EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

License to use: ELEXIS survey on licensing lexicographic data and software

Kosem I., Nimb S., Tiberius C., Boelhouwer B., Krek S.

Jožef Stefan Institute, Slovenia, Det Danske Sprog- og Litteraturselskab, Denmark, Dutch Language Institute the Netherlands
iztok.kosem@ijs.si, sn@dsl.dk, Carole.Tiberius@ivdnt.nl, Bob.Boelhouwer@ivdnt.nl

Abstract

Lexicographic resources are extremely valuable, not only for the general public but also for other applications, such as natural language processing, linked open data, etc. As many resources are still not available or are only available under very strict conditions, it is important to understand their owners' or creators' stance towards data sharing. This is particularly relevant for the European Lexicographic Infrastructure (ELEXIS) project, which has as one of its main aims the development of the Dictionary Matrix that will be formed of extensive links between key elements found in different types of dictionaries. This paper reports on a survey on licensing lexicographic data conducted amongst partner and observer institutions in ELEXIS. The results show that there are many differences on how institutions in different countries approach data licensing. Moreover, the differences can be observed at the level of dictionary microstructure, as institutions are more protective towards certain types of lexicographic data. Using a case study, it is demonstrated how a more open approach to sharing data can benefit the community of a particular language, and the ELEXIS community in general.

Keywords: licensing; survey; data; ELEXIS; Dictionary Matrix; lexicographic resources; dictionary; corpus

1 Introduction

The creation of a dictionary of quality requires a large amount of highly skilled labour. Therefore, such a product is of high value to its creators, sponsors, and users. The owners or compilers of dictionaries will want to protect their data, but for different reasons. If the owner is a private organisation, the reason will probably lie in the commercial value of the data. Public organisations, on the other hand, will have other reasons, such as the need to prove their relevance to the funding provider by reporting visits to their dictionary website. Furthermore, certain organisations may not be allowed by their funding provider or governing organisation to hand the data to others. The quality of the data may also be a consideration; some organisations 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 organisation's reputation.

As a result, general (open) access to lexicographic data is still extremely limited, which prohibits reuse of valuable datasets in other fields, such as natural language processing, linked open data and the Semantic Web, as well as in the context of digital humanities. One of the main objectives of the European Lexicographic Infrastructure (ELEXIS) is to address these issues, i.e. to enable access to lexicographic data and to promote an open access culture in lexicography, in line with the *European Commission Recommendation on access to and preservation of scientific information*. Serious efforts have been made within ELEXIS to address these Intellectual Property Rights issues currently preventing the inclusion of lexicographic data into open access infrastructures. In the context of this work, a survey was conducted among partner and observer institutions in order to get information on, and understanding of, their existing licensing practices.

This paper presents some of the main findings of the survey, both from the perspective of current practices and situations at different types of institutions, and in terms of their future plans and concerns. Also, a case study of licensing practices at one of the ELEXIS partner institutions is presented in more detail. The paper concludes by presenting the licensing options that the ELEXIS consortium prepared for partners/observers in order to facilitate the content sharing process, and to obtain enough content for the Dictionary Matrix.

2 European Lexicographic infrastructure (ELEXIS)

ELEXIS (Krek et al. 2018, 2019; Pedersen et al. 2018; Woldrich et al. 2020) is a Horizon 2020 project dedicated to creating a sustainable infrastructure for lexicography. The main objectives of ELEXIS are to (1) enable efficient access to high quality lexicographic data so that it can also be used by other fields including NLP, AI and digital humanities, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources.

To realise these goals, ELEXIS has an inclusive multi-layered organisation that aims at engaging different user groups with various levels of intensity during the project, shown in Figure 1. The core of the organisational structure consists of 17 consortium partners. The consortium is composed of content-holding institutions and researchers with complementary backgrounds: lexicography, digital humanities, standardisation, language technology, Semantic Web and artificial intelligence. Furthermore, the consortium cooperates with existing infrastructures, i.e. CLARIN and DARIAH.

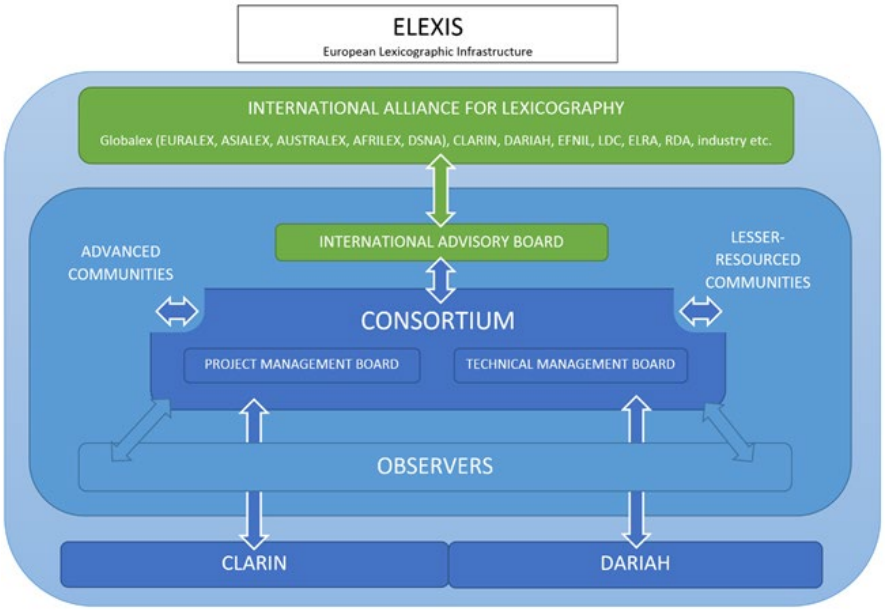

Figure 1: ELEXIS organisational structure.

Another organisational layer is formed by observer institutions that are directly included in outreach and dissemination activities through various channels. The central group of institutions that fall under the observer category are those producing quality lexicographic data and resources. Typically, but not exclusively, these institutions include (European) national language institutes, large dictionary publishers and other prominent producers of lexicographic data. At the time of writing, ELEXIS had 58 observers.

One of the main data deliverables of ELEXIS will be the Dictionary Matrix that will be formed of extensive links between key elements found in different types of dictionaries - monolingual, multilingual, modern, historical etc. With the Dictionary Matrix, ELEXIS is creating a universal lexicographic metastructure spanning across languages and time. The focus is on (direct or indirect) linking of existing lexicographic resources, minimally on the headword and part of speech level, but where possible also on the level of senses, examples, translations, collocations, and other types of information in lexicographic resources. The Dictionary Matrix will be available as a public service, and the links between dictionary elements will be shared as Linguistic Linked Open Data (LLOD) enabling other fields to exploit the high-quality semantic data from lexicographic resources.

Given that obtaining enough datasets from partner and observer institutions is a key prerequisite for the Dictionary Matrix, and licensing lexicographic content is one of the important topics of discussions between consortium partners, as well as between partners and observers, a survey was conducted among partner and observer institutions in order to gain an insight into their existing licensing practices. The aim of the survey was not only to gain an insight into the licensing situation at different institutions, but also to get an idea of common concerns and problems connected with licensing and sharing lexicographic data.

3 Survey on licensing practices of ELEXIS institutions

The survey was conducted in the final months of 2019. We used the 1ka survey system,[1] which was previously used for several other surveys conducted by members of the team, in ELEXIS as well as in other projects. The survey consisted of 36 questions, 12 of them were multiple-choice and the rest open-ended. Several open-ended questions (many of them optional) were used in order to offer the respondents a possibility to elaborate on their answers. The average time of survey completion was approximately 10 minutes, which was less than expected but probably the consequence of the fact that due to the specific nature of the questions, the respondents had to gather the information in advance, and then enter it into the survey.

The survey was completed by 38 ELEXIS partner and observer institutions from 25 different countries, predominantly from Europe. Almost half of the institutions (18) were public, and 13 of them were universities or university departments. Four responding institutions were non-profit organisations, and two were private companies. Two of the institutions reported to be a mixture of public and private.

3.1 Source of funding

Most of the institutions (28) reported using public funding at the national level for the creation of their lexicographic

[1] https://www.1ka.si/

resources. In most cases, the reported source of funding was the ministry responsible for science, or a research agency/council. Some of the institutions also reported combining national funding with their internal funding or private funding.

Nine institutions reported using public funding at the international level to create their lexicographic resources, either as the only source of funding or in combination with other sources. The reported 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.

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

3.2 Intellectual property rights (IPR)

29 institutions (80.5%) reported having a cleared IPR status for their lexicographic data, with 16 institutions having cleared IPR status for all their lexicographic data, and 13 institutions only for some. The reasons for not having the IPR status cleared varied: 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.

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 (formerly) employed or collaborating externally by contract). Out of ten institutions that commented on how difficult it 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 licenses for their data. The problems mentioned by some of the institutions were bureaucracy, vague initial contracts with the authors, and the connections between resources (derivatives etc.).

Half of the institutions (N=32) reported having a special person or department dealing with IPR issues. A closer look at further explanations of the answers, however, revealed that none of the institutions had a person or department that specialised solely in IPR. 15 out of 32 institutions had a legal department or a legal expert dealing with all legal issues, and two more institutions reported 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 reported on the benefits of being in the national CLARIN consortium as the CLARIN department deals with all IPR issues for them.

3.3 Distribution of lexicographic data

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 did not have a policy on the distribution of lexicographic data. The percentage of public institutions without such a policy is significantly lower (30%).

We asked the institutions on how they make their data available, either as content to the language users for consultation purposes, or as datasets. Moreover, we were interested in getting a better understanding of which types of data they are willing or not willing to share. The results are presented in the following subsections.

3.3.1 Current status of data availability

On the question of how they make their lexicographic resources available to the language users, 84% of the institutions reported offering their lexicographic resources to the users online for free, which is in line with the fact that most of the institutions are publicly funded. Only five institutions reported offering (some of) their lexicographic resources online for a fee. Interestingly, many institutions reported publishing paper dictionaries, over half of the respondents in fact (20 out of 38).[2] Six institutions (four public, one non-profit and one a mixture of public and private) reported using a publication model where they publish the paper version first, and provide the online version after some time for free. It is noteworthy that universities or university departments seemed to be more oriented toward (free) online access (Figure 2), whereas public institutions and non-profit organisations exhibited a similar balance of online and paper format.

[2] A similar finding was observed in the survey of user needs (Kallas et al. 2019; Kallas et al. 2020), where it was stated that the reason for still publishing print dictionaries was tradition, i.e. the previous volumes of a dictionaries being also published in print.

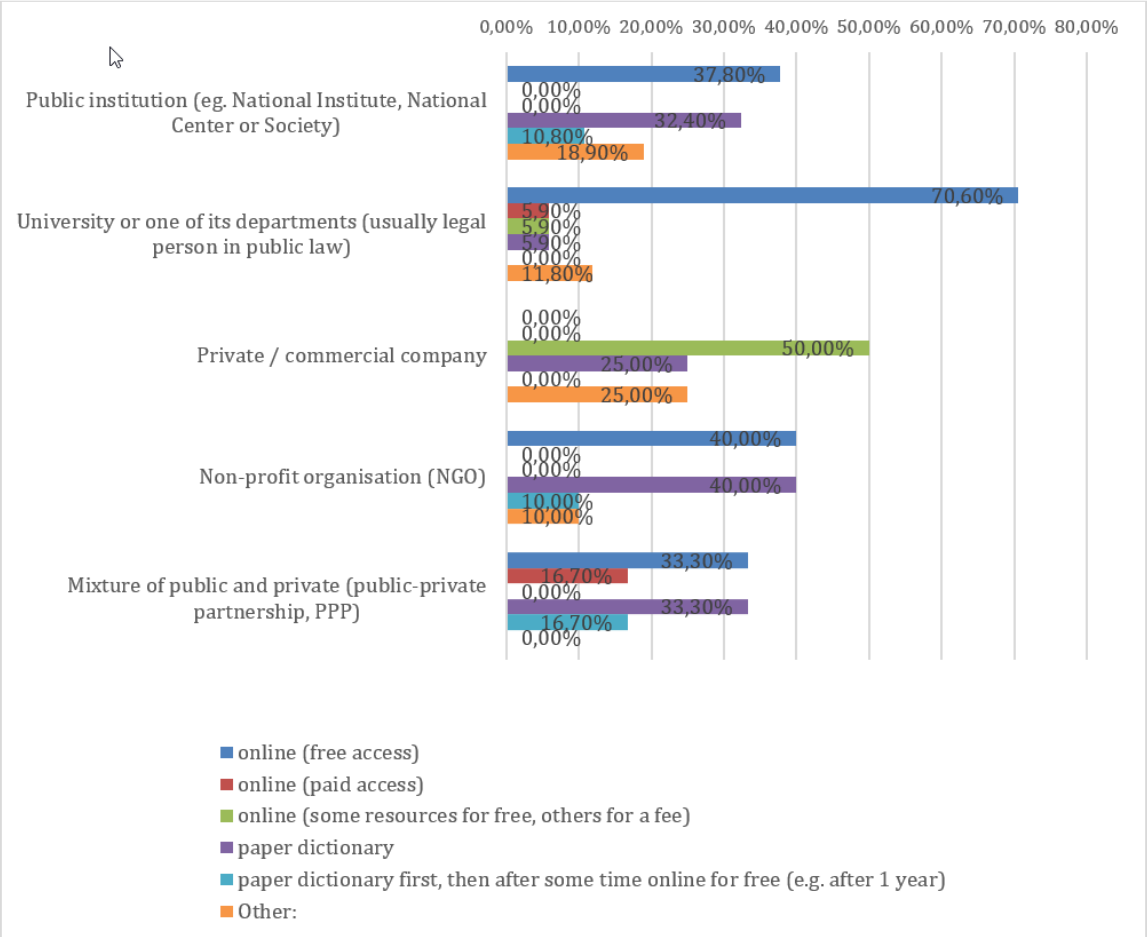

Figure 2: Availability of lexicographic resources by type of institution.

In terms of datasets, a high number of institutions reported making their lexicographic data available for reuse by others, with the majority of them (18) offering free download of the data. Only two institutions reported charging for download of their data. Creative Commons (most often CC-BY or CC-BY-SA) was used as a standard licensing schema for lexicographic data by the majority of institutions (86%; N=22). CLARIN licensing framework and Open Data Commons were used by only a few institutions, five and one respectively. A few institutions mentioned that they choose a licensing schema on a case-to-case basis.

17 institutions reported making their data available via API; 13 institutions were offering free API access and 4 institutions paid API access. A few institutions mentioned they were in the process of setting up API access. 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 to number of requests 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.

Customized services seem to be used often, with 13 institutions reporting they offer them. The customers are researchers or companies, and individual cases mentioned ranged from preparing lemma/headword lists with selected information from entries, lists of commonly misspelled words, audio files and images etc.

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

15 out of 31 institutions reported keeping track of the use of their datasets, the 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 the API use.

3.3.2 Willingness to share different types of data

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), pronunciation information (12), and frequently misspelled word forms of lemmas (9). It must be noted that lower numbers of institutions at certain types of data are also linked to the fact that some institutions 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 Figure 3, 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 a significantly smaller portion of the institutions, even smaller than the portion under the "would not share" option. Some institutions are willing to share (some) types of their data with other organisations 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 were 13 institutions that reported offering all their available types of data under public (open) access.

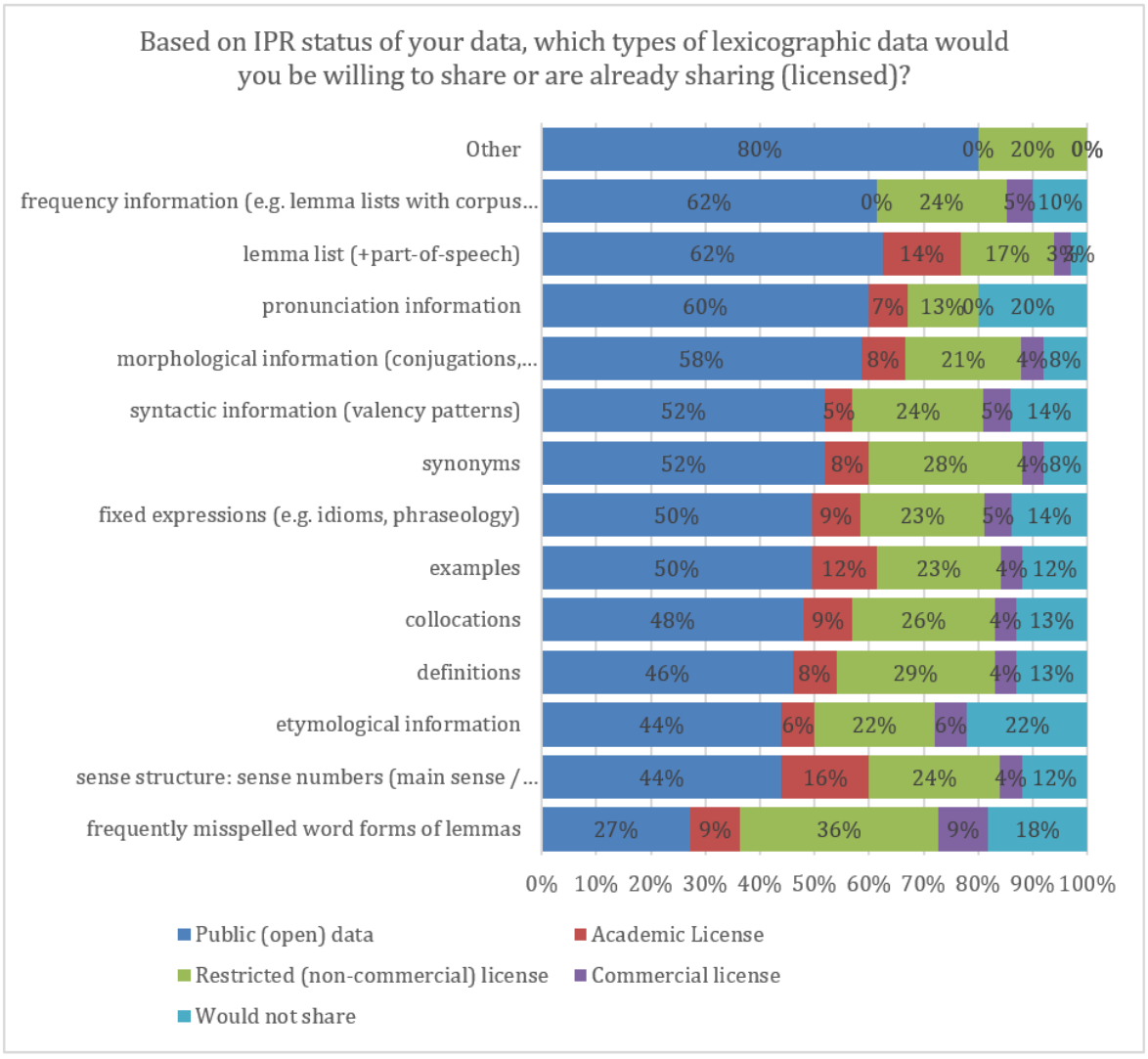

Figure 3: Sharing different types of lexicographic data.

3.3.3 Concerns about data sharing

The main concerns institutions have about sharing their data can be divided into three groups. Firstly, many institutions expressed concerns about how the data might be used by others, in particular they pointed to:

- Commercial use of their data, especially by competitors. The concerns are especially connected with producing low quality products for profit generation only.
- Misuse by others, e.g. use beyond the purposes allowed by the license. Also, misuse may result in breach of contract with data providers, e.g. when making a corpus.
- Fear of someone beating them to analysis or source preparation.

Then, there were concerns about the unclear status of their data because they were obtained from corpora with licensing restrictions. Finally, some institutions pointed to the lack of standardized documentation for sharing lexicographic data.

Despite various concerns about data sharing, the majority of the responding institutions (90%) had never taken legal action related to the use of their lexicographic data, indicating that this seems to be rare in the lexicographic world, especially as far as publicly funded institutions are concerned. 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 reported on a case of forensic linguistic analysis of various bilingual dictionaries, which had been conducted to determine their originality, i.e. to assess the possibility of theft of intellectual property.

3.4 Case study - DSL: How to approach licensing of lexicographic content

DSL (the Society for Danish Language and Literature) is one of the two partners in ELEXIS that reported to be a mixture of public and private organisation. When the ELEXIS project was initiated in 2018, DSL did not have cleared IPR status for their lexicographic data. The society has edited and published monolingual Danish dictionaries since 1911, and several of these projects have throughout the years been funded partly by the Danish Ministry of Culture, and partly by private funding, i.e. the Carlsberg Foundation. Since 2009, DSL has published the Danish dictionary DDO (Den Danske Ordbog) freely online, 4 years after the last volume of the printed DDO dictionary was released for sale. Today, the DDO dictionary, which is well known in the Danish society, is also freely available via API, but only for non-commercial purposes and under special (time limited) agreement.

Over the last decade, there have been an increasing number of inquiries regarding the spin-off data from the dictionary compilation process, e.g. lemma lists and frequency lists. These data were often given out for free for non-profit purposes, i.e. research or study projects, but always as stand-alone resources, that is without internal links between different types of data. No specific person at DSL was responsible for dealing with copyright issues concerning the digital lexicographic data. A legal advisor is affiliated with the society, however before the focus had been on copyright issues regarding publications in print. Moreover, the personal views of the DDO dictionary editors represent a variety of different opinions from being rather open, e.g. wanting to share any type of lexicographic data for research purposes, especially to partners in a research project, to the completely opposite position, i.e. wanting to keep the data in-house in order to guarantee the future lexicographic business of DSL. One major reason for not wanting to share the data is fear of abuse, e.g. use beyond the purposes allowed by the license, and that misuse, e.g. of corpus citations, may result in the breach of contracts with the data providers. DSL has quite restricted agreements with several key text providers which have been delivering texts for more than 25 years - texts that are typically only allowed to be used for internal corpus investigations at DSL and as citations in the DDO dictionary. However, it was unclear whether the providers would really have the right to protest if the citations in the dictionary, many of them dating back to the 1990's, were used for other purposes, e.g. for research.

In 2018, at the same time as the ELEXIS project started, the Danish Ministry of Culture set up a language technology committee, which included a DSL representative. The purpose was to clarify the major problems preventing Danish language technologies from being developed in line with the English language technology industry. The Danish datasets and lexical resources that existed at the time were described in a concluding report in 2019, including under which conditions they could be of benefit to the private business community. One of the main conclusions of the report was that a large, open source resource, which would integrate the existing lexical data from major Danish language institutions, was very much needed in order to facilitate the development of language technology products for Danish (Kirchmeier et al. 2020).

Combined with the participation in the survey on licensing practices of ELEXIS institutions described in this paper where many questions could not be answered clearly, it became clear that DSL needed a detailed policy regarding the sharing of more complex and integrated lexicographic resources, both to be able to fulfil the data requirements in the ELEXIS project and to be able to play a role in the development of an open source consolidated resource for Danish. Moreover, with such a policy, DSL would be able to answer the growing number of requests for the computational lexicographic data developed at DSL in a more homogeneous and standardised way.

3.4.1 Applying the ELEXIS survey for internal purposes

In order to produce a more specific policy in line with the opinion of the DDO editors, we carried out an internal survey. We reused most of the questions from the ELEXIS survey reported in this paper. The editors were asked individually about their views on the sharing of different types of data in the DDO dictionary. While answering the questions, they were allowed to consult the general results from the ELEXIS survey in Figure 3. They were for example able to see that two thirds of the lexicographic institutions involved in ELEXIS were willing to share lemma as well as frequency lists as open source data, and that most of the institutions also considered it acceptable to share synonyms, valency information and morphological information. The hypothesis was that the insight into which type of data most lexicographic institutions were in fact willing to share would probably influence the views of the editors.

The results of the internal survey showed that the views of the editors are, to a much higher degree than expected, in line with the views of the majority of the ELEXIS partner and observer institutions; in fact, in some cases the DDO editors were willing to share even more data, e.g. data on misspellings and sense structure. However, definitions and etymologies were considered by the DDO editors to be authored data, which cannot be shared, except for research purposes in projects where DSL is a partner, or against a substantial payment. The editors were also willing to share the manually selected citation examples in DDO, but as already mentioned this is a more complicated matter due to copyright issues.

The results of the internal survey led to several actions at DSL. Firstly, the legal consultant was involved in order to clarify the more complicated case of citation examples. He guaranteed that it would be legal to hand over the example citations to a third party as long as they remained integrated in other lexical data (e.g. in a sense structure); this, however,

would not hold true if examples were extracted separately, for example for the purpose of creating a stand-alone text corpus. Secondly, a number of lexicographic datasets, which were formerly given for free by DSL only if the specific (non-commercial) purpose stated by the user was acceptable, were made directly available for download online at korpus.dsl.dk. Today, the only restriction on the data is that it must not be used to publish an independent dictionary that is in competition with DSL's products. The interested parties must accept this condition before making the download. Furthermore, the website for downloads has been improved, and the number of open source lexical resources that can be directly downloaded has increased, now also including lists of synonyms and common misspellings. Also, a new strategy on customized services was confirmed by the DSL board: namely, DSL decided to contribute to the development of technologies for the Danish language by allowing users (including companies) to request special datasets based on DDO as long as they cover DSL's payroll expenses for the preparation of the dataset, and under the condition that the dataset is afterwards made freely available on the DSL website korpus.dsl.dk.

Furthermore, the clarification of the editors' standpoints and the knowledge on which data types they are willing to share has led to a new, nationally funded project. In 2020, it significantly facilitated the work on preparing the project application, and in March 2021, a three-year project with the University of Copenhagen, the Danish Language Council and the Danish Agency for Digitisation was approved. DSL contribution lies in a high amount of already linked lexicographic data consisting of lemmas, morphology, misspellings, synonyms, sense structure, and examples from DDO, as well as semantic links to the Danish WordNet which were developed from (and at the same time linked to) DDO data in the DanNet project 2004-2010 (Pedersen et al. 2009).

Last, but not least, the clarification of the editors' views and the general DSL policy also facilitated DSL's contribution to the collection of lexicographic data in the ELEXIS project. Accordingly, the DDO data that DSL submitted consists of linked lexicographic information, while authored text, e.g. full definitions, was not included. Still, a sample of approximately 5,000 lemmas with full sense descriptions has been provided for research purposes.

4 Conclusions

As the survey among ELEXIS partners and observers has shown, there are many differences across different countries, without any clear patterns of, for example, similar practices by similar types of institutions. While there are many institutions that promote open access for all, or nearly all, of their data, there are on the other hand still several institutions that are very protective of their data. Interestingly, there seem to be different levels of concern for different types of lexicographic data, with definitions, examples, synonyms, collocations, and frequently misspelled forms of lemmas being the most protected.

The majority of the institutions reported using Creative Commons licensing schema, while few others reported (also) using the CLARIN licensing framework or Open Data Commons. In addition, just under 50% of the institutions reported keeping track of their datasets, and 50% of the institutions reported having a special person dealing with intellectual property rights issues. Given the concerns reported and the importance of licensing mentioned by the institutions, this percentage can be considered quite low.

The case study of DSL clearly shows that raising awareness on the benefits of opening the data, and sharing experiences and opinions on data sharing can lead to important changes in the approach to licensing lexicographic content. We can report that other institutions are adopting a similar approach, opening their datasets more as compared to their status in the ELEXIS agreement. To date, 118 different datasets, e.g. general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), and lemma lists have been collected from 32 ELEXIS partner and observer institutions. A sample list of the datasets can be found in the ELEXIS Deliverable 6.3 Intermediate interoperability report (Kosem et al. 2021).

These datasets will be used for linking purposes in creating the Dictionary Matrix. Since some of the data contributors are still very protective of their data, the ELEXIS consortium has come up with a number of flexible and diverse licensing options to encourage the institutions to contribute their data (or parts of it) to the Dictionary Matrix:

1. The owners of background data can decide how much data they want to contribute to the Dictionary Matrix. They have the option to contribute entire dictionaries or only parts of them, e.g. full entries for a certain letter only, or only certain elements for all entries. Minimally, a headword list with part-of-speech information is required to develop the links for the Dictionary Matrix.
2. Owners of a set of dictionary data can choose an (open access) license of their liking.
3. Dictionary content that will be presented to users will be accompanied by the information on the rights that rest on the data, as well as the appropriate attribution if the chosen license for the dictionary requires that.
4. If the owners of dictionary content still hesitate to share their data directly, it will be possible to link from the Dictionary Matrix to external resources. These links, when queried, will direct the users to the data that will be served up at the proprietary websites of the owners.

This licensing scheme is the result of a special task in ELEXIS which was dedicated to providing guidelines and solutions for handling copyright and authorship protection, resulting in a deliverable entitled *Recommendations on legal and IPR issues for lexicography* (Boelhouwer et al. 2020).

The process of making lexical resources more openly available has already started in the lexicographic community, but more promotion and raising awareness is needed. As voiced by one of the respondents of the survey: "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."