elexis european lexicographic infrastructure

D6.2

RECOMMENDATIONS ON LEGAL AND IPR ISSUES FOR LEXICOGRAPHY

Author(s): Bob Boelhouwer, Iztok Kosem, Sanni Nimb, Miloš Jakubíček, Carole Tiberius, Simon Krek, Morten Rosenmeier

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D6.2 RECOMMENDATIONS ON LEGAL AND IPR ISSUES FOR LEXICOGRAPHY

Deliverable Number:	6.2
Dissemination Level:	Public
Delivery Date:	31. 1. 2020
Version:	1.0
Author(s):	Bob Boelhouwer
	Iztok Kosem
	Sanni Nimb
	Miloš Jakubíček
	Carole Tiberius
	Simon Krek

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
21/05/2019	First draft	Bob Boelhouwer, Carole Tiberius
21/12/2019	Review of legal aspects	Morten Rosenmeier
07/01/2020	Feedback	Sanni Nimb, Iztok Kosem, Miloš Jakubíček
15/01/2020	Chapter on survey results added	Iztok Kosem
29/01/2020	Proofreading, incorporation of feedback	Simon Krek, Sanni Nimb, Bob Boelhouwer, Carole Tiberius
31/01/2020	Final version	Iztok Kosem

Table of Contents

List of Tables

List of Figures

Summary

The document offers a detailed overview of different licenses for data and software, describes and recommends the license selection for different purposes within the ELEXIS infrastructure, and presents the results of a survey that focused on current licensing situation in institutions that produce lexicographic content. All this can be read in detail in the document, however, in this introductory section we summarize all the findings in the form of recommendations for good practice in licensing lexicographic data and software:

1. It is recommended that open licenses are used whenever possible, using standard licensing schemas.
2. Licensing and Intellectual Property rights issues connected to data or software should be carefully considered at the very beginning of a lexicographic project, i.e. at the planning or proposal writing stage. Consulting legal experts and possibly lexicographic community is recommended.
3. If an open license cannot be used for the entire lexicographic dataset, using different licenses for different types/parts of lexicographic data should be considered.
4. For software, one of the FLOSS licenses is recommended, preferably GNU Public License v3.0, GNU Library or Lesser General Public License v.3.0 and Apache License v2.0.
5. To address potential safety concerns and IPR abuse, it is recommended to use established repositories that use authentication protocols.

1 General introduction, legal background[1]

The creation of a dictionary of quality requires a large amount of highly skilled labor. Such a product therefore is of high value to the creators, the sponsors, and to the users. The owners of these dictionaries seem to have a reason to protect their data. If the owner is a private organization, the reason will probably lay in the commercial value of the data. For public organizations, however, there are also reasons to protect their data. Some organizations may need to prove their relevance to the funding provider by reporting visits to the website where their data can be consulted. Some organizations may even generate income from paid access or commercial publication of their data. Furthermore, certain organizations may not be allowed by their funding provider or governing organization to hand the data to others. The quality of the data may also be a consideration; some organizations might want to maintain tight control over their data in order to avoid that diluted or deprecated versions of the data undermine its usability and the organization's reputation.

This document provides general information on licensing methods and forwards recommendations for certain licensing schemes. Even though some legal information is included, it is important to underline that this work is not intended to provide legal advice.

The document investigates licenses that particular institutions might want to attach to their products. The document therefore does not investigate licenses for products received/obtained from other parties (e.g. content from publishers, software from developers).

There are basically two ways to regulate the use of content:

- By means of an individually formulated license, i.e. a license in which the copyright owner defines the acts which can be performed without permission.
- By means of “standardized” licenses, the content of which is not formulated by the copyright owner, but by a standardization organization.

[1] This section and the next is based on a report previously published as part of the Succeed Project (FP7-ICT grant 600555). See https://www.digitisation.eu/training/recommendations-for-digitisation-projects/recommendations-on-licensing-recommendations/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1.1 Aspects of intellectual property legislation

Intellectual property (IP) rights, very broadly, means the legal rights which result from intellectual activity in the industrial, scientific, literary and artistic fields. Countries have laws to protect IP for two main reasons. One is to give statutory expression to the moral and economic rights of creators in their creations and the rights of the public in access to those creations. The second is to promote, as a deliberate act of Government policy, creativity and the dissemination and application of its results and to encourage fair-trading, which would contribute to economic and social development[2].

There is an extensive body of information on IP legislation. We will not attempt to cover all of it. Some notions, however, are important and will be discussed next.

1.2 Copyright on original works

The basic legal framework for copyright at the international level is the „Berne Convention for the Protection of Literary and Artistic Works", 1886, generally known as the „Berne Convention"[3]. Signatory countries are required to recognize copyright of works originating in other signatory countries in the same way as they recognize the copyright of their own nationals. Copyright protection is automatic, not subject to any formality. In addition, the agreement establishes certain minimum standards of protection concerning the rights acknowledged (moral rights and economic rights), limitations to the exclusive rights and the duration of the copyright.

Whenever an author produces a creative work, under the Berne Convention he/she automatically becomes owner of the copyright of this work[4]. Owners of a creative work have the exclusive right to use the work and to authorize others to use it on agreed terms.

[2] Chapter 1, WIPO Intellectual Property Handbook: Policy, Law and Use: https://www.wipo.int/about-ip/en/iprm

[3] https://www.wipo.int/treaties/en/ip/berne

[4] This may be different, depending on the legislation of the country in which the work has been produced, when the author is hired by a person or organization to produce the work.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Under the Berne Convention, the general minimum duration of the copyright is life of the author plus 50 years after his/her death[5], but contracting parties are free to provide longer terms. The European Union extended that term with the 1993 Directive on harmonizing the term of copyright protection[6]. In the European Union a work is protected for 70 years after the death of the author[7]. If the author is not known, it is protected for 70 years after its first lawful publication. The Berne Convention authorizes countries to allow certain free uses of copyrighted works. This includes, for instance, the reproduction of limited parts of copyrighted material for certain purposes (e.g., review, news reporting, teaching or scholarly research) without obtaining permission from the author and without paying a fee or royalty.

1.3 Copyright on derivative works

Works derived from original works can also bear copyright[8]. Digitized versions of an originally printed work can sometimes establish derivative works. To be eligible for copyright, a derivative work must be different enough from the original to be considered a "new work" or must contain a substantial amount of new material. So simply reproducing a public domain text in digital form would not create a derivative work. However, by enhancing the work with links, annotations, sound recordings or images it likely will. It is important to note that when the original work is not in the public domain, its copyright holder has to authorize the creation and exploitation of the derivative work.

[5] There are a few exceptions to this general term for certain categories of works, like cinematographic works (minimum protection of 50 years after the work has been made available to the public, or, if not made available, 50 years after the making of such a work) or photographic works and works of applied art (minimum 25 years from the making of the work).

[6] Directive 93/98/EEC was repealed and replaced by Directive 2006/116/EC, amended by Directive 2011/77/EU. See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32011L0077

[7] Same general protection term applies also in the United States since the Copyright Term Extension Act, 1998.

[8] See: https://www.wipo.int/tk/en/resources/glossary.html#19, https://en.wikipedia.org/wiki/Derivative_work and http://www.publicdomainsherpa.com/derivative-work.html

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1.4 European Copyright Directive 2019/790

The new directive has three main objectives. First, protecting press publications from "abuse" by large internet platforms, second, creating copyright exemptions for (academic) text and data mining and teaching and, finally, enabling heritage institutions to (re)publish orphan works.

The first objective mentioned above (cf. Art. 17 of directive) seems not of concern for the ELEXIS project, since it pertains to 'online content-sharing service providers', which by definition (cf. Art. 2-6 of directive) are commercial organizations.

1.4.1 Copyright exemption for (academic) text – and data mining.

Article 3 of the directive broadens the possibility for academic researchers to harvest data for text research and data mining. This pertains to texts to which they have legal access (par. 1), also when the rights owner explicitly prohibits it (par 2). Of course, the data can only be used for analysis and cannot be redistributed in any form.

1.4.2 Enable heritage institutions to (re)publish orphan works

According to Article 7, works that are out-of-commerce, which are not available and are not likely to become available to the public through customary commercial channels, must be the subject of "extended collective licenses". This means that when a collective rights management organisation allows a cultural heritage institution to digitize or in other ways exploit out-of-commerce works, this permission should not only be binding for the copyright owners representened by the organization, but in some instances also for the copyright owners not represented by the organisation. A number of requirements must be met. Among other things, right holders should have a possibility to object.

This might open up possibilities to include out-of-print dictionaries into the public deliverables of the project.

1.5 Database rights

The so-called "sui generis" database right is also a property right, which is to a certain extent comparable to but distinct from copyright. According to the Directive 96/9/EC on the Legal Protection

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

of Databases[9], a database is a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. As such, it can be protected by copyright if there is originality in the selection or arrangement of the contents, and/or by the sui generis right. The latter right is granted to the maker of the database, and for it to apply, there must have been a substantial investment in obtaining, verifying or presenting its contents. It is possible that a database will satisfy both these requirements so that both copyright and the sui generis right apply.

There is no registration required for database rights - it is an automatic right like copyright. However, the term of protection under database right is in principle much shorter than under copyright, as it lasts for 15 years from the making of the database[10]; if made available to the public before expiry of that period, then the term is 15 years from the making it available to the public.

Many databases are a collection of copyright works, such as a database of poetry from the last fifty years where each poem will also in itself be protected by copyright. So people compiling databases need to make sure that they have permission from the copyright owners for use of their material, and people using databases need to be aware of the rights of the owners of underlying works as well as database rights owners.

[9] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML

[10] It should be noted that any substantial change to the contents of a database, which would result in the database being considered to be a substantial new investment, shall qualify the database resulting from that investment for its own term of protection (art. 10.3 Database Directive).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2 Common public licensing schemes

In this chapter we will present options for institutions to license materials that they want to publish online. We will assume that there are no other parties that have copyright on those materials. Publishing material that is copyrighted might need a license that is agreed on by the rights owner and it is not possible to give general recommendations for such licenses. In the following sections we make a distinction between software, content (original creative material) and data (a collection of information). We will not further consider licensing of metadata (descriptions of data).

In this section we will present a number of licensing frameworks with a noncommercial nature. The organizations that provide these frameworks all have the ideal that sharing content, data and metadata is advantageous for the common good.

As was mentioned in Chapter 1, the owner of a work can authorize others to use that work. Usually such authorization is formalized in the form of a license. A license is a contract between the owner (or its representative) of copyrighted material and its user. The license specifies what the user can do with the material and also what the user is not allowed to do. If the user has to pay a fee for the use of the material, we consider it a commercial license. If no fee is involved, we consider it a non-commercial license. Licenses typically differ in the types of usage they allow to the user and the types of usage that are prohibited.

2.1 Licensing content

2.1.1 Creative Commons (CC)

Creative Commons[11] is a non-profit organization that has released several standardized copyright licenses known as “Creative Common Licenses” free of charge to the public. With these licenses the owner of a work can easily express which rights they want to reserve and which rights they waive for the benefit of users of the work, without having to formulate their own license provisions. All CC licenses grant the user the right to redistribute the work under certain conditions.

[11] See: https://creativecommons.org/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

The six licenses are combinations of four conditions. Below we explain these conditions.

Attribution (BY). Users of the work need to credit the author.

NoDerivs (ND). The work should be passed along unchanged and in whole.

NonCommercial (NC): The work may not be sold or in other ways be used commercially.

ShareAlike (SA): The work may be changed and build upon, as long as the users license their new creations under identical terms.

All six CC licenses have at least the „Attribution" condition (which is an expression of one of the basic moral rights of the author, i.e., the paternity right). The combination of „NoDerivs" and „ShareAlike" does not occur because these conditions exclude each other. Consequently, the six licenses are:

CC BY: *Attribution*. This license lets others distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the author for the original creation. This is the most accommodating of the licenses offered. Recommended for maximum dissemination and use of licensed materials.

CC BY-ND: *Attribution-NoDerivs*. This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to the author.

CC BY-NC-SA: *Attribution-NonCommercial-ShareAlike*. This license lets others remix, tweak, and build upon the work non-commercially, as long as they credit the author and license their new creations under the identical terms.

CC BY-SA: *Attribution-ShareAlike*. This license lets others remix, tweak, and build upon the work even for commercial purposes, as long as they credit the author and license their new creations under the identical terms. This license is often compared to "copyleft" free and open source software licenses. All new works based on this one will carry the same license, so any derivatives will also allow commercial use.

CC BY-NC: *Attribution-NonCommercial*. This license lets others remix, tweak, and build upon the work non-commercially, and although their new works must also acknowledge the author and be non-commercial, they do not have to license their derivative works on the same terms.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

CC BY-NC-ND: *Attribution-NonCommercial-NoDerivs*. This license is the most restrictive of the six main licenses, only allowing others to download the work and share it with others as long as they credit the author, but they cannot change the work in any way or use it commercially.

In addition to the licenses above, Creative Commons provides a tool (CC0) to waive all rights on a work. It is difficult to place a work in the public domain as long as the automatic copyrights and database rights have not yet expired. But CC0 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.

2.2 Licensing data

2.2.1 Creative Commons version 4.0

Version 4.0 of the CC license suite addresses database rights in addition to copyright and other copyright-like rights. Because database rights can impede user's ability to share, reuse, and modify a work in the same way that copyright can, 4.0 makes it clear that these permissions apply to works that would otherwise be restricted by database rights as well.

2.2.2 Open Data Commons (ODC)

Open Data Commons[12] provides a set of legal tools to provide open data. The licenses make an explicit distinction between "data" and "content". Data would pertain to collections of content or information typically organised in a database (see section 2.3 on database rights). The ODC database license (ODC-ODBL) grants the user the rights to copy or redistribute the database, to produce works from it, and to modify, transform or expand it. There are, however, provisions to these rights. If a derived work is published, the user must attribute the original database. Derived (published) work should also be shared under the ODC database license and it is possible to redistribute the database (or a derived version) in a closed form as long as an open version is made available as well.

[12] See: https://opendatacommons.org/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Connected to the license for databases, ODC provides a separate license for database content (ODC-DBCL). That license is to be used in combination with the database license.

Next, ODC provides an attribution license (ODC-BY). This license gives the user the same options for application of the database as the ODC database license, but with less provisions. Only attribution is required.

Finally, ODC provides the Public Domain Dedication and License (ODC-PDDL). Using this license will place the database in the public domain. Thus, users can use and redistribute the database without provisions.

2.2.3 End user licensing

In the previous sections we described a number of licensing frameworks that allow open access, which means that the data can be used by anyone without a fee and without requiring consent of the owner. Moreover, both frameworks allowed redistribution (or sharing) of the data.

Allowing redistribution means that the owner loses a lot of control of the data. It is more difficult to measure the number of users of the data (impact) and manage version control of the data.

To remedy the above problems it is possible to employ end user licenses. The main difference with the CC and ODC frameworks is that the user is prohibited from redistributing the data to others.

End user licenses exist in many forms and often include a fee. But they can also be of a more accessible kind in the form of an end user agreement in which a user only has to agree to the license by clicking a button after which they will gain free access to the product.

CLARIN has made available a number of 'End User Licensing Agreements' (EULA's)[13] that can be used for these purposes.

2.3 Licensing software

In general software license provides information on the possibilities and restrictions related to usage, modification or distribution of the software. There is a large variety among available standard licenses.

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

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Depending on the restriction level of the license under which the software tool is released, we can distinguish several types of software licenses:

- Free/Libre and Open Source Software (FLOSS) – this is the most liberal approach to software distribution and it means that the software is open and free. Open means that the source code is available to any user. Free means freedom in terms of usage, modification and distribution of the software.
- Proprietary software – it usually has a lot of restrictions put on the usage and there is usually no access to source code. For example, users cannot modify the software and can use it for personal purposes only.
- Multi-license software – it can be used in various scenarios to provide software to different groups of users with different licenses, which is a mixture of the two above.

There are many licenses that can be used in the context of multi-licensing or FLOSS and it is useful to identify the most common ones. Table 1 presents a list of popular licenses. There are three levels of restrictions identified in this table:

- Permissive – indicates that the license poses minimal requirements about how the software can be distributed/used or modified. All permissive licenses in the summary were approved by the Copyfree initiative[14].
- Copyleft – the most restrictive license type assigned in the summary to the GPL license. The reason for that is that only GPL puts the obligation on the licensee that any derived and publicly distributed work needs to be released under GPL. Derivative work can be both modified version of the software as well as the software that incorporates the licensed tool. Because by definition a certain copyleft license requires derivative work to be licensed under the very same license, it is incompatible with other copyleft licenses. This is why, in practice, GPL licensed code cannot be incorporated in a distributed software tool that is not licensed with GPL.
- Partial copyleft – it is a type of copyleft license that is less restrictive and allows licensed software to be used in software licensed differently. In other words, software tools that are released under partial copyleft license can be used in other software tools which are licensed using a different license (but still some restrictions may apply, e.g. in the case of the Mozilla Public License (MPL), the source code licensed with partial copyleft license needs to

[14] See: http://copyfree.org/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

be released under the same license, but other components can be licensed with different licenses).

The last column in Table 1 provides additional notes on each license. One of the especially interesting remarks is GPL compatibility, meaning whether the tool licensed with a certain license can be used in a software tool released using GPL; the most popular copyleft license for software. Not all licenses are compatible with GPL, especially those posing more restrictions on usage than GPL. GPL is not compatible with any other license, except GPL itself.

License	Level of Restriction	Notes
Apache License v.2.0	Partial copyleft	Compatible with GPL v.3
3-Clause BSD license[15]	Permissive	Compatible with GPL v.3
2-Clause BSD license[16]	Permissive	Compatible with GPL v.3
GNU General Public License v3 (GPL)	Copyleft	The most popular copyleft license for software
GNU Library or "Lesser" General Public License v.3 (LGPL)	Partial copyleft	Compatible with GPL v.3; modified source-code need to be released under the same license
Affero GPL v3	Copyleft	Compatible with GPL v3
MIT license	Permissive	Compatible with GPL v.3
Mozilla Public License 2.0 (MPL)	Partial copyleft	Compatbile with GPL v.3.0; source code files which are under MPL need to remain under MPL
Common Development and Distribution License	Partial copyleft	Incompatible with GPL v.3.0; due to additional restrictions on notes in the source code it is incompatible with GPL
Eclipse Public License v.1.0	Partial copyleft	Incompatible with GPL v.3.0; it is more restrictive than GPL in the context of patent retaliation thus not compatible with GPL

Table 1: List of popular licenses

[15] Also known as New/Revised/Modified BSD license

[16] Also known as FreeBSD or Simplified BSD license

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2.3.1 Compatibility/proliferation of licenses

Compatibility of licenses is one of the biggest troubles that tool developers face when releasing software. It is important to remember that when one develops a new tool it is relatively common to use various software libraries to support particular functions, e.g. logging, mathematical computations, linguistic analysis, etc. It is convenient to use software libraries that are open source. Unfortunately, one can only choose from those software libraries that are compatible with each other in terms of licenses. For example if tool A is licensed using Apache License v.1.1 and one wants to use it in tool B which will be licensed using GPL then it is legally impossible as the licenses are incompatible. This is why various licenses have been enhanced and modified to make them more compatible with each other. Take for example the Apache License version 2.0. It is compatible with GPL, but only one-way. That means that software components released with Apache License v.2.0 can be used in GPL licensed tools, but not vice versa. This is because GPL is a copyleft license and requires derivative works to be licensed only with the use of GPL. The common name for this problem is license proliferation. Although the problem still exists, especially in large tools consisting of many components, it is presently reduced by various recommendations and suggestions from the developer community. An example is Google Developers platform, which restricts the types of licenses that can be used in projects and also recommends the GPL or Apache License. To have deeper understanding of the differences, similarities and compatibility issues between the most common licenses, a short summary of each is provided in the following subsections. The intention is to characterise the newest versions of the licenses, therefore only those versions are considered in the descriptions.

2.3.2 GNU Public License (GPL)

The GPL license is one of the most popular licenses used for releasing FLOSS software. The main characteristic of this license is that it is a strong copyleft license. That means that if a tool is GPL licensed, then any derivative work needs to be licensed with GPL as well. One of the key issues with the GPL license is the interpretation of the term derivative work. There is still debate whether linking a GPL-licensed program to another one (using as a static or dynamic library) yields a derivative work or not[17]. For the time being the safe approach (same as interpretation of Free Software Foundation -

[17] See: https://www.linuxjournal.com/article/6366

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

FSF) is to assume it is a derivative work, although courts under a certain jurisdiction might decide differently. Because GPL is strong copyleft, it is not compatible with licenses that pose similar restrictions (e.g. Mozilla Public License). On the other hand, all permissive licenses are compatible with GPL (e.g. the MIT License). Because GPL is a popular license, some licenses (which are still in use, attract new developers and gain synergy) were modified to become compatible with GPL (e.g. Apache License). The current version of GPL is 3.0 which was released in 2007. GPL v.3.0 is recommended by the FSF.

2.3.3 GNU Library or Lesser General Public License (LGPL)

LGPL was created with a special focus on software libraries and it was created as a compromise between the strong copyleft GPL license and the more permissive ones like the MIT License. The main idea is that the software licensed with LGPL can be used in (linked with) another software tool. The new software can be licensed with a different license, including a proprietary one. Nevertheless, LGPL is not fully permissive since all derivative work from the LGPL-licensed software needs to be released with the same license. LGPL is recommended by the FSF for special cases only, such as when the functionality of a library is already available in other software libraries licensed with a more permissive approach. In such situations there is no reason to apply GPL, as it will limit the number of users of the specific software library (because there are other libraries released with permissive approach, meaning that proprietary tools can use them).

2.3.4 Affero General Public License (AGPL)

GPL pertains to software that is publicly distributed. For products that incorporate GPL licensed components but that are used privately there are no requirements. In the time of the Internet that leaves a loophole for derivative works that are not distributed as standalone tools but are made available as online tools. The AGPL aims to fill that gap. AGPL v3 is compatible with GPL v3.0, with one additional requirement: if you run a modified program on a server and let other users interact with it, your server must also allow them to download the source code corresponding to the modified version running on that server.

2.3.5 Apache License

The Apache License is a free software license developed by the Apache Software Foundation and initially based on BSD License. The license is commonly known as a permissive one because a modified

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

version of the Apache licensed software can be released using a different license and the licensed software can be used in proprietary tools. Nevertheless, the Apache License requires that every file that has not been modified is licensed using the original Apache License and also that special notes need to be present in the modified files. Apache License v2.0 is the current one. It is compatible with GPL v.3.0, meaning that software licensed with Apache License can be used in tools released with GPL v.3.0 (but not vice versa). FSF recommends Apache License v2.0 when there is a necessity to use a non-copyleft license (permissive).

2.3.6 Berkeley Software Distribution License (BSD)

The BSD License is one of the best-known permissive licenses. It has three main versions, but only two are approved by the Open Source Initiative (OSI) and FSF. The two approved versions are 3-Clause BSD License and 2-Clause BSD License. Initially the BSD License had 4 main clauses posing restrictions. Due to one of those clauses (so-called advertising clause) OSI rejected it. In 1999 the advertising clause was removed and a so-called New/Modified/Revised BSD License was created. There is also a 2-Clause BSD License (also known as FreeBSD license), which omits the so-called non-endorsement clause. Both the 2-Clause and 3-Clause are GPL compatible. For the BSD License it is important to compare the different variants and decide carefully which to choose.

2.3.7 MIT License

The MIT License is a permissive license created by the Massachusetts Institute of Technology. It is sometimes called the X11 License as it was designed for the X Window System. The MIT License is similar to the 2-Clause BSD License. It is compatible with the GPL license.

2.3.8 Mozilla Public License (MPL)

The Mozilla Public License is maintained by the Mozilla Foundation. It is a weak copyleft license, somewhere between the Apache License v2.0 and the GNU General Public License. MPL licensed software can be used by a differently licensed tool, but a modified version of the software needs to be released with MPL license. The current version is MPL 2.0 and this version is compatible with GPL (in contrast to prior versions).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2.3.9 Eclipse Public License (EPL)

The Eclipse Public License is maintained by the Eclipse Foundation. It is a weak copyleft license that is weaker than GNU General Public License. In case of EPL, additions and modifications to the software can be licensed with a different license only if these cannot be considered to be a derivative work. In case of derivative work, the software needs to be licensed with EPL and it should be made available to everyone. The current version 2.0 of the EPL is compatible with GPL if a special option is selected. Otherwise it is incompatible because the GPL is too restrictive.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3 Project requirements

In this chapter we outline the legal provisions that are required within the Elexis project.

We will only consider deliverables since these will be made available to communities outside the consortium. We presume that all IPR (Intellectual Property Rights) matters with respect to products that are shared within the consortium are covered by the Consortium Agreement.

3.1 Software deliverables

The project is due to publish a number of software tools: D1.3, D1.4, D2.2, D2.4, D2.5, D3.1, D3.3, D4.1, D4.2, D4.3, D4.4. In the project proposal it is mentioned that the software deliverables will be published as open source. As is explained in Section 2.3, there are several licensing schemes available for open source software. Software tools are rarely developed from scratch. They usually are built on existing software or use existing components. As was also explained in Section 2.3.1, the license for the new tool may depend on the licenses that rest on the building blocks that have been used in building it.

3.1.1 What do we need for building them?

Software is rarely built from scratch. Most likely the software tools will be derived from existing software. The tools will mainly be used for data processing. Therefore we need data to test and calibrate the tools and possibly to develop data models that will be used by the tools. The type of data required for building tools will be decided by the groups responsible for these tools.

3.1.2 Who will have access to the product?

The software will be made available to the general public.

3.2 Data deliverables

The proposal mentions two data deliverables:

D4.5, **Sample Dictionary Drafts:** Sample dictionaries drafted for consortium languages using the D4.2 module which will be publicly available for reviewing the methodology by the community.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

D4.6, **Semantically Annotated Corpora:** Provision of text corpora (mostly based on online texts) with semantic annotation, not older than 3 years for consortium languages (and other EU languages, including lesser resourced).

Beside that, two important datasets that are at the core to the project are the Dictionary Matrix and the Matrix Dictionary[18].

3.2.1 What data do we need?

Deliverables D4.5 and D4.6 will probably not require background data.

One of the main data deliverables is the Dictionary Matrix which provides extensive links between key structural elements found in different types of dictionaries. Therefore, the focus is on (direct or indirect) linking of existing lexicographic resources, on the lemma level and other "simpler" levels, but also on the sense level by pivoting through one of the existing semantic resources – BabelNet. Ultimately, linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in different types of existing lexicographic resources, monolingual, multilingual, modern, historical etc. For the Dictionary Matrix we need digital dictionary content from monolingual dictionaries in multiple languages. These dictionaries are provided as background data by project members. According to the Consortium agreement, the members are committed to provide their data for sharing within the consortium, but not for sharing beyond that. Apart from the members of the consortium, the project aims at drawing in Observing Institutions which will be asked to contribute data to the project.

Figure 1 shows the model for the ELEXIS services. The main public facing datasets are the Dictionaries (TEI Lex-0, Ontolex) and Dictionary Matrix (Links). The Dictionary Matrix contains foreground data, while the Dictionary component contains mainly background data. The latter will at a later stage also contain foreground data in the form of user generated content.

[18] See http://ihjj.hr/mreznik/page/simon-krek/21

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

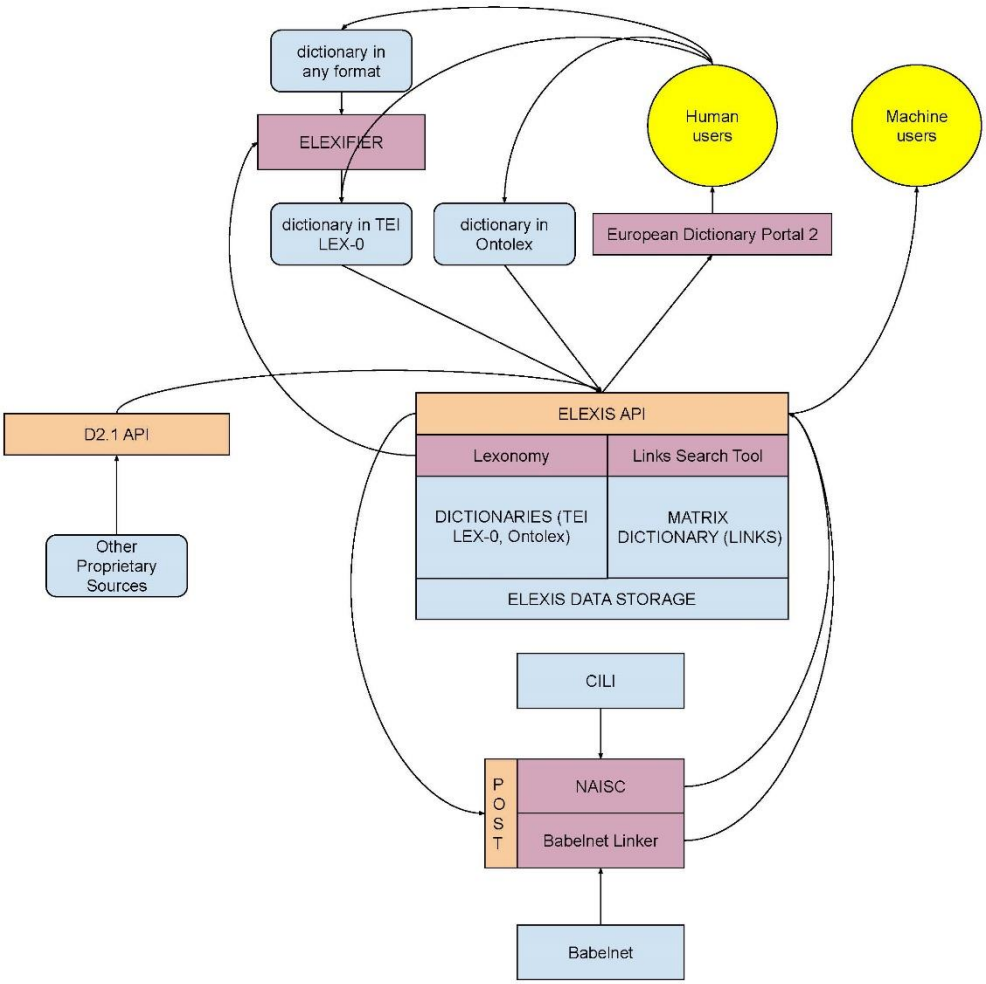

Figure 1: Model for ELEXIS services.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4 Recommendations

4.1 Licensing software

ELEXIS will be an active contributor to open-source software development, additionally encouraging technology uptake as a consequence. The pre-existing tools that partners/observers will bring into the project are mostly open-source and they will be further developed in the project.

What is needed is a recommended public license for the software developed within the project. In light of the aim of the project to empower the community as much as possible, FLOSS licenses seem to be the most appropriate option. Of course, there are many FLOSS licenses. Of the available options, GNU Public License v3.0, GNU Library or Lesser General Public License v.3.0 and Apache License v.2.0 are the most widely used and should be recommended. In case a software tool uses building blocks that carry a license that is in conflict with the default license mentioned above, the tool can be published with a different, but open, license.

4.2 Licensing data for developing tools.

Using data to train tools within the project seems to be the least problematic, provided that it will be impossible to extract a significant part of the original data from the tool and if the tool does not provide services that are in direct competition with the services of the data owner. The licensing conditions for this purpose are in fact laid down in the Consortium agreement for the members of the consortium. For the Observer institutions, however, there should be a license drawn up between the consortium and the owner that describes clearly what data is involved, what use will be made of the data, by whom it will be used and what the result will be.

Since the data will only be used by members of the consortium, an end-user license seems to be the best option. We could opt for the CLARIN-DELA-RES-2014-10 license[19].

[19] See: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/ClarinSA/CLARIN-DELA-RES-2014-10.rtf

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4.3 Licensing Dictionary Content.

There will be two classes of content in the Dictionary component: Background data provided by consortium partners and observer institutions, and content contributed by users in the context of crowdsourcing programs.

4.3.1 Licensing Background Dictionary Content.

The data of the Dictionary Matrix will be shared as Linguistic Linked Open Data (LLOD). The licensing structure used by LLOD and mentioned in the proposal is Creative Commons Attribution-Non Commercial-Share Alike 3.0 License.

The Dictionary component (see Figure 1) contains the dictionary content that is served when users query the Dictionary Matrix.

Many members of the consortium have placed restrictions on sharing the data beyond the consortium as can be seen in Attachment 1 of the Consortium Agreement. The same will probably hold for the Observer institutions. Since not all participants will be eager to publish their background data under an open license, we will have to opt for a more diverse licensing plan.

Owners of a set of dictionary data can choose an open access license of their liking. It might be of importance to build in safeguards into the service architecture that will prevent users from downloading complete or large data sets from the Dictionaries component (see Figure 1). Moreover, when a user is shown dictionary content, they should be informed about the rights that rests on the data and the appropriate attribution of the license of the dictionary requires that.

Note that it is possible to allow to link to 'Other Proprietary Resources' (see Figure 1) that are not included in the Dictionaries component. These links, when queried, will direct the user to data that will be served up at the proprietary websites of the owners.

The owners of background data can decide how much data they want to commit to the Dictionaries component and with which license. Minimally, the headword list with part-of-speech information is required to develop the links for the Dictionary Matrix.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4.3.2 Licensing Crowdsourced Dictionary Content.

It is important to be mindful of potential intellectual property (IP) issues that can surround crowdsourced data.

In the first place it is important to be aware of possible third party rights that rests on the provided content. There should be provisions and procedures in place to deal with such matters. One could consider creating a service where abuse of proprietary data can be reported. Furthermore, it might be useful to only allow registered users to contribute data and record which user has contributed which part (e.g. Wikipedia style). If rogue data has been reported, it might be sensible to inspect all other contributions of the user.

Secondly, there should be appropriate Terms and Conditions (T&C) that define the rights and obligations of the contributors and the ELEXIS consortium and future proprietor of the ELEXIS services. These terms and conditions should be strict enough to allow the ELEXIS organization to use the data for the intended purposes while also be acceptable to participants. The more aggressive the legal approach is towards managing intellectual property, the less likely the crowd is willing to participate[20].

Below we will list some components that could be part of the T&C[21].

- Should the right to the data be transferred to the ELEXIS organization? This does not seem to be necessary if the data is distributed under an open license. The creator remains the owner. Recommendation: no.
- Should it be exclusively? This is not recommendable since it might deter contributors and it is not required for the purpose of the ELEXIS organization. Recommendation: no.
- Should there be attribution? Since multiple parties will provide data to the Dictionaries component and different licenses might apply, all data parts must be linked to the source and owners. From this perspective it seems useful to retain links between crowdsourced content and the contributors. Recommendation: yes.

[20] De Beer, J., I. P. McCarthy, A. Soliman, E. Treen (2017); Click here to agree: Managing intellectual property when crowdsourcing solutions; https://doi.org/10.1016/j.bushor.2016.11.002

[21] See: Ball, A. (2014). 'How to License Research Data'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

- Should the data be provided with an open license? This does not seem to be necessary, since the dictionary content is not intended for distribution, only for consultation. On the other hand, it might also be useful to allow other parties to use the crowdsources content in combination with the Dictionary Matrix while it is not likely to deter participants. Recommendation: yes.
- Should it include a clause prohibiting that third party IP is submitted? Recommendation: yes.

A good starting point for drafting the T@C are the GitHub Terms of Service[22].

[22] https://help.github.com/en/articles/github-terms-of-service

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

5 GDPR

5.1 Overview

Since May 12th 2018 the EU General Data Protection Regulation (GDPR) has entered into force for the whole of the European Union. That means that all member states of the EU have the same regulations in place.

The regulation details requirements for organizations that store or manage personal data.

Personal data are data that are connected to an individual or that can identify an individual. Examples are: name, picture, telephone number, address, bank account number, e-mail address, IP-address, finger print, etc.

The principles of proper data management are:

- *Transparency*. The person whose data is processed is informed about that, has given explicit consent, and has been informed about her/his rights.
- *Purpose limitation*. The data can only be used for the legitimate purpose explicitly specified to the data subject.
- *Data minimization*. Only data can be collected that is absolutely necessary for the purposes specified.
- *Accuracy*. The data must be kept accurate and up to date.
- *Storage limitation*. Data can only be stored for as long as is necessary for the specified purposes.
- *Integrity and confidentiality*. Data must be processed in such a way as to ensure appropriate security, integrity, and confidentiality.
- *Accountability*. The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

5.2 GDPR issues within ELEXIS.

In the light of the above we have to make an inventory of what personal data is stored and processed within ELEXIS.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

First, we have the personal data of the team members that have been collected by the project management. Although this data is for internal use only, the principles mentioned above should be respected.

Second, ELEXIS is tasked to organise and conduct training activities (WP5). This will require management of personal data from those that enroll in these activities. The same principles of proper data management as described in 6.1 are germane here.

Finally, there is the data that will be released as public deliverables. The main ingredient of those deliverables (Dictionary Matrix and Matrix Dictionary) will be dictionary content. This content will probably contain personal information; think of names of authors whose works have been cited and quotations from public figures. In principle, a person whose name is mentioned in the public deliverables can demand that the information is removed from the source. While that is very unlikely to happen, there is probably also an exemption to the prohibition where the person has manifestly made that information public. The latter, however, has not been tested in court yet.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

6 Survey on licensing

In order to obtain an overview of current licensing practices of main producers and providers of lexicographic content, a survey was conducted among ELEXIS partner and observer institutions. The aim of the survey was to gain an insight into the licensing situation at different institutions, and common and concerns problems with licensing and sharing lexicographic data. The survey was conducted in the final months of 2019, also to ensure the topicality of its findings.

6.1 Respondent institutions

The survey was completed by representatives of 38 different ELEXIS partner and observer institutions (Table 2). The majority of institutions were based in Europe.

Country	Frequency
Bulgaria	3
International (UK/US)	3
Italy	3
Spain	3
Croatia	2
Denmark	2
Germany	2
Lithuania	2
Slovenia	2
Sweden	2
Czech Republic	1
Estonia	1
Finland	1
Iceland	1
Ireland	1
Israel	1
Latvia	1
Netherlands	1
North Macedonia	1
Norway	1
Portugal	1
Romania	1
Russia	1
Switzerland	1

Table 2: Respondent organisations by country

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Almost half of the institutions (18) were public, and 13 of them were universities or university departments. Four responding institutions were non-profit organizations, and two were private companies. Two of the institutions reported to be a mixture of public and private.

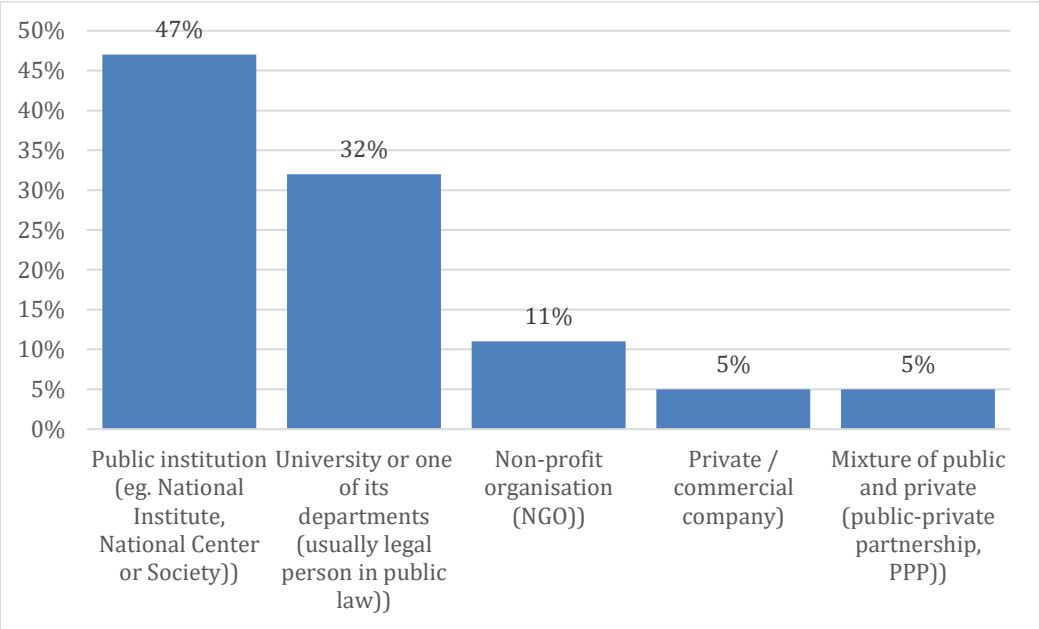

Figure 2: Institutions by type

6.2 Funding

Most of the institutions (28) reported using public funding at the national level for the creation of their lexicographic resources. In most cases, the source of funding is the ministry responsible for science, or a research agency/council. Therefore, lexicographic resources seem to be mostly considered by governments as a scientific undertaking. Some institutions combine national funding with their internal funding or private funding.

Nine institutions reported using public funding at the international level, either solely or in combination with other sources, to create their lexicographic resources. The funding sources included H2020 funding, Marie Curie and ERC grants, European social fund, and the European regional development fund. Other types of public funding, used by seven institutions, included specific regional funding, small-scale project funding, or scholarships.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

More than third of the institutions (13) reported using private funding to create their lexicographic resources. The funding sources include sponsorship (by companies, foundations), collaboration with private companies such as publishers, and institution's own funds. One institution reported investing profit from investments into stock market and real estate into lexicographic resources.

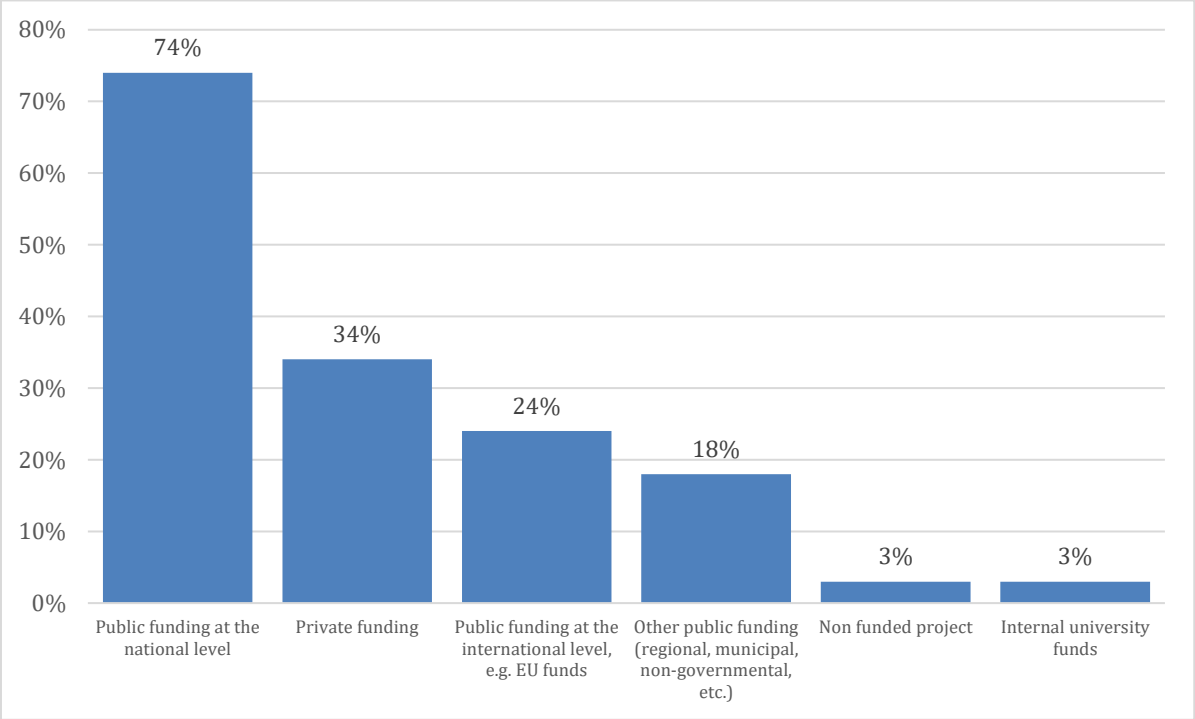

Figure 3: Funding of lexicographic resources (multiple answers were possible) (N=38).

6.3 Providing access to lexicographic resources

84% of the institutions offer their lexicographic resources online for free, and only five offer (some of) their lexicographic resources online for a fee. On the other hand, there are still many institutions that publish paper dictionaries, over half of the respondent institutionss in fact (20 out of 38). Six institutions reported using a publication model where they publish the paper version first, and provide the online version after some time for free. Among the six institutions, four are public, one is non-profit organisation and one is a mixture of public and private. A few institutions also mentioned providing lexicographic resources via mobile apps.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Type of access	No. of institutions
online (free access)	32
online (paid access)	2
online (some resources for free, others for a fee)	3
paper dictionary	20
paper dictionary first, then after some time online for free (e.g. after 1 year)	6
other	11

Table 3: Type of access to lexicographic resources (multiple answers were possible)

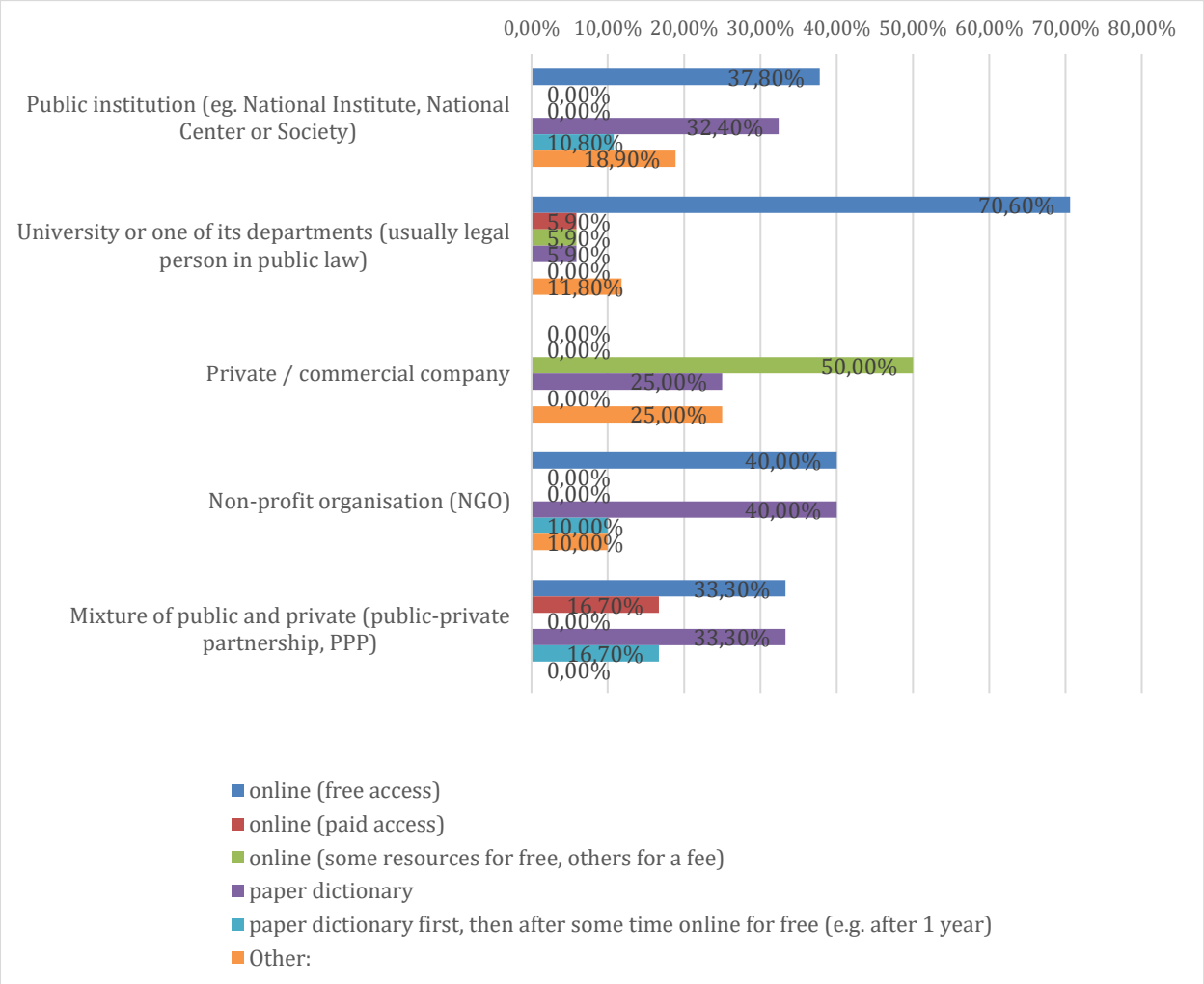

Figure 4: Availability of lexicographic resources based on organization type

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

A high number of institutions make their lexicographic data available for reuse by others, with majority of them (18) offering free download of the data, which is provided under a certain license. Only two institutions reported charging for download of their data. 17 institutions reported making their data available via API; 13 institutions offering free API access and 4 institutions paid API access. Seven institutions reported of offering customised preparation of datasets to interested parties on request. It is important to point out that when explaining their answers several institutions reported making only certain parts of their lexicographic resources available to others (e.g. headword lists, lists of typical misspellings), and/or introducing usage limits by number of (API) calls or amounts of data. Paid API access is thus used as an additional service to free access, for example for substantial usage, or for using lexicographic data for commercial purposes. A few institutions mentioned under Other that they were in the process of setting API access.

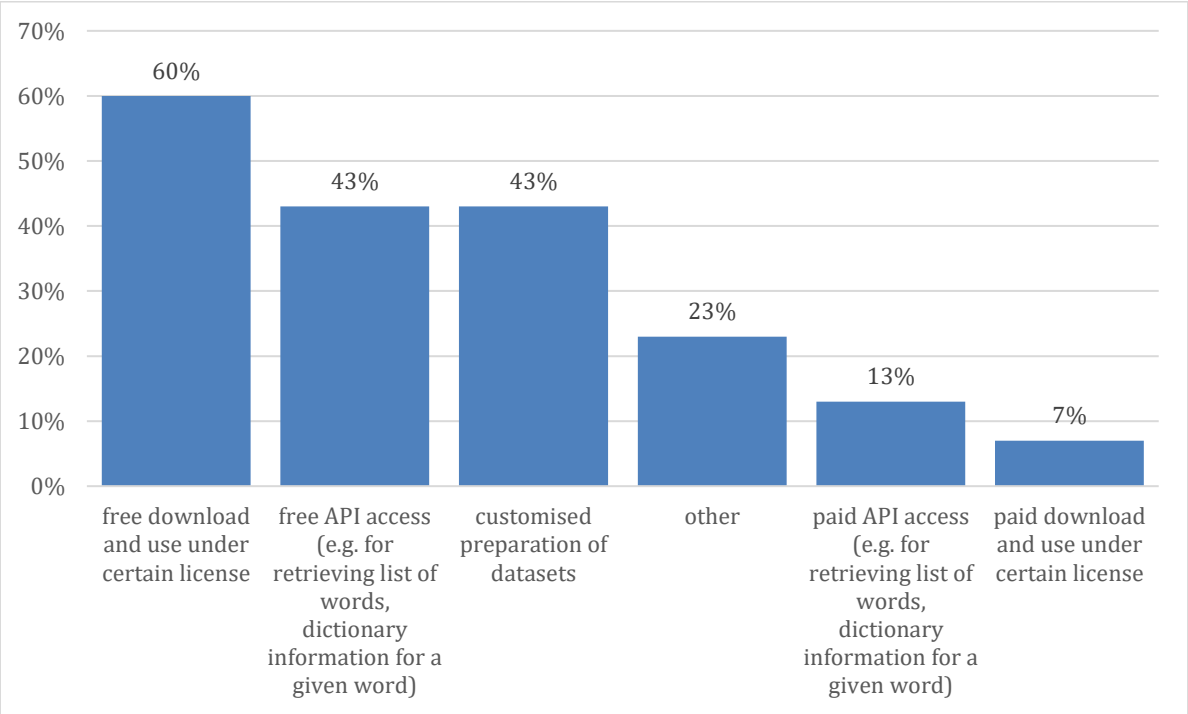

Figure 5: Availability of lexicographic data for reuse

Among the institutions that offer free download under certain license, many use CC-BY or CC-BY-SA for their lexicographic resources. Several institutions use CLARIN repositories for making their data available (and other data, in case they are CLARIN centres).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Customized services are quite common, with 13 institutions reporting they offer them. The customers are researchers or companies, and individual cases mentioned range from preparing lemma lists with selected information from entries, lists of commonly misspelled words, audio files and images etc.

6.4 Licensing lexicographic data

Majority of institutions (80.5%) reported having a cleared IPR status for their lexicographic data, with 16 institutions having a cleared IPR status for all their lexicographic data, and 13 institutions only for some. The reasons for not having the IPR status cleared vary from and still trying to negotiate the agreement with the authors of the resources, lack of time, and considering the data as not interesting for external parties. On the other hand, seven institutions reported not having a cleared IPR status for their lexicographic data; many were in the process of sorting it out. It is noteworthy that out of those seven institutions, six receive public funding for their lexicographic resources.

	Frequency	Percent (%)
Yes.	16	44.4
For some of the datasets only.	13	36.1
No.	7	19.4

Table 4: Do you have a cleared IPR status for all your lexicographic data? (N=36)

Only nine institutions provided details on the copyright holders of their data. In most cases, the institutions themselves are the copyright holders, with exceptions mainly being limited to particular datasets where the copyright holders are the authors (either working at the institutions or externally).

Out of ten institutions that commented on how difficult was to obtain the copyright clearance most said that it was easy; this was related to the fact that they compiled the resources themselves and did not need any external clearance. It should be noted that almost all of these institutions use open licence for their data. The problems mentioned by some of the institutions were linked to bureaucracy involved, vague initial contracts with the authors, and having numerous linked resources.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

As far as the advice or recommendations on obtaining copyright clearance goes, it was often mentioned that the copyright status of the datasets should be made clear at the beginning of the project. In terms of dealing with copyright holders, especially individual authors, the advice is to sometimes use less jargon. Finally, one institution also stressed the importance of having a data management plan.

Only seven institutions reported making their lexicographic data available through brokers; two used ELRA, one META-SHARE, and one ELRC-share. Three reported using CLARIN repository.

Creative Commons is used as a standard licensing schema for lexicographic data by majority of institutions (86%; N=22). CLARIN licensing framework and Open Data Commons are used by only few institutions, five and one respectively. A few institutions mentioned that they choose licensing schema on a case-to-case basis.

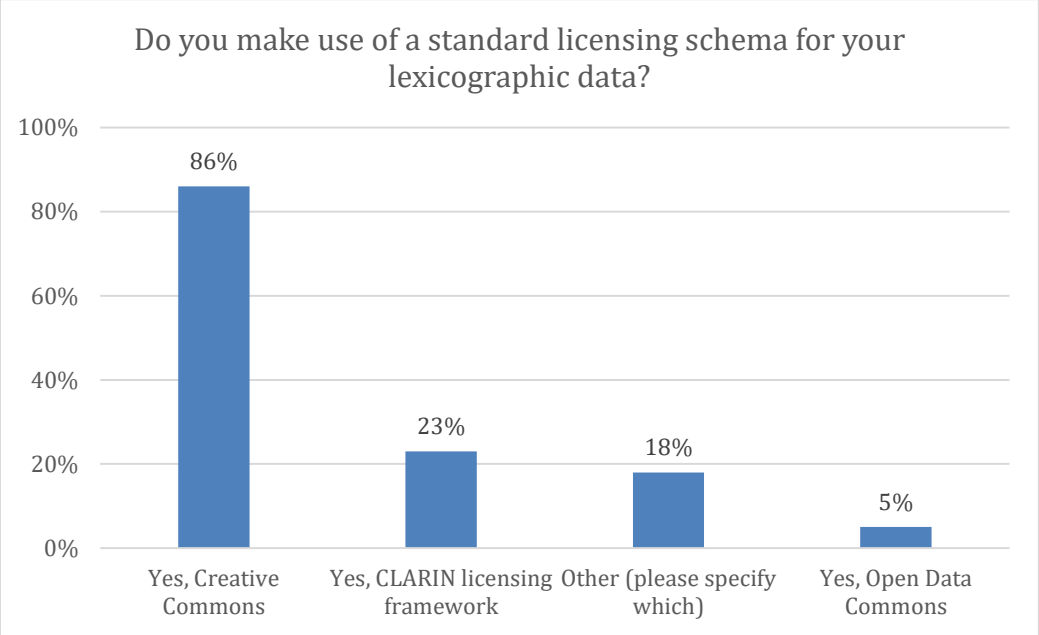

Figure 6: Standard licensing schemas used by institutions for lexicographic data (N=22, multiple answers were possible)

When asked about different types of lexicographic data, most institutions reported to be willing to share, under different licenses, lemma lists (28 institutions). Many institutions would also be willing to share examples (23), synonyms (22), sense structure (22), morphological information (22), definitions (21), collocations (20), fixed expressions (19), frequency information (19), and syntactic information (18). Fewer institutions reported willingness to share etymological information (14),

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

pronunciation information (12), and frequently misspelled word forms of lemmas (9). It must be noted that lower number of institutions at certain types of data is also linked to the fact they do not have such types of data. Nonetheless, the aforementioned three types of lexicographic data with the lowest number of institutions being willing to share them also exhibit the highest percentages of institutions selecting the "would not share" option.

As shown in Table 5, the most frequently selected license for nearly all the data types was public (open) data, followed by restricted (non-commercial) license. Taking the ratio between these two licenses into account, the institutions seem to be more protective of frequently misspelled word forms of lemmas, definitions, synonyms, and collocations. Several institutions commented on the problematic or unclear status of corpus examples, Academic license and commercial license were selected by significantly smaller percentages of institutions, even smaller than "would not share" option. Some institutions are willing to share (some) types of their data with other organizations only by using specially prepared contracts between institutions.

It is interesting to note that if non-applicable answers are excluded (as they mean that such types of data are not made or available at the institutions), there are 13 institutions that reported offering all their available types of data under public (open) access. Perhaps a good “advice” to others is voiced by one of the respondents: “A culture of data sharing across institutions is still to come. According to many, now (generational change, end of paper-based publishing) could be the moment for an initiative.”

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

		Public (open) data	Academic License	Restricted (non-commercial) license	Commercial license	Would not share	Not applicable	**Total**
lemma list (+part-of-speech)	N	18	4	5	1	1	2	**31**
	%	58 %	13 %	16 %	3 %	3 %	6 %	**100 %**
sense structure: sense numbers (main sense / subsense) for each lemma, without any other data	N	11	4	6	1	3	6	**31**
	%	35 %	13 %	19 %	3 %	10 %	19 %	**100 %**
definitions	N	11	2	7	1	3	7	**31**
	%	35 %	6 %	23 %	3 %	10 %	23 %	**100 %**
examples	N	13	3	6	1	3	5	**31**
	%	42 %	10 %	19 %	3 %	10 %	16 %	**100 %**
fixed expressions (e.g. idioms, phraseology)	N	11	2	5	1	3	9	**31**
	%	35 %	6 %	16 %	3 %	10 %	29 %	**100 %**
collocations	N	11	2	6	1	3	7	**30**
	%	37 %	7 %	20 %	3 %	10 %	23 %	**100 %**
synonyms	N	13	2	7	1	2	6	**31**
	%	42 %	6 %	23 %	3 %	6 %	19 %	**100 %**
morphological information (conjugations, inflections)	N	14	2	5	1	2	7	**31**
	%	45 %	6 %	16 %	3 %	6 %	23 %	**100 %**
syntactic information (valency patterns)	N	11	1	5	1	3	10	**31**
	%	35 %	3 %	16 %	3 %	10 %	32 %	**100 %**
frequently misspelled word forms of lemmas	N	3	1	4	1	2	20	**31**
	%	10 %	3 %	13 %	3 %	6 %	65 %	**100 %**
frequency information (e.g. lemma lists with corpus frequency or frequency rank)	N	13	0	5	1	2	10	**31**
	%	42 %	0 %	16 %	3 %	6 %	32 %	**100 %**
pronunciation information	N	9	1	2	0	3	16	**31**
	%	29 %	3 %	6 %	0 %	10 %	52 %	**100 %**
etymological information	N	8	1	4	1	4	13	**31**
	%	26 %	3 %	13 %	3 %	13 %	42 %	**100 %**
Other:	N	4	0	1	0	0	7	**12**
	%	33 %	0	8 %	0 %	0 %	58 %	**100 %**

Table 5: Based on IPR status of your data, which types of lexicographic data would you be willing to share or are already sharing (licensed)?

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

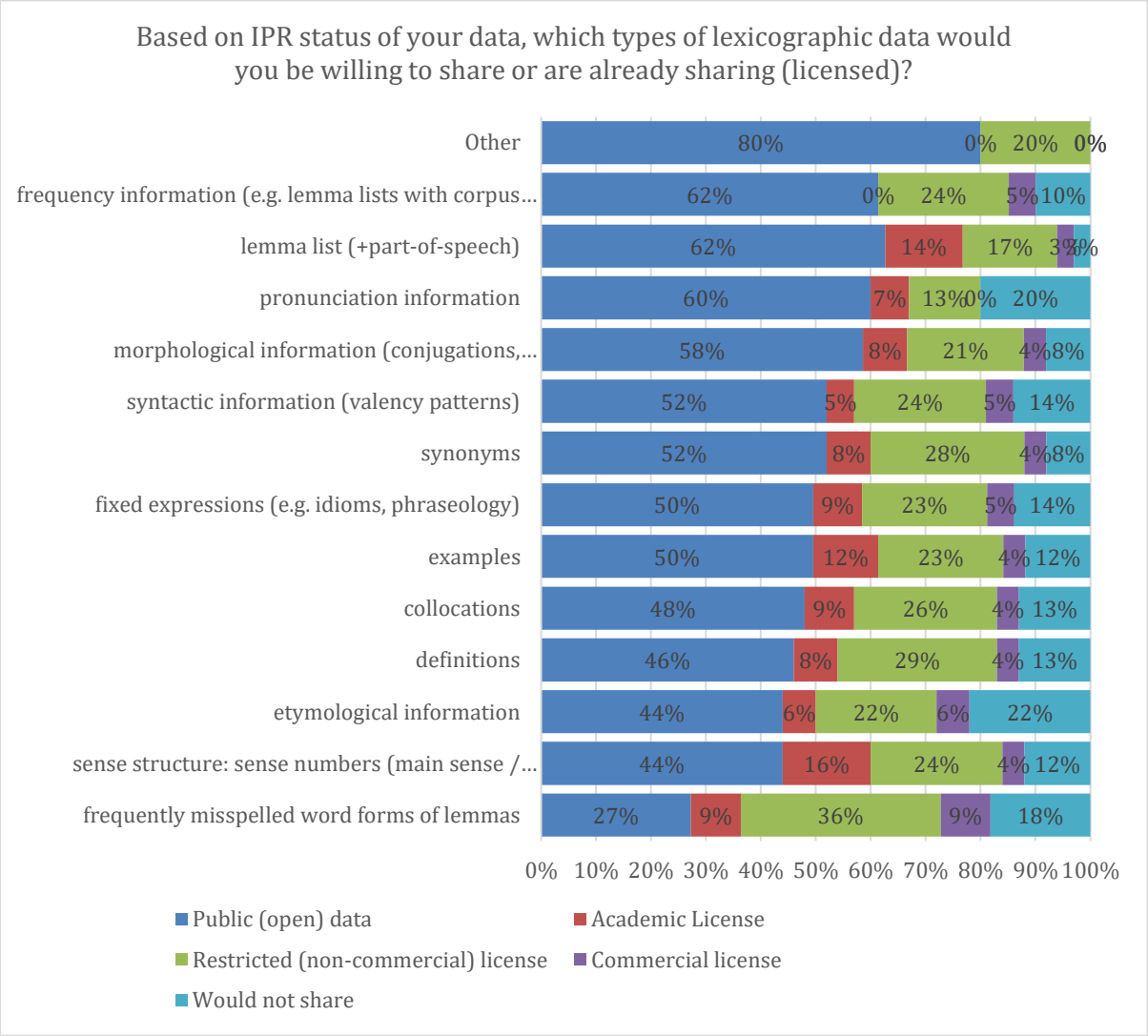

Figure 7: Licenses used by institutions for different types of data

The main concerns institutions have with regards to sharing their data can be summarized as follows:

- Commercial use of their data, especially by competitors. The concerns are especially connected with producing low quality products for profit generation only.
- Unclear status of data because they were obtained from corpora with licensing restrictions.
- Lack of standardized documentation for sharing lexicographic data.
- Misuse by others, e.g. use beyond the purposes allowed by the license. Also, misuse may result in breach of contract with data provider, e.g. when making a corpus.
- Fear of someone beating them to analyses or source preparation.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

15 out of 31 institutions keep track of the use of their datasets, other 16 reported they do not. The use of datasets is monitored in one of three ways: by requiring users to register, by asking users to report on how they used the data, or by monitoring API use.

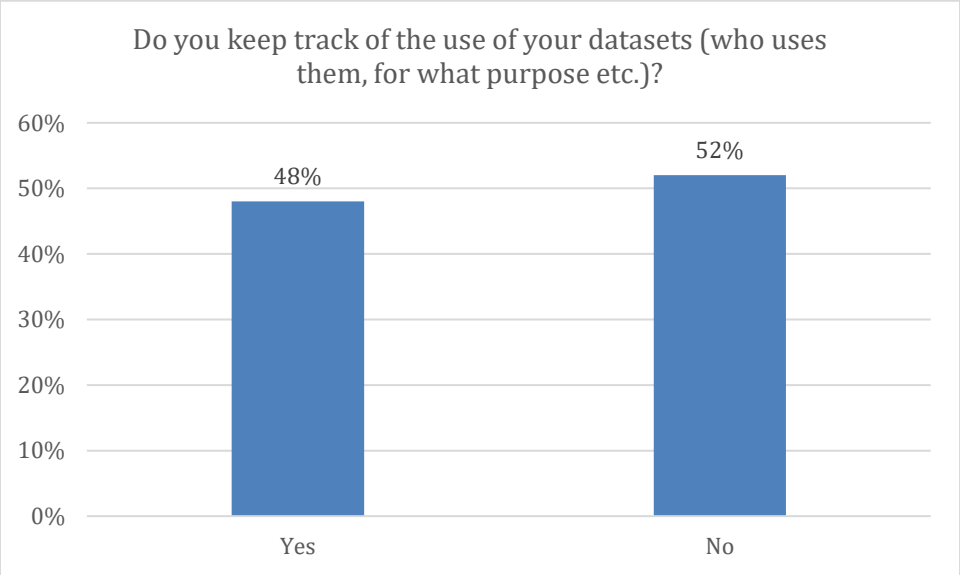

Figure 8: Tracking the use of datasets (N=31)

63% of the institutions (20 out of 32) reported having a policy on the distribution of lexicographic data. Five of those institutions reported having the policy publicly available, whereas others have an internal policy only. On the other hand, 12 institutions reported not having such a policy (public or internal). It is also noteworthy that seven out of ten universities (70%) that answered this question do not have a policy on the distribution of lexicographic data. The percentage of public institutions (e.g. institutes) without such a policy is significantly lower (30%).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

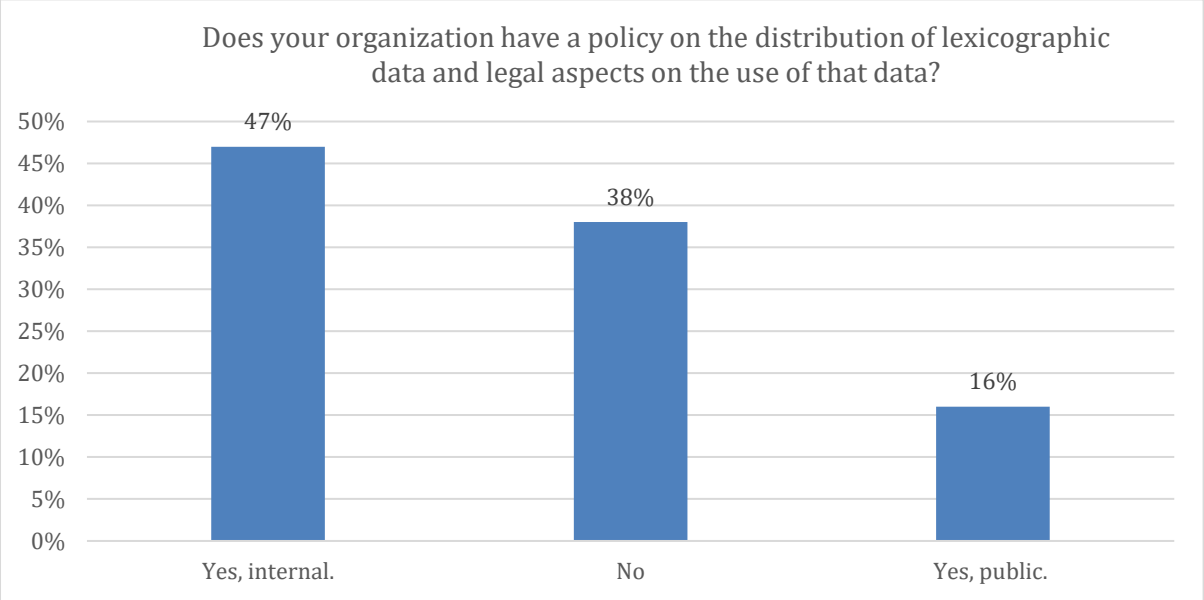

Figure 9: Institutional policies on the distribution of lexicographic data (N=32)

Exactly half of the 32 institutions that answered Question 33 reported having a special person or department dealing with IPR issues. A closer look at the further explanations of the answers, however, reveals that none of the institutions have a person or department that specialised solely in IPR. 15 out of 32 institutions have a legal department or legal expert dealing with all legal issues, and two more institutions reported on hiring a legal expert when necessary. At five institutions, a non-legal person - such as deputy director, manager of language resources or head of the centre - deals with IPR issues. One institution reports on benefits of being in the national CLARIN consortium as the CLARIN department deals with all IPR issues for them.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

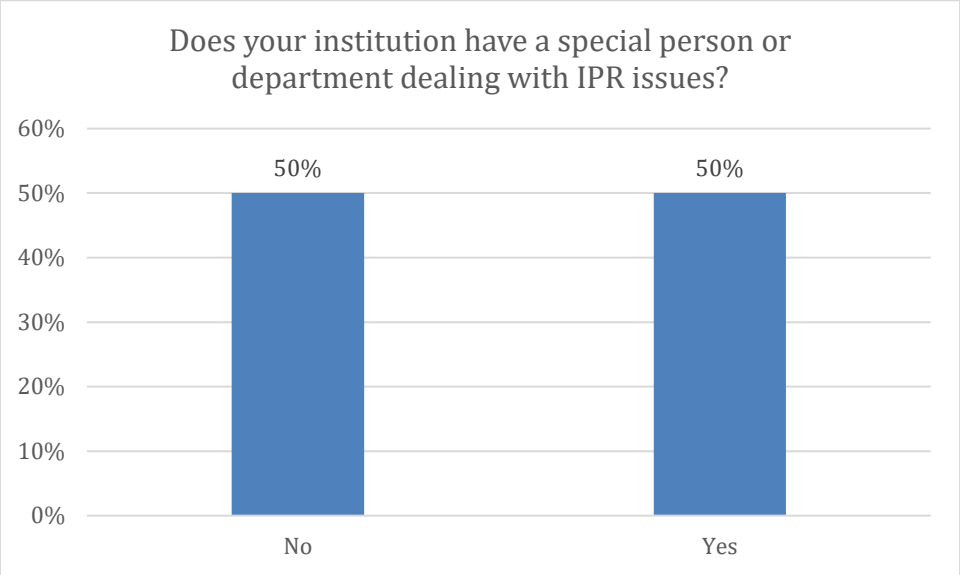

Figure 10: Availability of legal support for dealing with IPR issues (N=32)

Majority of the responding institutions (90%) have never taken legal action related to the use of their lexicographic data, indicating that this is rare in lexicographic world. One institution reported on reaching a settlement (in court) related to some of their dictionaries, and the other on the fact that they conduct surveillance of possible bad practices or illegal use and take action when necessary. The third institution reports on a case of forensic linguistic analysis of various bilingual dictionaries, which was conducted to determine their originality, i.e. to assess the possibility of theft of intellectual property.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.