The Public License Selector: Making Open Licensing Easier

Paweł Kamocki, Pavel Straňák, Michal Sedlák
Institut für Deutsche Sprache, Mannheim / Université Paris Descartes / WWU Münster
Charles University in Prague, Charles University in Prague
Address1, Address2, Address3
kamocki@ids-mannheim.de, stranak@ufal.mff.cuni.cz, sedlak@ufal.mff.cuni.cz

Abstract

Researchers in Natural Language Processing rely on availability of data and software, ideally under open licenses, but little is done to actively encourage it. In fact, the current Copyright framework grants exclusive rights to authors to copy their works, make them available to the public and make derivative works (such as annotated language corpora). Moreover, in the EU databases are protected against unauthorized extraction and re-utilization of their contents. Therefore, proper public licensing plays a crucial role in providing access to research data. A public license is a license that grants certain rights not to one particular user, but to the general public (everybody). Our article presents a tool that we developed and whose purpose is to assist the user in the licensing process. As software and data should be licensed under different licenses, the tool is composed of two separate parts: Data and Software. The underlying logic as well as elements of the graphic interface are presented below.

Keywords: Open Source, Open Data, Open Access, License, IP Rights, Copyright, Creative Commons

1. Introduction

Researchers in Natural Language Processing, just like in Digital Humanities, rely on availability of data and software, ideally under open licenses, but little is done to actively encourage it. Unlike many other projects, CLARIN has been paying particular attention to licensing questions. Most of us have experienced situations when having read a good paper, we have an idea to test or improve its hypotheses, but cannot get access to the underlying data. Or, having created a dataset we cannot make it broadly available because we not sure what we are allowed to do with the data. As far as software is concerned, the situation is similar, and even systems reported as state-of-the-art in the field need not be available. In computation linguistics, ACL has the option to submit data or software with a paper, but permanent data and software availability to all, not just to a reviewer, is still not required.

In fact, the current Copyright framework grants exclusive rights to authors to copy their works, make them available to the public and make derivative works (such as annotated language corpora). Moreover, in the EU databases are protected against unauthorized extraction and re-utilization of their contents.

Therefore, proper public licensing plays a crucial role in providing access to research data. A public license is a license that grants certain rights not to one particular user, but to the general public (everybody). However, the choice of a proper license is an uneasy (and often neglected) task. Therefore, tools like licentia.inria.fr or ELRA License Wizard[1] have been created. As a part of this movement, we developed a tool that assists the user in the licensing process and that might be an alternative to other license choosers – the Public License Selector.[2]

2. The Public License Selector

Before the License Selector can be presented, it is essential to define the notion of a public license. A public license is a license that grants certain rights not to an individual user, but to the general public (every potential user). Public licenses for software has been known since 1980s (when software licenses such as BSDL, MIT or GNU GPL emerged). However, public licenses for other categories of works (including datasets) only appeared in the 21st century, mostly due to the creation of the Creative Commons foundation. The latest version of the CC license suit (including six licenses, a waiver and a public domain mark), CC 4.0, is well adapted for datasets, as it covers not only copyright, but also the sui generis database right, but older versions are still in use. While choosing a license, one has to keep in mind that the licenses which are appropriate for software are not appropriate for data and vice versa. Moreover, not all public licenses are 'open', i.e. not all of them meet the requirements for Open Access/Open Data/Open Source label.

Software is a very particular category of copyright-protected works. In fact, unlike in case of other works, using software consists of making reproductions of the code in the memory of a computer; without these reproductions, software is completely useless for a human being. Its utilitarian character and the ways in which is created (often by a team of developers, rather than by an inspired individual) also distinguishes software from other categories of copyright-protected works. These particularities are reflected in the legal framework that applies to software. In fact, in the EU the copyright protection of software is regulated not by the Copyright Directive (2001/29/EC), but by a special Software Directive (2009/24/EC) whose first version (adopted as early as 1991) predates the Copyright Directive. Therefore, the restricted acts, the statutory exceptions and the rules on authorship of software may differ from those concerning other categories of works. Furthermore, unlike most copyright-protected works, software

[1] http://www.elda.fr/en/dissemination/press-releases/elra-license-wizard-pr/

[2] https://github.com/ufal/public-license-selector

might be patentable (at least in some jurisdictions). All these particularities have to be taken into account in the licensing process. As a consequence, data licenses (like Creative Commons) differ substantially from software licenses (like GNU GPL). It is reflected even in the language used by these licenses; while CC use the general expression "licensed material", software licenses use specific terms like Program (in GNU GPL) or software (MIT license). The Creative Commons foundation itself does not recommend the use of its licenses for software[3]. Exceptionally, a software license may cover documentation and data that cannot be separated from the software itself (like in an XML file), but whenever the separation is possible, it is strongly recommend to respect the dichotomy between data and software licenses.

Since data and software (applications) should be licensed under different licenses, the tool is in fact divided into two parts: selecting software licenses or selecting data licenses. This distinction is not present in other tools that we are aware of. We believe that the distinction between data and software is both intuitive and important. Our tool provides a selection of popular Open Source[4] licenses for software and Creative Commons 4.0 for data. We avoid redundant licenses, trying to pick the best license within each category.

2.1. Data Licensing

[htb] Creative Commons is a US foundation created by Larry Lessig, a famous copyright scholar and activist. Since 2001, the foundation proposes a series of licenses built of four building blocks, i.e.: attribution (BY), share-alike (SA), non-commercial (NC) and no derivatives (ND). In our view, the latest version of these licenses, Creative Commons 4.0, is the best tool for data licensing, as it covers (unlike the previous versions) not only copyright, but also related rights, including the database right. This is why CC licenses are the core element of our License Selector, although it is still possible to choose other licenses.

The first – and arguably the most complicated – question that the user has to answer while choosing a data license is ‘Is your data within the scope of copyright and related rights?’. As IP law in the European Union is merely harmonized and not unified, the exact scope of copyright and similar rights may differ between Member States (e.g. some Member States recognize an exclusive right for ‘scientific and critical editions’, while others don’t). The answer to this first question requires some basic knowledge from the user (in particularly complicated cases, the answer may not be obvious even to an expert lawyer). This is why the tool provides the user with some information that appears on the screen while the user places the cursor over the phrase ‘scope of copyright and related rights’ – see Figure 2.

If the user’s answer is in the negative, the data is identified as being in the public domain and the License Selector suggests the use of a CC Public Domain Mark.

If the answer is in the positive, the next question that the user has to answer is ‘Do you own copyright and similar rights in your dataset and all its constitutive parts?’. The user shall answer in the positive (according to the instructions that show when he places the cursor over the phrase) if he is the author of the dataset and/or if he is the producer of the database in which the data are contained. In such a case, the user can move directly to the choice of a license. If this is not the case, the user is asked if all the elements of the dataset are licensed under a public license (such as Creative Commons or Open Data Commons) or in the public domain (i.e. the work is within the scope of an exclusive right, but its term expired; the definitions of both notions can be seen if the user places the cursor over the respective notions). If the answer is in the negative, additional permission is required. The License Selector cannot suggest any license, but instead suggests the user to contact the legal helpdesk in his institution.

If all the elements of the dataset are licensed under a public license or in a public domain, the user is then requested to choose the licenses that are present in his dataset. The list of possible choices include all CC licenses (regardless of the version), Open Data Commons Licenses, CC0 and CC Public Domain Mark as well as ‘unmarked’ public domain (for works that belong to the public domain, but are not marked with a CC Public Domain Mark). Multiple options can be checked, see Figure 3. The output of this stage influences the choice of the license by the License Selector; in short, the most restrictive license present in the dataset is at the same time the least restrictive (or the only) license that can be used for the whole dataset, see Figure 4 for the result of choices as shown in Figure 3. Moreover, data licensed under a license containing an ND (no derivatives) requirement cannot be ‘mixed’ into a larger dataset with other data, as this would violate the terms of the license (compilation being an Adapted Material under section 1 a) of the CC BY-ND 4.0 license).

The next stage is the actual selection of the license. In this part of the process, the user is asked a series of questions, each of which allows to determine whether a given license requirement (BY, NC, ND, SA) should be included in the selected license or not (i.e. whether the requirement is ’picked up’ on the way). Depending on the outcome of the previous stage, some questions may not be asked (the requirement is ’picked up’ automatically to comply with other licenses present in the dataset).

At the end of this stage, the License Selector suggests one CC 4.0 license (or a CC0 waiver). While the tool allows the user to choose between all the CC licenses, only two of them meet the definition of Open Data (according to the Open Definition, Open means anyone can freely access, use, modify, and share for any purpose – subject, at most, to requirements that preserve provenance and openness); those licenses are marked with the Open Data label.

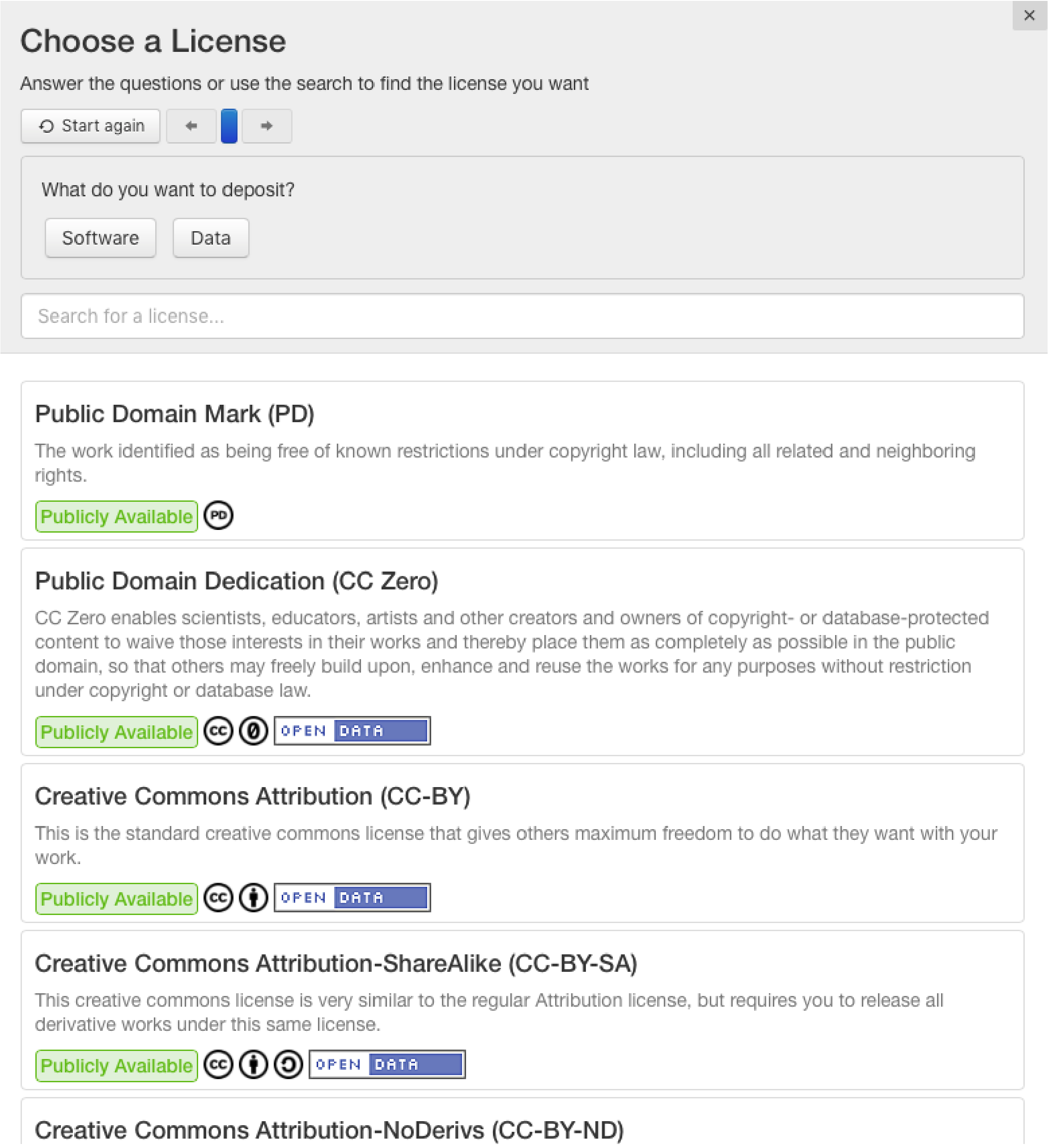
2.2. Software Licensing

The Open Source Initiative, created in 1998, adopted the Open Source Definition (OSD), according to which software licensed under a license that meets a set of specific criteria [5] can be labeled with an open-source certification

[3] https://wiki.creativecommons.org/index.php/Frequently_Asked_Questions#Can_I_apply_a_Creative_Commons_license_to_software.3F

[4] http://opensource.org/osd

[5] see https://opensource.org/osd-annotated



Choose a License

Answer the questions or use the search to find the license you want

What do you want to deposit?

Public Domain Mark (PD)

The work identified as being free of known restrictions under copyright law, including all related and neighboring rights.

Public Domain Dedication (CC Zero)

CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Creative Commons Attribution (CC-BY)

This is the standard creative commons license that gives others maximum freedom to do what they want with your work.

Creative Commons Attribution-ShareAlike (CC-BY-SA)

This creative commons license is very similar to the regular Attribution license, but requires you to release all derivative works under this same license.

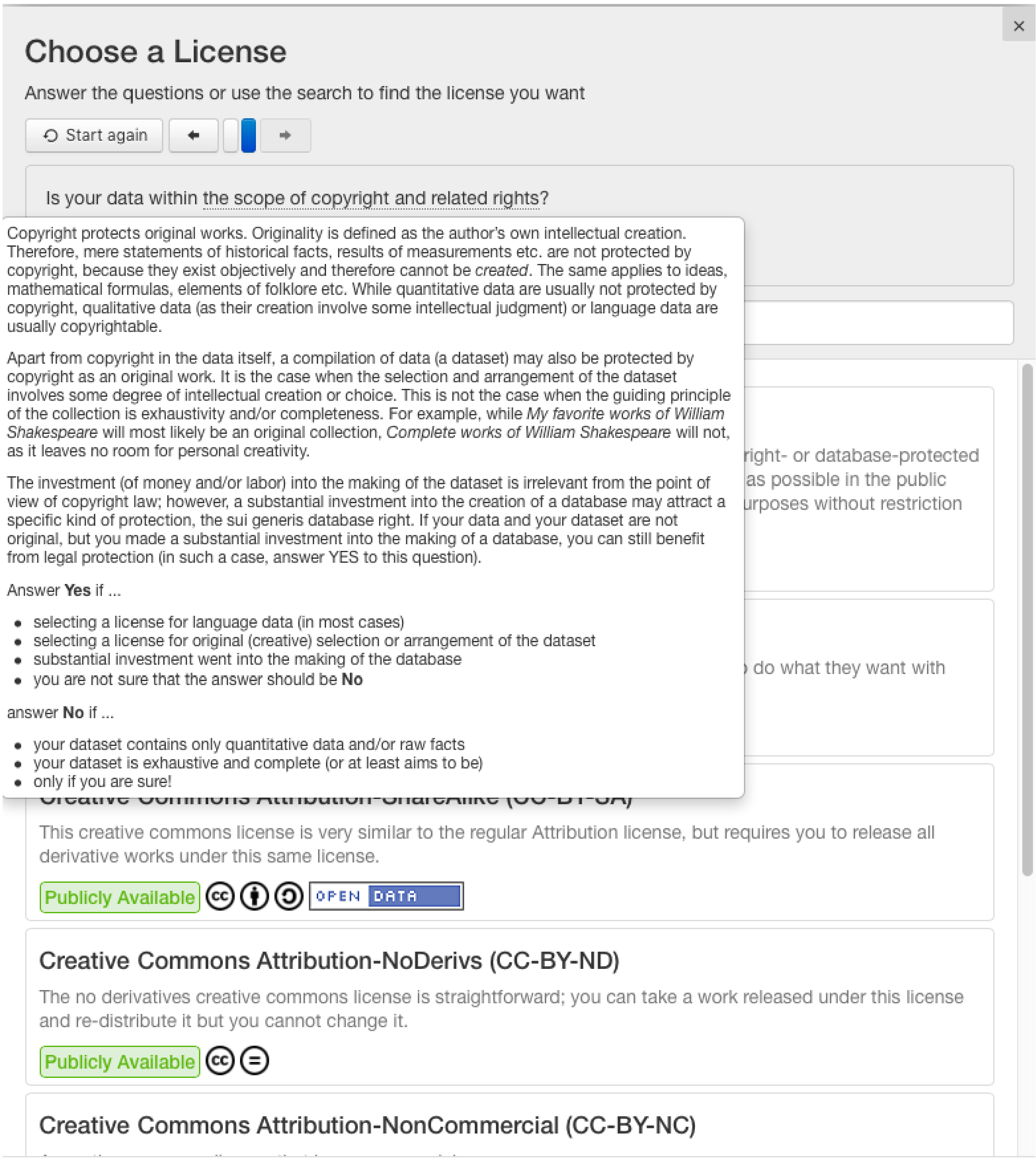
Creative Commons Attribution-NoDerivs (CC-BY-ND)

Figure 1: The first question influencing all the rest: Are you licensing software or data? As explained in Section 2. the difference is crucial.

mark. Each of the licenses that the License Selector allows the user to choose from have been approved as Open Source by the Open Source Initiative.

The part concerning software licensing follows a fundamentally different logic. Because of the plethora of available software licenses and the fact that sometimes it is difficult to make a clear distinction between their requirements, instead of being pointed towards one particular license, the user is pointed towards a group of licenses. In order to facilitate the choice, some policy decisions were taken to determine a hierarchy of licenses – the most recommendable licenses appear first, while the least recommendable come at the end of the list. The essential criterion for 'recommendability' was the number of licenses that a given license is compatible with – in most cases a license compatible with more licenses should be chosen over a license providing for less compatibility. See Figure 5 for such an ordered list of compatible licenses.

The licenses are divided into three groups: permissive licenses (including – in that order – the MIT License, the Free BSD License (2-clause), the New BSD License (3-clause), Apache License 2.0 and Artistic License 2.0), weak copyleft licenses (including – in that order – GNU LGPL 2.1, GNU LGPL 2.1 or later, GNU LGPL 3.0, Mozilla Public License 2.0, Eclipse Public License 1.0, Common Development and Distribution License 1.0), strong copyleft licenses (including – in that order – GNU GPL 2.0 or later, GNU GPL 3.0, GNU GPL 2.0) and an additional group of network copyleft licenses (Affero GPL v. 3 and Affero GPL v. 2). The choice of licenses is therefore relatively broad,



Choose a License

Answer the questions or use the search to find the license you want

Is your data within the scope of copyright and related rights?

Copyright protects original works. Originality is defined as the author's own intellectual creation. Therefore, mere statements of historical facts, results of measurements etc. are not protected by copyright, because they exist objectively and therefore cannot be *created*. The same applies to ideas, mathematical formulas, elements of folklore etc. While quantitative data are usually not protected by copyright, qualitative data (as their creation involve some intellectual judgment) or language data are usually copyrightable.

Apart from copyright in the data itself, a compilation of data (a dataset) may also be protected by copyright as an original work. It is the case when the selection and arrangement of the dataset involves some degree of intellectual creation or choice. This is not the case when the guiding principle of the collection is exhaustivity and/or completeness. For example, while *My favorite works of William Shakespeare* will most likely be an original collection, *Complete works of William Shakespeare* will not, as it leaves no room for personal creativity.

The investment (of money and/or labor) into the making of the dataset is irrelevant from the point of view of copyright law; however, a substantial investment into the creation of a database may attract a specific kind of protection, the sui generis database right. If your data and your dataset are not original, but you made a substantial investment into the making of a database, you can still benefit from legal protection (in such a case, answer YES to this question).

Answer **Yes** if ...

- selecting a license for language data (in most cases)
- selecting a license for original (creative) selection or arrangement of the dataset
- substantial investment went into the making of the database
- you are not sure that the answer should be **No**

answer **No** if ...

- your dataset contains only quantitative data and/or raw facts
- your dataset is exhaustive and complete (or at least aims to be)
- only if you are sure!

right- or database-protected as possible in the public urposes without restriction

do what they want with

Creative Commons Attribution-ShareAlike (CC-BY-SA)

This creative commons license is very similar to the regular Attribution license, but requires you to release all derivative works under this same license.

Creative Commons Attribution-NoDerivs (CC-BY-ND)

The no derivatives creative commons license is straightforward; you can take a work released under this license and re-distribute it but you cannot change it.

Creative Commons Attribution-NonCommercial (CC-BY-NC)

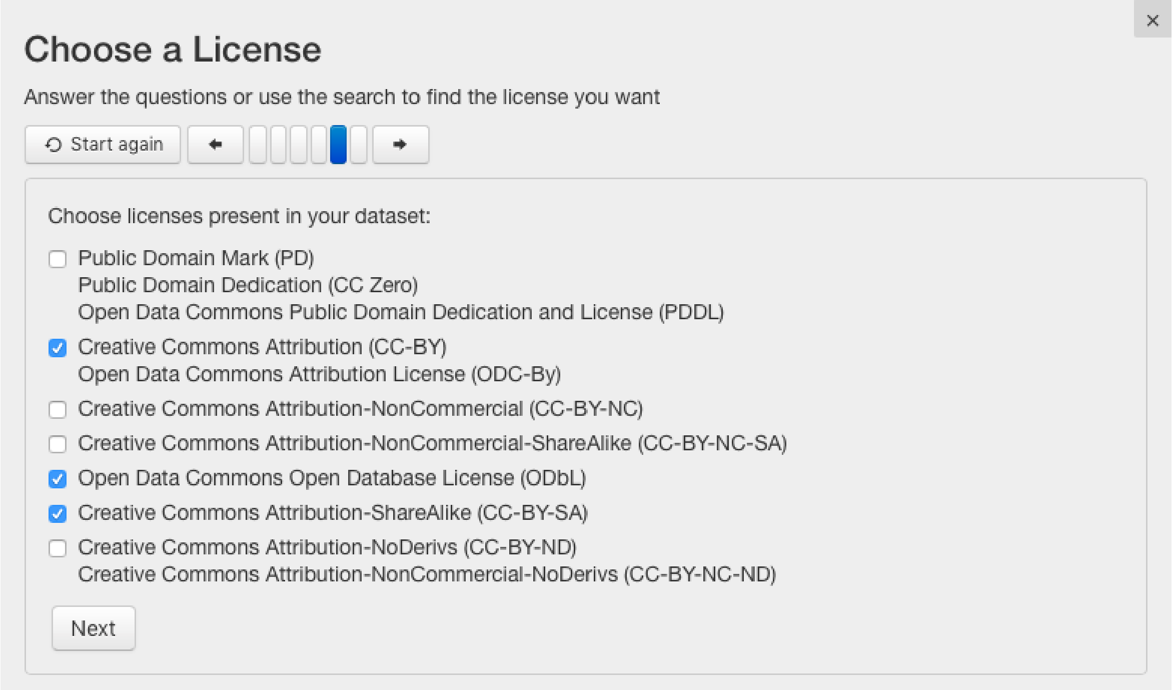
Figure 2: Explanations of legal terms are provided. For the Scope of Copyright the explanation is long, since it is a complicated issue. Other explanations are much shorter.

but it is impossible to include them all – especially that some software licenses are in practice only used in certain specific projects. The list, however, can be easily expanded in order to take other licenses into account.

The compatibility chart, which eliminates certain 'incompatible' licenses at this stage of the process is based on the information provided by the authors of every given license, and not on our subjective assessment.

The first question asked to the user at the next stage is 'Is your code based on existing software or is it your original work?'. The purpose of this question is to determine whether the license can be freely chosen (this is the case if the user is the only author of original code), or whether the choice is limited by other licenses present in the code. If the user answers 'my own code', the tool goes directly to the group selection. If, however, the answer is 'based on existing code', the tool asks the user to choose the licenses present in his code from a list containing all the licenses listed above, as well as 'public domain' and 'other license/no license' options. The choices made at this stage narrow down to possible selection of licenses. Moreover, it is possible that the user chooses two or more incompatible licenses – in this case the tool says that no license can be chosen and the user is pointed towards the legal helpdesk, which may be able to help him get additional permissions from the right holders. In the second part of the process, the user is asked two questions:

Do you want others who modify your code to be forced



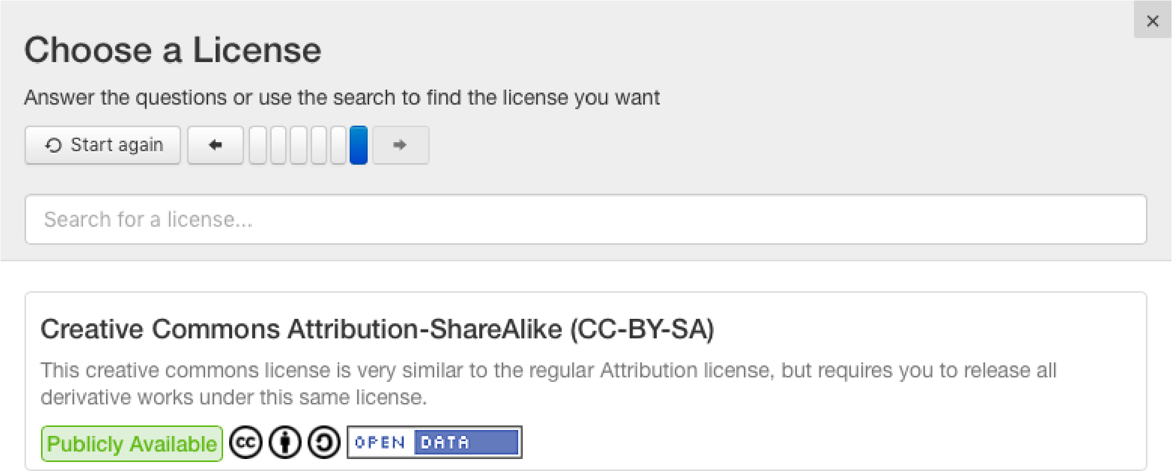
Choose a License

Answer the questions or use the search to find the license you want

Choose licenses present in your dataset:

- □ Public Domain Mark (PD)
Public Domain Dedication (CC Zero)
Open Data Commons Public Domain Dedication and License (PDDL)
- ☑ Creative Commons Attribution (CC-BY)
Open Data Commons Attribution License (ODC-By)
- □ Creative Commons Attribution-NonCommercial (CC-BY-NC)
- □ Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA)
- ☑ Open Data Commons Open Database License (ODbL)
- ☑ Creative Commons Attribution-ShareAlike (CC-BY-SA)
- □ Creative Commons Attribution-NoDerivs (CC-BY-ND)
Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND)

Figure 3: Indicating licenses of the datasets that the new dataset is derived from.



Choose a License

Answer the questions or use the search to find the license you want

Creative Commons Attribution-ShareAlike (CC-BY-SA)

This creative commons license is very similar to the regular Attribution license, but requires you to release all derivative works under this same license.

Figure 4: The final list of possible licenses for a work derived fom existing works under licenses as shown in Figure 3

to release the modified code under an open source license? (NO = the user is directed towards the group of permissive licenses; YES = question 2 is asked);

Is your code used directly as an executable or are you licensing a library (your code will be linked)? (EXECUTABLE = the user is pointed towards strong copyleft licenses AND network copyleft licenses; LIBRARY = the user is pointed towards weak copyleft licenses).

It should be pointed out that strong copyleft licenses and network copyleft licenses are presented together (with strong copyleft licenses appearing higher on the list) and the choice is in practice left to the user. We assume that those who want to deposit 'network software' will have enough knowledge to choose a 'network' license from the list. If necessary, this issue can be easily fixed by adding a hypothetical third question, such as e.g. 'Is your software intended for use over a computer network?' (YES = the user is directed towards network copyleft licenses; NO = the user is directed towards strong copyleft licenses).

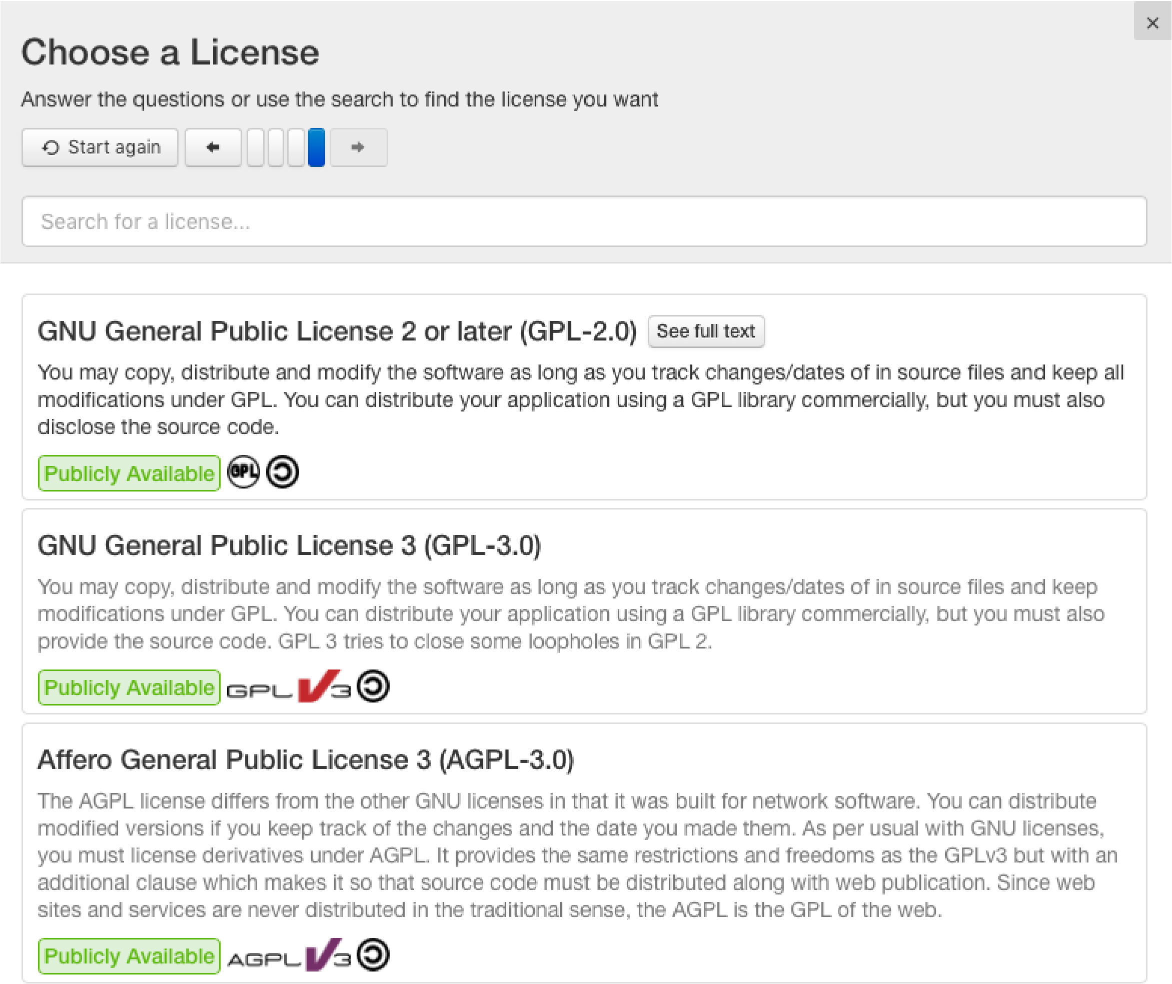
3. Conclusion

The Public License Selector is fully implemented and available; it has been integrated in the submission workflows of two repository systems: The LINDAT fork of DSpace[6] used currently by four Clarin centres,[7] and the EUDAT B2SHARE project.[8] It is available under the very permis-

[6] `https://github.com/ufal/lindat-dspace`

[7] `clarin.cz`, `clarin.si`, `clarin-pl.eu`, `clarino.uib.no`

[8] `https://b2share.eudat.eu`, `http://hdl.handle.net/11346/G0VE`



Choose a License

Answer the questions or use the search to find the license you want

GNU General Public License 2 or later (GPL-2.0)

You may copy, distribute and modify the software as long as you track changes/dates of in source files and keep all modifications under GPL. You can distribute your application using a GPL library commercially, but you must also disclose the source code.

GNU General Public License 3 (GPL-3.0)

You may copy, distribute and modify the software as long as you track changes/dates of in source files and keep modifications under GPL. You can distribute your application using a GPL library commercially, but you must also provide the source code. GPL 3 tries to close some loopholes in GPL 2.

Affero General Public License 3 (AGPL-3.0)

The AGPL license differs from the other GNU licenses in that it was built for network software. You can distribute modified versions if you keep track of the changes and the date you made them. As per usual with GNU licenses, you must license derivatives under AGPL. It provides the same restrictions and freedoms as the GPLv3 but with an additional clause which makes it so that source code must be distributed along with web publication. Since web sites and services are never distributed in the traditional sense, the AGPL is the GPL of the web.

Figure 5: After answering the questions, an end result may be a list of several compatible licenses. As explained in Section 2.2. the list is ordered by our level of recommendation.

sive MIT License, thus free to modify, extend and use also in other ways.[9]

4. Bibliographical References

Kamocki, P. and Ketzan, E.). Creative commons and language resources: General issues and what's new in CC 4.0.

Stodden, V. (2009). The legal framework for reproducible scientific research: Licensing and copyright. *IEEE Computing in Science and Engineering*, 11(1):35–40.

[9] http://opensource.org/licenses/mit-license.php