0*[13]. This license provides resource users the right to use and publish data but grants no other rights.

3.3 Right to Redistribute

The distribution rights of a copyrighted work are the exclusive rights granted to the copyright owners. Copyright owners or allowed licensors can either distribute their work through proprietary licenses where they may restrict the distribution rights or through "open-source" licenses which can allow third parties to redistribute the work.

This right to redistribute is essential in open science to promote the works that have been produced and allow others to evaluate the quality of research.

Therefore, most "open-source" licenses provide third parties obtaining content placed under those licenses the right to redistribute the original work. In opposition, as an example of proprietary license, the LDC User Agreement for non-members[14] does not allow redistribution of the work protected by the license.

3.4 Right to Distribute Derivatives

We can define a derivative work as a work that includes major elements of copyrighted work that would otherwise be infringing the law if not authorized by the creator of the original work.

This right is essential, especially in research, where we usually need to rework on preexisting works. These preexisting works may be existing copyrighted works on language resources or software that are available prior to any new licensing to third parties. Researchers may need to combine and reuse data available and be allowed to create new works to be distributed to the public.

Usually in "open-source" licenses, this will be provided as a right to rework upon the original work which grants the licensee the right to use a part or the entirety of a work in a derivative work.

However, in the case of the Creative Commons CC-BY-ND license[15], the "ND" denomination stands for "No Derivatives". This can be misleading as the license allows the creation of derivatives works but not their distribution.

3.5 Patent License

Some "open-source" licenses which are used mostly in relation with software and code, such as the Apache License or the General Public License, grant the user a right to modify content protected by patent claims from the original author.

A patent is another exclusive proprietary right that is granted to creators of innovative process and may be attached to some software.

3.6 Right to Grant Sub-Licenses

The copyright owner can grant licenses to third parties and allow them in turn to grant sub-licenses to others so that the content can be wider spread.

In the context of "open-source" research, this ability to sub-license is also crucial as it would allow users to license the content to third parties.

4. Restrictions on Redistribution

The licenses we studied balance the rights that we detailed above with certain obligations that bear on the licensees when dealing with the content.

Therefore, in this section, we will detail the restrictions that are used in "open-source" licenses and that we gathered in

[13] https://cdla.dev/permissive-1-0/

[14] https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf

[15] https://creativecommons.org/licenses/by-nd/4.0/

the “Requirements on redistribution and publications” category of our metadata schema.

4.1 Attribution Requirement

This condition is one of the most often used conditions in open-source licenses. The best-known form of this requirement is the “BY” designation in Creative Commons licenses[16]. This requirement compels the user to attribute the original creator of the work when reusing his content either in derivative work or whenever the content is reused in any way. This is done by reproducing a statement inserted by the original author with its work and comes with a sentence such as “[Title] by [Author] licensed under CC-BY 4.0”.

4.2 Documentation of Modifications

Licensees can also be compelled to document modifications they bring to the original content.

This condition is not based upon any traditional category of rights granted by copyright law. It is specific to software development where documentation is needed especially when a version of a software changes. For example, some GNU-GPL licenses (GPL 3.0 as for the latest version)[17] provide that any new version must carry notices that the content has been modified.

Indeed, this documentation can give essential information on code changes that might change the performance of a piece of software and its features and how they interact together with earlier versions.

Therefore, we thought it was mandatory for us to include this condition as it is essential in reusing or redeveloping software upon available content.

4.3 Retention of Copyright Notice

The retention of the copyright notice means that all derivative works shall keep the attribution notice and full license text from the original works. This retention can also be required for subsequent redistribution of content containing the original work when made by a licensee.

This condition provides that a user inserting content made available under a license providing for this condition must reproduce and retain the copyright notice that is attached to the original content. This is done mainly to remind subsequent users that the original content is available with copyright restrictions and nudges subsequent users to keep their contributions available under such conditions.

4.4 Share-Alike Requirement

As its name suggests, this requirement mandates users of content shared under licenses containing this condition to share any derivative content that they may produce under the same license as the original content.

This is mainly done to keep some form of control over the usage of the content and to maintain the reusability of the work.

By sharing the derived content under the same license as the original content the copyright owner ensures that knowledge can continuously flow under the same licensing scheme.

4.5 Copyleft Requirement

The Copyleft philosophy bears similarities with the ShareAlike requirement. However, the former differs from the latter in the sense that the licensee is required to license the derived content under the same license or a compatible license. The licensee must not impose conditions that may impair the redistribution of the original works afterwards.

The best-known example is the GNU-GPL License that requires users to license the modified works under the same license as the original work.

5. Requirement on Reuse

In addition to the restriction on redistribution of original or derivative work, some of the licenses we studied for this task also provide for some obligations on how the derivative content can be reused by licensees.

5.1 Grant of Commercial Use License

As previously mentioned in Section 3, the original copyright is granted with a set of rights that they can exploit either for free or commercially.

Therefore, when making content available under an open-source license, copyright owners can also decide to allow third parties to make profit from redistribution of the original or derived content. This may be especially useful for developers of commercial applications relying on open-source content while maintaining the underlying content available to all interested users.

One major exception is the Creative Commons CC-BY-NC license[18] which forbids the sharing of data for monetary compensation or commercial advantage.

5.2 Reuse of Content for Specific Activities

This category is not usually mentioned literally in open licenses but due to the focus on research activities of the ELG platform we identified some metadata items that are linked to the reusability of content.

In this section we will detail the different items that fall within this category:

- **Evaluation Use**

This item refers to the possibility of academic or commercial stakeholders to use the resource for the evaluation of technologies. This evaluation can allow to ensure that a resource is suitable for certain purposes. It can also allow to evaluate a language tool in the light of certain measurements.

- **Academic Use and Research Use**

We thought it useful to clearly notify users whether a resource is usable only in academic settings and differentiate them from research use by all types of users.

Even though we can understand them as similar restrictions, Research Use can also cover research and

[16] https://creativecommons.org/licenses/by/4.0

[17] https://www.gnu.org/licenses/gpl-3.0.html

[18] https://creativecommons.org/licenses/by-nc/4.0

development activities undertaken by private enterprises as well as academic research.

- **Language Engineering Research Use**

In addition to the previous category, we thought that it would also be necessary to properly identify research in the Language Engineering field. Indeed, the ELG is a platform that is dedicated to language resources and tools and that helps foster a European innovation space for European Languages.

Therefore, we inferred this condition from the exploitation rights granted by the license. The Computational Use of Data Agreement[19] provides that the content must be used for Computational Use which could imply Language Engineering.

- **Machine Learning Training Use**

Recently, we saw the emergence of language models as being now the primary use of language resources. The enhancement of methods relying on neural networks and artificial intelligence results in a further need for legal certainty on these use cases.

During our study, we considered that the right to create derivatives includes the right to train models with resources, as we believe that a model is derived from the training performed thanks to the resources.

6. Use of Rights for Searching Licenses

The analysis of licenses has produced a long list of rights. Although they are important for understanding the requirements set for users when using an LRT, not all of them are necessary for discoverability purposes. Thus, for the facet "condition of use", we have used only a carefully selected subset of them, to ensure that they cover the most usual user queries.

Similar facets are used in the CLARIN VLO[20] with the facet "Availability" with the CLARIN license categories[21] (Kelli et al. 2018) and the Google dataset search engine[22], where the "usage rights" has only two values: whether commercial use is allowed or not. We have, therefore, restricted the list of conditions to six values, namely: *no conditions*, *commercial use not allowed*, *derivatives not allowed*, *redistribution use not allowed*, *research use allowed*. All rights that are not included in the facet are mapped to the value "other specific restrictions".

7. Conclusion

In this paper we have detailed the various items that we identified during our investigation of licenses and turned into metadata items to help build a "legal search" feature in the ELG platform search engine.

This feature was identified as crucial from the beginning to make sure that the rights of creators are respected and to help reuse and bring legal certainty to all stakeholders.

8. Acknowledgements

We would like to acknowledge the European Language Grid Project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 825627.

9. References

9.1 Bibliographical References

Choukri K., Piperidis S., Tsiavos P., Patrikakos T., Gavrilidou M., Weitzmann J.H. (2012). META-SHARE: Licenses, Legal, IPR and Licensing issues. Deliverable D6.1.3. In T4ME Net (META-NET) project. 24 February 2012.

Kelli A., Lindén K., Vider K., Labropoulou P., Ketzan E., Kamocki P. & Stranák P. (2018). Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? Linköping Electronic Conference Proceedings, 147, 102-111.

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G., Pérez, J. M. G., and Garcia-Silva, A. (2020). Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In Nicoletta Calzolari, et al., editors, Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 5. European Language Resources Association (ELRA).

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May. European Language Resources Association (ELRA).

Piperidis, S., Harris P., Spurk C., Rehm G., Choukri K., Hamon O., Calzolari N., del Gratta R., Magnini B., and Girardi C.(2014). META-SHARE: One year after. In: In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014). European Language Resources Association (ELRA), pp. 1532–1538.

Rehm G., Berger M., Elsholz E., Hegele S., Kintzel F., Marheinecke K., Piperidis S., Deligiannis M., Galanis D., Gkirtzou K., Labropoulou P., Bontcheva K., Jones D., Roberts I., Hajic J., Hamrlová J., Kačena L., Choukri K., Arranz V., Vasiļjevs A., Anvari O., Lagzdiņš A., Meļņika J., Backfried G., Dikici E., Janosik M., Prinz K., Prinz C., Stampler S., Thomas-Aniola D., Pérez J. M. G., Silva A. G., Berrío C., Germann U., Renals S., and Klejch O. (2020). European Language Grid: An Overview. In Nicoletta Calzolari, et al., editors, Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 5. European Language Resources Association (ELRA).

[19] https://spdx.org/licenses/C-UDA-1.0.html

[20] https://vlo.clarin.eu

[21] https://www.clarin.eu/content/licenses-and-clarin-categories

[22] https://datasetsearch.research.google.com/

Rodriguez-Doncel V. and Labropoulou P. 2015. RDF Representation of Licenses for Language Resources. In Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications, pages 49–58, Beijing, China. Association for Computational Linguistics.

Wilkinson, MD., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Growth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E,, Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

9.2 Related works

(License wizards, terms & conditions with licenses):

- https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-find-and-compare-software-licenses
- https://ufal.github.io/public-license-selector/
- https://tldrlegal.com/
- http://licentia.inria.fr/
- RDF representation of licenses: Rodriguez-Doncel and Labropoulou 2015