elexis european lexicographic infrastructure

D1.1

LEXICOGRAPHIC PRACTICES IN EUROPE: A SURVEY OF USER NEEDS

Author(s): Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius

Date: 11. 6. 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D1.1 LEXICOGRAPHIC PRACTICES IN EUROPE: A SURVEY OF USER NEEDS

Deliverable Number:	D1.1
Dissemination Level:	Public
Delivery Date:	31. 1. 2019
Version:	1.1
Author(s):	Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius

elexis european lexicographic infrastructure

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

Deliverable/Document Information

Deliverable No.:	D1.1
Deliverable Title:	Lexicographic Practices in Europe: A Survey of User Needs
Author(s):	Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius
Dissemination level:	Public

Document History

Version Date	Changes/Approval	Author(s)/Approved by
V0.1 15/11/2018	First Draft	Jelena Kallas, Svetla Koeva, Margit Langemets, Carole Tiberius
V0.2 10/12/2018	First Draft and Review	Iztok Kosem
V0.4 15/12/2018	Incorporation of Feedback	Jelena Kallas, Svetla Koeva, Margit Langemets, Carole Tiberius
V0.5 25/12/2018	Review	Miloš Jakubíček
V0.6 10/1/2019	Incorporation of Feedback	Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius
V0.7 15/1/2019	Review	Simon Krek
V1.0 31/1/2019	Final version	Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius
V1.1 11/6/2020	Revised version	Jelena Kallas, Svetla Koeva, Iztok Kosem, Margit Langemets, Carole Tiberius

Table of Contents

List of Tables

List of Figures

Glossary

AI – Artificial Intelligence
API – application programming interface
CMDI – Component MetaData Infrastructure
CQS - Corpus Query System
CSV – Comma-separated values (text format)
DTD – Document Type Definition
DWS - Dictionary Writing System
ENeL - European Network of e-Lexicography
GDEX – Good Dictionary Examples (tool)
GOLD – General Ontology of Linguistics Descriptions
GUI – Graphical User Interface
IPR – Intellectual property
L1 – First language
L2 – Second language
LMF – Lexical Markup Framework
LOD – Linked Open Data
N – number of respondents
NLP – Natural Language Processing
OCR – optical character recognition
OLAC – Open Language Archives Community
Q – question in the survey
RDF – Resource Description Framework
SVN – the Subversion file format
TEI – Text Encoding Initiative
TSV – Tab-separated values (text format)
URL – Uniform Resource Locator
UX – User Experience
XSL – Extensible Stylesheet Language
XSLT – Extensible Stylesheet Language Transformations

1 Introduction

This deliverable presents the results of task T1.1 User Needs. The aim of this task was to generate an overview of lexicographic practices across Europe both for born-digital and retrodigitised resources. This is particularly important as the lexicographic landscape in Europe is currently rather heterogeneous. On the one hand, it is characterised by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts (as cooperation on a larger European scale has long been limited). On the other hand, there is a significant variation in the level of expertise and resources available to lexicographers across Europe.

To obtain an overview of lexicographic practises, two surveys have been carried out focussing on different aspects of the lexicographic workflow (e.g. software and tools, publication, retrodigitisation, metadata and data formats).

The results of the surveys provide an insight in what is needed by lexicographers and lexicographic institutions in terms of tools, functionalities and training. As such the results feed back into the ELEXIS project, especially into WP4 “NLP for Lexicography” and WP5 “Training and Education”.

The work of task T1.1 built on the results of the COST action European Network of e-Lexicography (ENeL)[1]. The Aim of the ENeL COST Action was to increase, coordinate and harmonise European research in the field of e-lexicography and to make authoritative dictionary information on the languages of Europe easily accessible, which resulted in the European Dictionary Portal[2]. The COST Action took place between 2013 and 2017 and, in the course of these four years, grew into a congregation of 30 participating countries with more than 280 members, successfully uniting lexicographers in Europe.

Within the Action, several workshops and surveys took place, some of which are particularly relevant in the context of ELEXIS, i.e. the workshop on the Workflow of Corpus-Based Lexicography[3], the survey on Dictionary Writing Systems and Corpus Query Systems[4] and the survey on the Automatic Acquisition of Lexicographic Knowledge[5]. Below, a brief summary of the results of these three surveys is given.

[1] http://www.elexicography.eu/
[2] http://dictionaryportal.eu
[3] http://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf
[4] http://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf
[5] http://www.elexicography.eu/wp-content/uploads/2015/10/ENeL_WG3_Survey-AKA4Lexicography-TiberiusHeylenKrek.pptx

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 2014 survey on the lexicographic workflow showed that overall the different projects could fit their lexicographical process into the phases proposed by Klosa (2013)[6], but that it was sometimes difficult to put exact time labels on the different phases as sometimes a phase could continue without requiring full-time effort.

The project descriptions also showed that even although lexicography became more and more computer-assisted, compiling dictionaries remained a highly labour-intensive task. The general monolingual dictionaries of the 2014 study had the longest time span with an average of fourteen years. The duration of the compilation of specialised dictionaries/databases was much shorter with an average of just over three years. Of the different phases the analysis phase took the longest for all types of projects. The majority of the projects mentioned a lack of IT support at the time. This was also the case for the more computational projects mentioned under the specialised dictionaries.

The 2015 survey on Dictionary Writing Systems and Corpus Query Systems provided an insight in the use of lexicographic tools by members of the ENeL COST Action. 70% of the respondents indicated that they or their institution use some kind of specialized software to produce dictionaries, i.e. a Dictionary Writing System (DWS). Although a number of off-the-shelf systems were used, e.g. SDL Multiterm; iLex; T-LEX; IDM DPS; Protege Ontology Editor and Termeki (termbases.eu), using a customised or in-house editor was quite common. Just over 70% indicated that they also use specialised software to query a text corpus. Although Sketch Engine was the most mentioned Corpus Query System (CQS), most institutes still used and/or developed their own system (e.g. Korp, COSMAS II, BlackLab, Poliqarp). The following open-source or off-the-shelf systems were mentioned: (no) Sketch Engine, IMS Open Corpus Workbench (CWB) and Folio Views.

From the 2015 survey on the Automatic Acquisition of Lexicographic Knowledge we learnt that automatic extraction of knowledge was more and more finding its way into lexicography. Key lexicographic tasks, such as finding collocations, definitions, example sentences, translations, were more and more beginning to be transferred from humans to machines. The respondents were also quite positive about the quality of the automatically acquired data. Automatic extraction of lemma lists, frequency information, example sentences and grammatical patterns were the most common types of automatic knowledge acquisition mentioned, whereas extraction of definitions and knowledge rich contexts were not so common. The survey showed that usually there is some sort of human intervention in this workflow. However, data such as lemma lists, frequency information, example sentences, translation equivalents and lexical-semantic relations are sometimes integrated

[6] Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herberst Ernst (eds.): Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin, Boston: de Gruyter, S. 517-524. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.4).

in a lexicographic product without any human intervention. In our analysis, the results of the ENeL surveys will be compared to the results of the 2018 ELEXIS surveys.

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 Methodology

The original idea was to carry out one European-wide survey focussing on lexicographic workflows, metadata and data formats used in lexicographic projects within Europe. However, whilst preparing the survey, it became clear that one survey could not cover all the aspects we were interested in. One of the problems was that a certain type of information could only be provided by a specific type of personnel, usually the ones with a clearer overview of project(s) and future plans, for example editors-in-chief or project leaders. Another issue was the potential length of the survey; with all the questions included, the survey would be very long, which would likely put off potential respondents, or we would get many partially completed surveys. Therefore, it was decided to conduct two separate surveys, one targeted at institutions and one targeted at individual lexicographers. To get as many responses as possible from individual lexicographers (and not just the opinion of their institutions), the survey targeted at individual lexicographers was limited in length. Having two surveys also enabled us to use different dissemination approaches, and to avoid duplication or overlap of information.

The survey targeted at institutions was more limited in terms of respondents (initially, the focus was on lexicographic partner institutions), and required a more personalised dissemination approach. We contacted the relevant people directly via email or in person at conferences. On the other hand, the survey targeted at lexicographers had to be distributed as widely as possible, through many different channels such as international and national mailing lists, social networks (e.g. ELEXIS Facebook and Twitter profiles), group or individual emails, a booth at the EURALEX 2018 conference), etc. We made a decision not to limit the survey to lexicographers in Europe, as we were also interested in lexicographic practices around the world. Nonetheless, most of our efforts when sending (personalised) reminders closer to the survey deadline were focussed on European countries with few or no respondents.

The main aim of the surveys was to get a good overview of different tools and methods used by lexicographers around Europe, as well as the needs that they have now or anticipate to have in the short-term and long-term future. It was important to get a good coverage of countries to enable comparisons, and more importantly, help us in preparing more targeted activities with the ELEXIS projects such as training workshops and materials, tools etc. Equally important was the attempt to get several respondents from the same country, in terms of institution, age, role in the team, dictionary project, etc. to ensure that the data would be representative of a country and not of a single institution, generation, project and so forth. Still, we knew from the start that this objective would be difficult to achieve, given that in several countries there are very few institutions, or just one, that compile dictionaries.

The method chosen for the surveys was an online questionnaire. Questionnaires have already proven to be a very effective and useful method of approaching the lexicographic community in the ENEL Cost Action. Several survey tools were considered for the implementation of the surveys, and

in the end Google Forms was chosen as it is simple to use and manage, and it covered the majority of our needs. Google Forms does not offer advanced analysis support, however that was not an issue as it was decided in advance to conduct the analysis in a different tool, mainly on account of a relatively high number of open-ended questions requiring manual coding and analysis.

The survey for institutions was opened on 11 July 2018. It remains open as we expect to extend it to observers as one of the steps for obtaining information about their projects, workflows and infrastructures. The survey for lexicographers was publicly announced on various mailing lists on 13 July 2018 and was closed on 1 October 2018. No more responses were accepted after that date.

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 Analysis of the results

3.1 Survey for lexicographers

The *ELEXIS Survey of Lexicographers' Needs for Lexicographers* contained 44 questions divided into 6 sections, i.e. (1) general information; (2) ongoing work; (3) software and tools; (4) publication; (5) retrodigitisation; (6) past and future. There were three different types of questions used in the survey: (1) "yes/no" questions, (2) multiple choice questions, and (3) open-ended questions. Not all questions were obligatory.

The survey was completed by 159 lexicographers, both across and outside Europe. As some questions were optional, not all questions were answered by each respondent. For this reason, we provide the number of responses for each question (i.e. N = number_of_responses) in our analysis. Next to each title we also provide the number of the question in the survey for lexicographers (e.g. Q3, Q25-27). These numbers relate to the questions in the survey which can be found in Appendix 1.

3.1.1 General information (Q2)

159 respondents came from a total of 45 countries, comprising of 36 European countries (140 respondents, Table 1) and 9 countries outside Europe (19 respondents, Table 2). We decided to categorise under European countries also countries with close cultural ties to Europe (and inclusive status in EU-funded initiatives such as COST Actions) and countries with active partners in the ELEXIS consortium.

COUNTRY	NO. OF RESPONDENTS	COUNTRY	NO. OF RESPONDENTS
Albania	1	Italy	3
Austria	2	Latvia	2
Basque Country	2	Lithuania	1
Belgium	1	North Macedonia	2
Bulgaria	6	Netherlands	6
Croatia	10	Norway	2
Czech Republic	7	Poland	2
Denmark	7	Portugal	2
Estonia	8	Romania	7
Finland	6	Russia	11
France	2	Scotland	1

Georgia	1	Serbia	5
Germany	7	Slovakia	6
Greece	3	Slovenia	6
Hungary	2	Spain	1
Iceland	2	Sweden	4
Ireland	2	Switzerland	1
Israel	1	UK	8
TOTAL		**140**	

Table 1: Countries and institutions across Europe

COUNTRY	PEOPLE
Australia	2
Brazil	1
Ghana	1
Cuba	1
Kuwait	1
Malaysia	2
Peru	1
South Africa	2
USA	8
TOTAL	**19**

Table 2: Countries and institutions outside Europe

3.1.1.1 Educational background (N=159, Q3)

Figure 1 shows that more than half of the respondents have a PhD (61%) and the majority has a degree in language/linguistics (81.1%).

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.

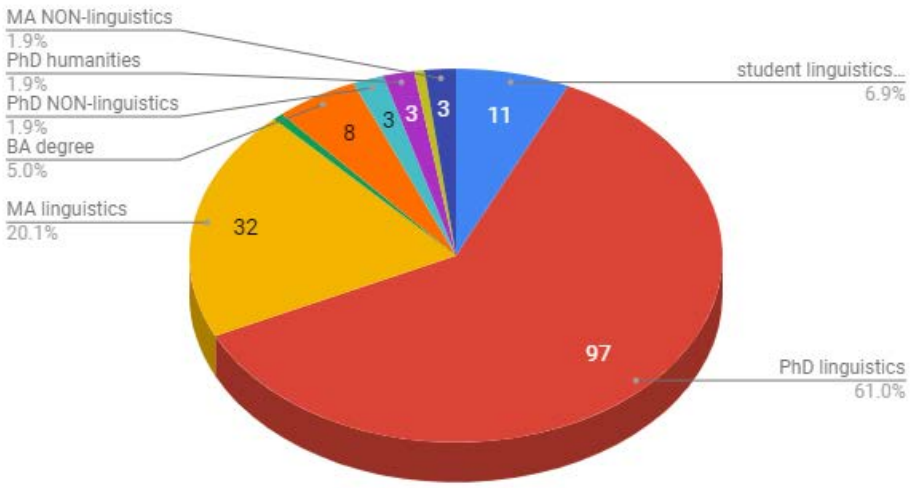

Figure 1: Educational background

3.1.1.2 Working years as a lexicographer (N=159, Q4)

The respondents range from very experienced lexicographers to those with little experience. The diagram shows that more than one third of respondents have more than 20 years of work experience in the field of lexicography (35.8%), every fourth lexicographer has 10-20 years of work experience (24.5%) and every fifth has 5-10 years of work experience (20.1%). These responses may be an indication that people who have started working as a lexicographer stay in the field for a long time. Every tenth respondent (10.1%) has very little work experience, having worked in the field for 1-3 years.

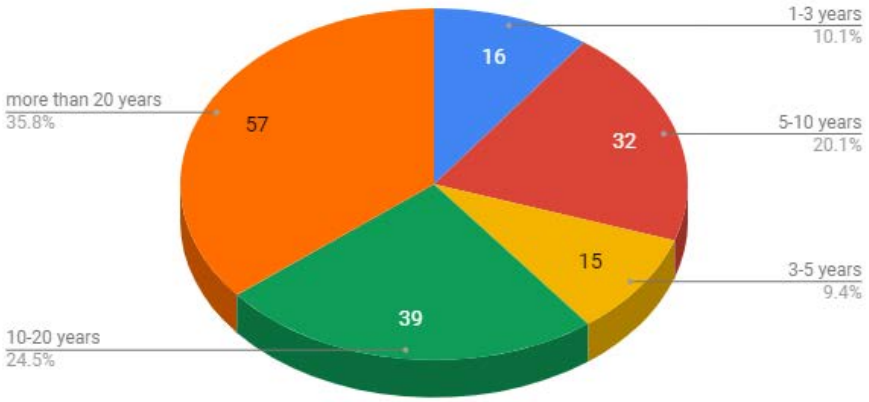

Figure 2: Working years as a lexicographer

3.1.1.3 Employment (N=154, Q5)

The diagram shows that the majority of respondents work as full-time in-house employees (68.6%). There are also quite a lot of freelance lexicographers among the respondents (22.6%).

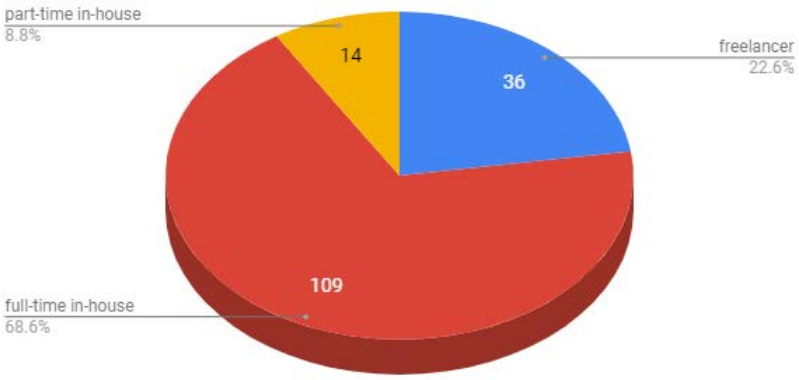

Figure 3: Employment

3.1.1.4 Type of institution or company (N=122, Q6)

Figure 4 shows that the majority of respondents came from public institutions or non-governmental organisations (77.9%). 17.2% respondents work at the universities. A small number of responses (4.9%) came from lexicographers working for private/commercial companies in Europe.

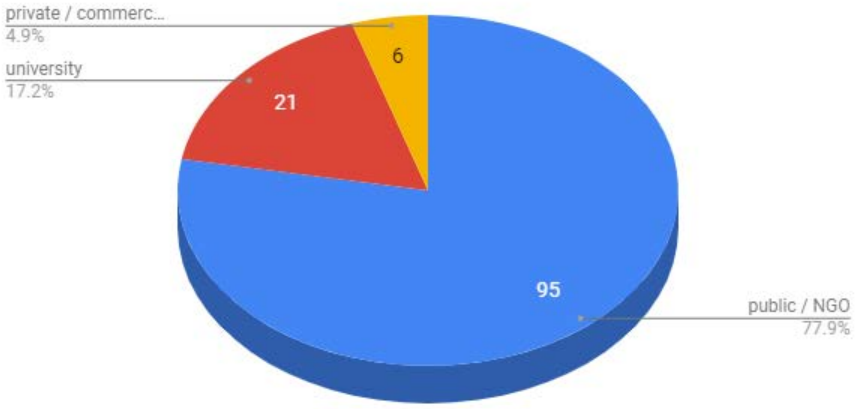

Figure 4: Type of institution or company

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.

3.1.1.5 Specific training as a lexicographer (N=159, Q7)

Figure 5 shows that more than one third of respondents have been trained within their own institute, usually by a tutor or a senior lexicographer (34.6%). One fourth of respondents have attended special courses or several courses (25.8%) since starting to work in lexicography. Only 11.3% of respondents report studying lexicography at the university, either as part of an MA course on lexicography or as a special course.

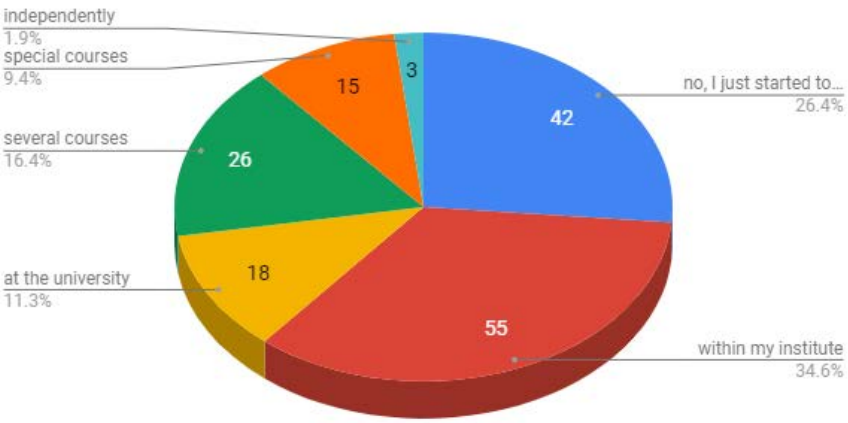

Figure 5: Specific training as a lexicographer

3.1.2 Ongoing work

3.1.2.1 Team size (N=157, Q8)

Figure 6 shows that the respondents work in teams of different sizes, with relatively similar shares being reported across all team sizes. The predominant team size among the respondents is 3-6 people (27.4%). There are also a few respondents that work in teams with more than 50 people (2.5%), while on the other hand, many respondents (mostly freelancers) do not work in a team (13.4%). Overall, we can comment that the majority of our respondents work in teams with less than 10 members.

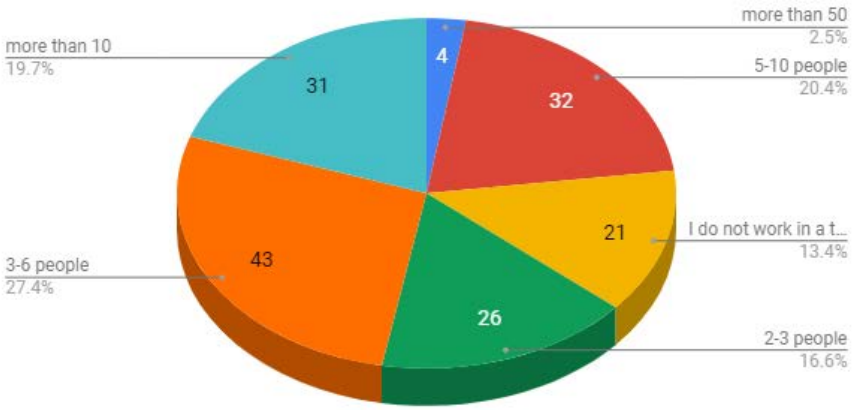

Figure 6: Team size

3.1.2.2 Joint teams (N=143, Q9)

The respondents were asked if their team includes people from different institutions or countries.[7] Figure 7 shows that more than half of the respondents belong to a team that consists only of people from their own institution (56.6%) and less than half are working together with people outside their institution (43.4%).

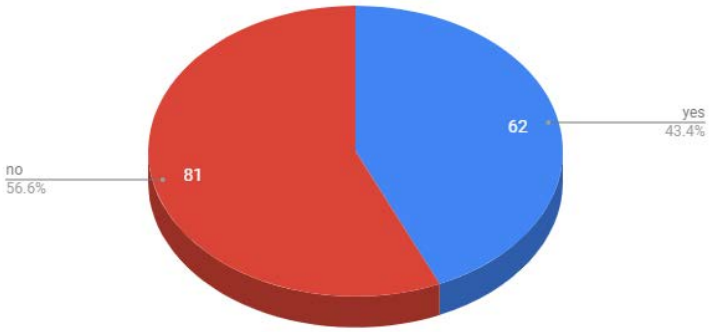

Figure 7: Joint teams

3.1.2.3 Types of projects (N=159, Q13)

Figure 8 shows that the majority of ongoing projects mentioned in the survey are monolingual dictionaries or databases (58.5%), either general, specific or dictionaries for learners. Much less respondents are involved in compiling bilingual (15.1%), multilingual (13.2%) and dialectal (8.8%) dictionaries or databases. There are a few projects that report combining monolingual data with

[7] All respondents could answer this question which means that it could also be answered by respondents that chose "I do not work in a team" in the previous question. Some of them did indeed answer this question.

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.

bilingual or multilingual data (altogether 4.4%). These projects might be monolingual projects with multilingual and multimodal extensions (linking with other languages) or they might be aggregated unified databases containing different kinds of data (lexicographic as well as terminological databases).

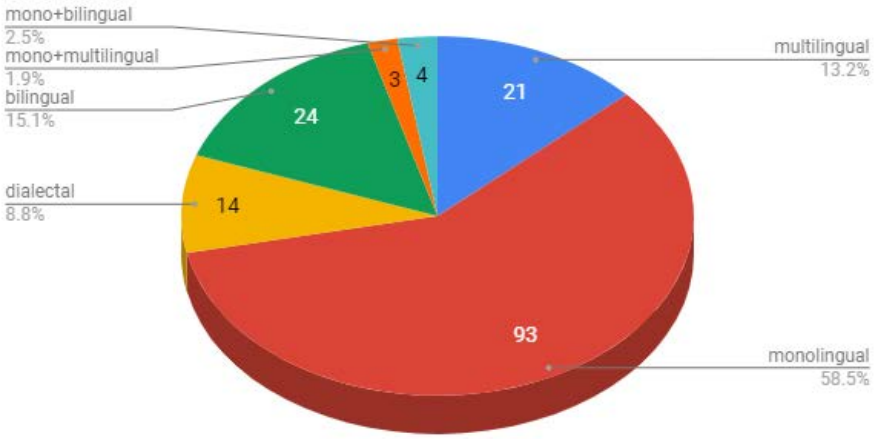

Figure 8: Types of projects

3.1.2.4 Kind of data (N=159, Q14)

Figure 9 shows that the majority of ongoing projects mentioned in the survey deal with general language (54.1%). One third of the projects deals with specific areas of language (32.1%), e.g. collocations, word-formation, word combinations, idioms, etc., either for general use or language learners, or either monolingual or bilingual. A terminological project was mentioned by every tenth respondent (10.1%).

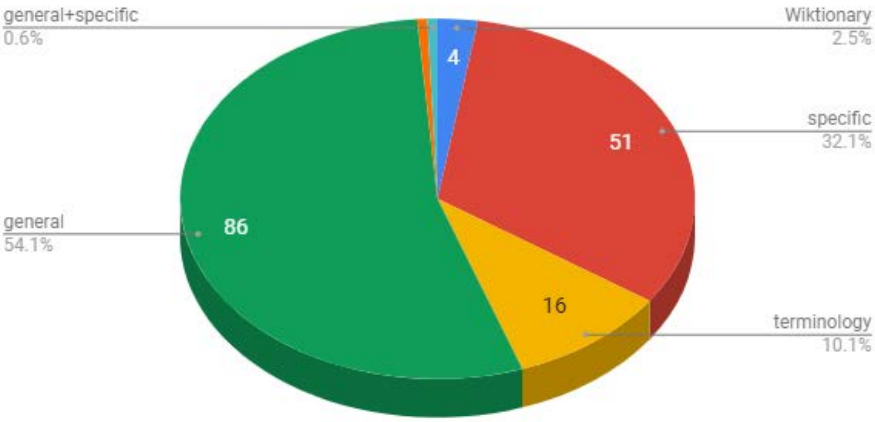

Figure 9: Kind of data

3.1.2.5 Specific dictionaries (N=53, Q14)

When looking more closely into the projects dealing with specialised dictionaries, we see that a variety of dictionary types (besides general dictionaries and terminological databases) were mentioned by the respondents. The most mentioned types are historical dictionaries (28.8%), dialectal dictionaries (17.3%), etymological dictionaries (13.5%), collocation dictionaries (9.6%) and idiom dictionaries (3.8%).

3.1.2.6 Duration of projects (N=158, Q11-12)

114 different projects were mentioned by the respondents. More than half of these projects are permanent projects (53 projects); these are mainly voluminous monolingual contemporary dictionaries, Wiktionaries, etymological and dialectal dictionaries, as well as some bilingual dictionaries, but also some specialised dictionaries (e.g. football expressions, neologisms, word combinations). Another 18 have a duration of 15-20 years; these are also mainly voluminous monolingual contemporary dictionaries, etymological and dialectal dictionaries, as well as some bilingual dictionaries. 22 projects have a duration of 5-10 years; these are mainly special or bilingual dictionaries. 21 projects have a duration of 3-4 years; these are mainly special dictionaries (e.g. spoken language, sign language, idioms, terminological dictionaries).

3.1.2.7 Organisation of the database (N=158, Q15)

Figure 10 shows that the majority of the project databases of the respondents are organised from word to meaning (word-based databases, 87.3%). Databases organised from meaning to word (concept-based, 8.9%) are used mainly when working with terminological data. There is also a small number of projects that combine both, word-based and concept-based organisation of the database (3.2%). One project mentioned being 'word-based and pattern-based' as "meanings are associated with patterns".

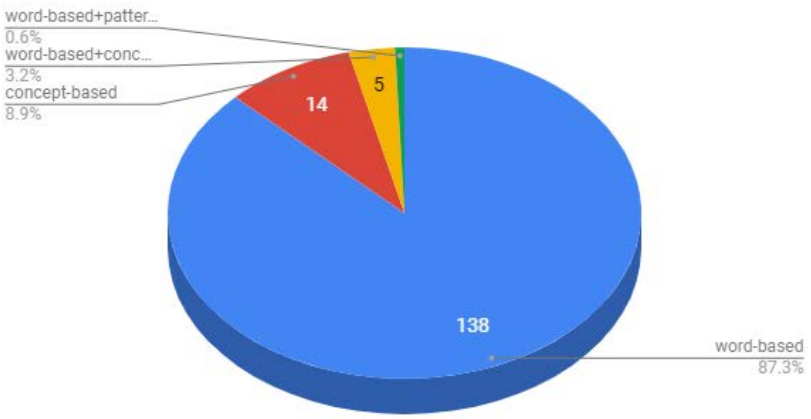

Figure 10: Organisation of the database

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.

3.1.2.8 Born-digital dictionaries (N=159, Q16)

The respondents were asked to give information on whether their project was born-digital or not. The term 'born-digital' was defined explicitly in the survey as 'a dictionary conceptualised for the electronic medium, offering radically different options for organisation and presentation of lexical information' in the survey. The options for answers were: "Yes", "No", "Other". Figure 11 shows that the majority of the respondents (54.1%) did not see their projects as born-digital. Some respondents (5%) reported their project being partly born-digital and left additional explanations, mostly that the project had started as a manual one, but developed into born-digital in a later stage. 40.9% of the respondents claim that their project is born-digital. When looking into the answers describing the compilation method of the databases in these projects, it seems though that not all these projects can be considered born-digital according to our definition of the term (see also section 3.1.2.10).

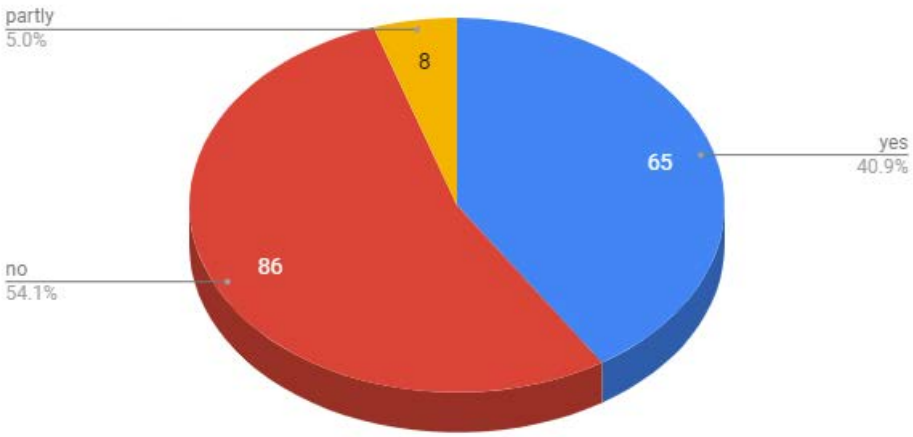

Figure 11: Born-digital dictionaries

3.1.2.9 Compiling methods for all projects (N=159, Q17)

The diagram shows that the majority of the respondents compile their dictionaries manually (57.9%). Nearly one third of the respondents work with semi-automatically collected data (30.8%) and some work manually while using some tools (3.8%). Only a few respondents indicated using fully-automatically collected data (7.5%). Altogether, less than half of the respondents (42.1%) use special tools in their dictionary projects.

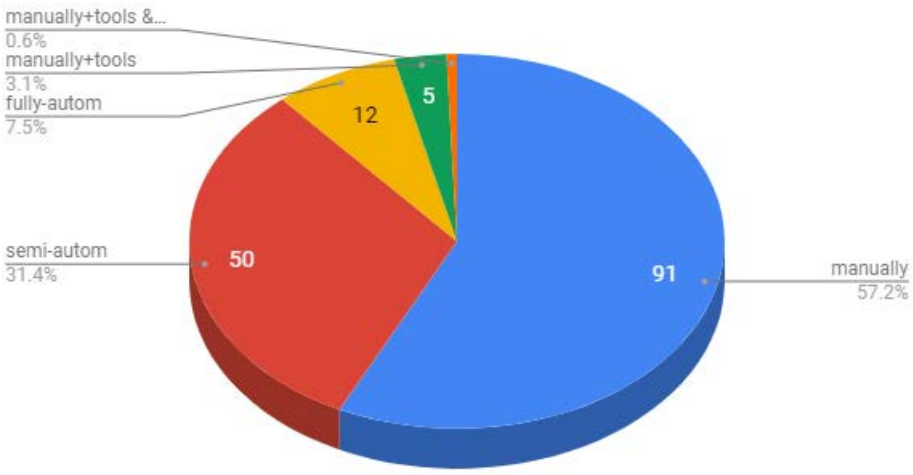

Figure 12: Compiling methods for all projects

3.1.2.10 Compiling methods for born-digital dictionaries (N=65, Q17)

The respondents who marked their project to be born-digital mentioned using different compiling methods: mainly semi-automatic (43.1%) and surprisingly also manual (!) (33.8%). It seems that although the term 'born-digital' was explicitly defined in the survey, the notion stayed somewhat unclear for many respondents who seem to consider a dictionary that is compiled using the computer as born-digital. Every seventh project (16.9%) was compiled fully automatically.

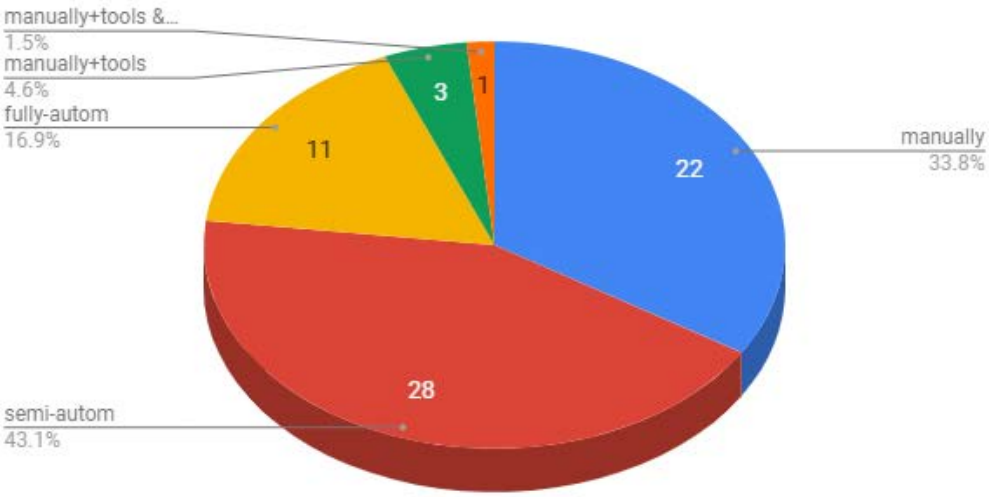

Figure 13: Compiling methods for born-digital dictionaries

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.

3.1.2.11 Compiling methods for not born-digital dictionaries (N=86, Q17)

Figure 14 shows that the respondents who marked their project to be *not* born-digital mentioned using different compiling methods: mainly manual (74.4%) but also semi-automatic (22.1%). One respondent answered even "fully-automatic" but probably meant working on a computer.

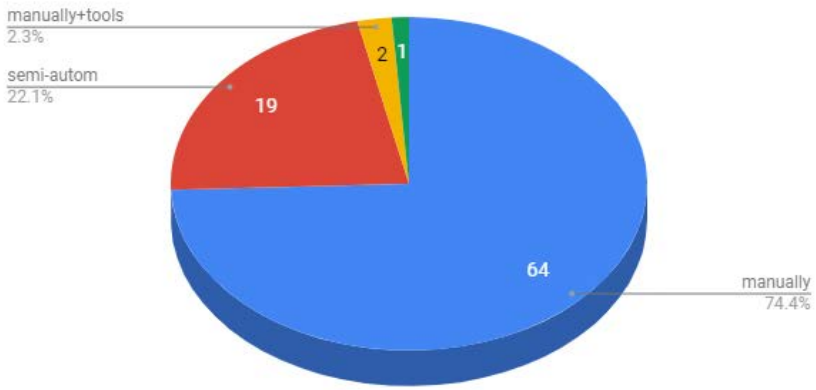

Figure 14: Compiling methods for not born-digital dictionaries

3.1.2.12 IT support (N=98, Q18-19)

Figure 15 shows that nearly half of the respondents claim having basic IT support for their work (43.9%). A fair number of respondents reported having good IT support (37.8%). However, the rate 'good' should not be overestimated as the analysis of the answers reveals that many lexicographers using manual compiling method for their work have answered that they are satisfied with their IT support ('good' or 'basic'). The second group of the respondents who chose the answer 'good' were those who use semi-automatic or fully-automatic methods but *wish for more*.

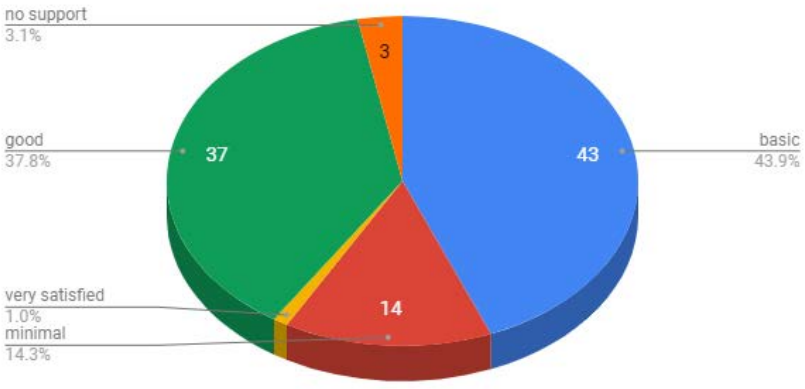

Figure 15: IT support

3.1.2.13 Outsourcing (N=159, Q18-19)

Figure 16 shows that the majority of the respondents do not use outsourcing for their projects (69.2%). 26.4% of the respondents work in projects where outsourcing is used. A small percentage of the respondents are not aware of whether outsourcing is used in their project(s) (4.4%).

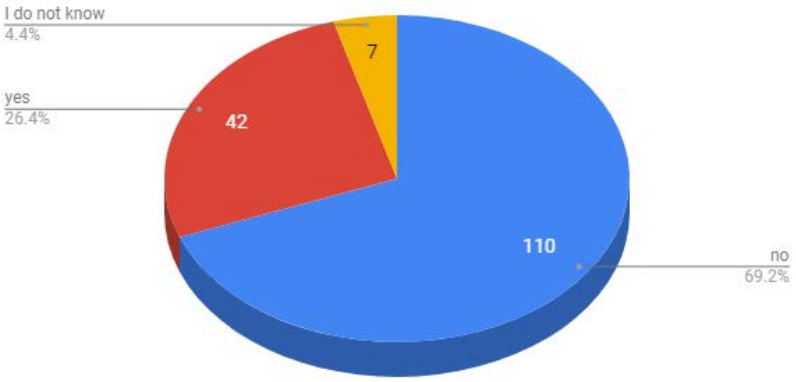

Figure 16: Outsourcing

3.1.2.14 Outsourcing affecting the workflow (N=29, Q21)

Of the 29 respondents who indicated that they have experience with outsourcing, 14 judged it as a very good or good experience. Nine respondents delivered the 'so-so' judgement; four respondents mentioned that this has brought a lot of extra work. However, as they commented, this extra work had to be done to improve the quality of their own data. And two respondents mentioned that outsourcing does not affect them directly as it mainly deals with online presentation of dictionary data.

Outsourcing seems to be mainly used for graphic design / online publishing; smartphone apps; Corpus Query Systems (e.g. Sketch Engine); new Dictionary Writing System development; constant development of tools. Trustworthy experts / efficiency and another view of the data and content (which might help to identify some lexicographic problems) were mentioned as positive experience. The cost (too expensive, lack of (regular) funding), more work (to teach and explain lexicographic details), delays and communication problems were mentioned as negative experience when outsourcing.

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.

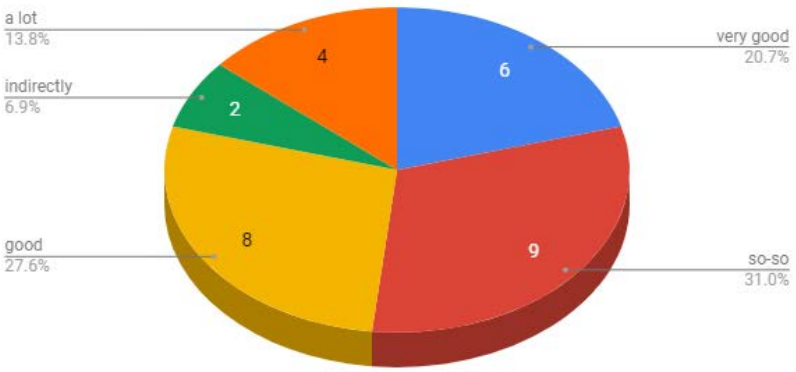

Figure 17: Outsourcing affecting the workflow

3.1.3 Software and tools

3.1.3.1 Software and tools supporting the workflow (N=89)

70 respondents did not answer the question whether they use tools to support their workflow, but from those who did (N=89) more than half (55.7%) reported that they use both a Dictionary Writing System (DWS) and a Corpus Query System (CQS) in their work.

There are mainly three types of combinations: commercial DWS and commercial CQS (e.g. IDM and Sketch Engine), in-house DWS and commercial CQS (e.g. EELex and Sketch Engine), in-house DWS and in-house CQS (LexDF and Corpus Workbench). The commonest model is the combination of in-house DWS with Sketch Engine.

Generally, the lexicographers in our survey use one CQS and one DWS, but some of them use several DWSs , e.g. iLex, Lexonomy and Tlex, and several CQSs at the same time (mostly Sketch Engine plus another system), e.g. Sketch Engine and KonText, Sketch Engine and Lexpan, Sketch Engine and Korp. Some institutions use Sketch Engine and noSketchEngine.

Altogether 54.8% of the respondents use Sketch Engine as CQS, other more commonly used CQSs are Corpus Workbench (CWB), CoRest, Korp, NoSketchEngine, AntConc, COSMAS II.

10.2% of the respondents only use a CQS (mostly Sketch Engine) and 15.9% only use a DWS (mostly in-house systems).

29.6% of the respondents use also special software for retrodigitisation, mainly for the compilation of historical, dialect and etymological dictionaries.

3.1.3.2 Dictionary Writing Systems (N=71, Q23-25) and Corpus Query Systems (N=78, Q26-28)

Altogether 15 Dictionary Writing Systems (DWS) and 22 Corpus Query Systems (CQS) were mentioned by respondents. In the table below, these systems are divided into three main categories: commercial, open-source and in-house. Online interfaces (only) and general purpose editors, dictionary publishing platforms and App Builders are considered as separate categories.

DWS / CQS	NAME	URL OR REFERENCE
COMMERCIAL		
DWS	IDM	http://dps.cw.idm.fr/index.html
DWS	iLex	https://issuu.com/jens.erlandsen/docs/ilex_brochure_120dpi
DWS	SDL MultiTerm	https://www.sdl.com/software-and-services/translation-software/term
DWS	TLex, Tlterm	https://tshwanedje.com/tshwanelex/ https://tshwanedje.com/terminology/
CQS	Archivarius 3000	http://www.likasoft.com/ru/document-search/
CQS	Folio Views	http://www2.iath.virginia.edu/elab/hfl0028.html
CQS	Lexis/Nexis Academic	https://academic.lexisnexis.eu/
CQS	Sketch Engine	https://app.sketchengine.eu/
CQS	TLex	https://tshwanedje.com/tshwanelex/
CQS	WordSmith Tools	https://www.lexically.net/wordsmith/
OPEN-SOURCE		
DWS	Alexis	http://alexis.fox1.cz
DWS/	FLEx (Fieldworks Language	https://software.sil.org/fieldworks/

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.

CQS	Explorer)	
DWS	leXkit	http://ixa.si.ehu.es/node/4462?language=en
DWS	Lexonomy	https://www.lexonomy.eu/
DWS	TermKate	http://termkate.elhuyar.eus/
CQS	AntConc	http://www.laurenceanthony.net/software/antconc/
CQS	BlackLab CQS	https://github.com/INL/BlackLab/blob/master/core/src/site/markdown/corpus-query-language.md
CQS	Corpus Workbench (CWB)	http://cwb.sourceforge.net/
CQS	COSMAS II	https://www.ids-mannheim.de/cosmas2/
CQS	CQPweb	http://cwb.sourceforge.net/cqpweb.php
CQS	Korp	https://spraakbanken.gu.se/eng/korp
CQS	Lexpan (Lexical Pattern Analyzer)	http://www1.ids-mannheim.de/lexik/uwv/lexpan.html
CQS	NoSketchEngine	https://www.sketchengine.eu/nosketch-engine/
CQS	TXM	http://textometrie.ens-lyon.fr/spip.php?rubrique49&lang=en
IN-HOUSE		
DWS	DEAF-DWS	http://www.deaf-page.de/st.php
DWS	EELex, since 2019 new system Ekilex	https://eelex.eki.ee https://ekilex.eki.ee
DWS	INT-DWS	Tiberius, Carole, Jan Niestadt and Tanneke Schoonheim (2014): 'The INL Dictionary Writing System'. In: Iztok Kosem and Michael Rundell (eds) *Slovenščina 2.0: Lexicography*, 2 (2): 72–93.
DWS	JMdictDB - Japanese	http://www.edrdg.org/jmdictdb/

	Dictionary Database	
DWS	LexDF	The product is not publicised, but registered with Inven2, The UiO patent and IPR organisation, since 2014.
DWS	Redigeringsapplikasjonen	https://www.hf.uio.no/iln/om/organisasjon/edd
CQS	CoREST	https://korpus.dsl.dk/corest/index.htm
CQS	DGD	https://dgd.ids-mannheim.de/DGD2Web/jsp/Welcome.jsp
CQS	ItzulTerm	http://itzulterm.elhuyar.eus/
CQS	KonText	https://kontext.korpus.cz/first_form?corpname=syn2015
ONLY ONLINE INTERFACE		
CQS	mtf3	http://clara.nytud.hu/mtsz/run.cgi/first_form
CQS	WhiteLab	https://github.com/TiCCSoftware/WhiteLab
GENERAL PURPOSE EDITORS, DICTIONARY PUBLISHING AND APP BUILDERS		
DAB (Dictionary App Builder)		https://software.sil.org/dictionaryappbuilder/
FrameMaker		https://www.adobe.com/ee/products/framemaker.htm
MediaWiki		https://www.mediawiki.org/wiki/MediaWiki
Oxygen		https://www.oxygenxml.com/
Webonary		https://www.webonary.org/

Table 3: Dictionary Writing Systems and Corpus Query Systems mentioned

When asked to describe their likes, dislikes, wishes and needs about the systems they use, the majority of respondents mentioned a number of requirements that apply both to DWS and CQS, i.e.:

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 system should be free, online, fast, open-source, browser independent, intuitive, easy to maintain
- the system should be interoperable with other resources, operating systems and tools. The majority of the respondents emphasised the need for automatic pre-compilation of entries and the possibility to integrate lexicographic information automatically from CQS into DWS
- the system should have API and script support
- the system should allow real-time collaborative input
- the system should enable real-time saving
- the system should be customisable, both in terms of functionalities and interface
- the system features should be localisable (e.g. Sketch Grammar and GDEX configuration in Sketch Engine)
- the system should enable an infrastructure for online publishing of the results
- the system should have proper documentation (not a black-box system)

In addition to the general requirements, there are a number of features that were mentioned as positives specifically for either DWS or CQS.

The following important functionalities were mentioned in relation to DWS:

- support for (automatic) data collection (simple import and export of files, mapping transcripts, inclusion of media files (e.g. audio files with a linked transcript)
- support for data management and data processing (unified data model, format standardisation, version history, assignment tools, change tracking, statistics, complex searching, advanced visualisation options, automatic validation tools, internal reference facilities, spell checker integration, bulk editing tools, easier mass updates, easy handling of subentries)
- support for (automatic) data extraction
- support for data publishing (e.g. print and export functions)
- tools for processing forum data and other user-generated content
- tools for the involvement of external experts: simple and low-cost (no-cost) solutions for external review and comments
- tools for crowdsourcing.

The following functionalities were considered important specifically for CQS:

- support for corpus compilation (new corpora creation (incl. spoken corpora), supporting various data formats, better access to certain types of texts (e.g. transcriptions), possibility to present legally sensitive data)
- support for corpus annotation (lemmatisation, tagging, multi-level annotation, incl. morphology, syntax, semantics)

- support for corpus annotation editing (corpus editing on the fly (e.g. tagging mistakes, mark-up), search-and-annotate function, misspellings detection, improved detection of noise in corpus data, support for data evaluation)
- support for corpus metadata editing
- support for data processing and data analysis (lemma list, word list, statistics, (advanced) CQL support, concordances, context filters, text types sorting, co-occurrence analysis, longest commonest match, neologism detection, diachronic analysis, (bilingual) term extraction, (syntactic) pattern detection)
- support for semantic analysis, enhanced sense disambiguation and semantic/sense clustering
- support for data acquisition (multi-level extraction, detecting language changes in real-time).

3.1.3.3 Data acquisition from CQS (N=84, Q29)

Altogether 17 different types of lexicographic data were proposed as possible answers of data types that can be obtained from a CQS. All these different data types are used by the respondents but not to the same degree.

The most commonly used data types obtained from a CQS are: dictionary examples (12%), collocations (10.3%) and frequency information (10.2%). Extraction of multi-word expressions (9.1%), patterns (7.5%), form variants (6.5%) and word senses (5.6%) are fairly common too.

Less than 5% of the respondents use CQS for the acquisition of lexical-semantic relations (4.9%), neologisms (4.6%), domain information (4.4%), definitions (3.9%), information on register (3.6%) and diachronic distribution of senses (2.6%).

Less than 2% of the respondents use CQS for the acquisition of multilingual data from parallel corpora (1.8%), knowledge rich contexts (1.6%), audio data from speech corpora (0.9%), clustering of data (0.5%) and regional varieties (0.2%). The last two data types were suggested by respondents.

3.1.3.4 Automatic data extraction / Automatic knowledge extraction (N=150, Q30)

The same list of types of lexicographic data (as in Q29) was used to see which types of data were automatically extracted from corpus data.

The most commonly mentioned data types are: automatic extraction of headword list (20.8%), collocations (12.7%) and frequency information (11.3%). Automatic extraction of multi-word expressions (8%), dictionary examples (7.5%) and form variants (6.1%) are fairly common too.

Less than 5% use automatic extraction for patterns (4.7%), neologisms (3.8%), lexical-semantic relations (3.8 %), domain information (4.4%), multilingual data from parallel/comparable corpora (3.8%), definitions (3.3%) and audio data from speech corpora (2.4%).

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.

Less than 2% use automatic extraction for knowledge rich contexts and regional varieties.

3.1.4 Publication

3.1.4.1 Publishing medium (N=150, Q31)

Figure 18 shows that almost half of the 150 respondents answering this question (46%), reported that the dictionaries they are working on are published online only. One third of the respondents (32%) reported both publishing online and in print. Every fifth (19.3%) respondent reported publishing his/her work in print only. Online dictionaries might be supplied with a dictionary app, print dictionaries with a CD. A small percentage (2%) of the respondents reported that they publish their dictionaries online, as an app, and in print.

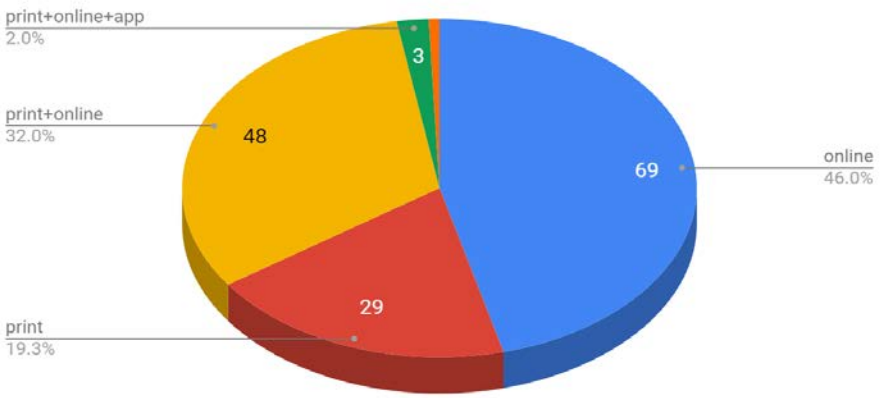

Figure 18: Publishing medium

These 150 respondents represent 124 different lexicographic projects altogether of which 100 (80%) will be published online and almost half of those (45 projects, 45%) will also appear in print. More than half, 54 projects (54%) are published online only. 24 projects (19.3%) will be printed only. Apps are reported for 4 projects (3,2%).

3.1.4.2 Involvement in online publication process and user research (N=63, Q32-33)

Lexicographers were asked to specify what kind of work they do when they are involved in online publication or user research. It was an open-ended multiple-answer question, but three options were proposed: 1. evaluating the user interface and providing new ideas; 2. creating add-on materials (e.g. blogs, slideshows, videos, quizzes, word games); 3. communicating with IT persons / user experience designer (UX) / interface designer (IX).

27% of respondents answered that they are not involved in online publication.

33.9% of those who are involved in online publication deal with user interface evaluation, and communication with IT specialists, including user experience designers and interface designers. In addition to user interface evaluation and communication with IT specialists, 16.9% of the

respondents are involved in the production of add-on materials. 11.9% are involved only in user interface evaluation and 8.5% only in IT communication.

Other tasks mentioned include:

- project management, e.g. communication with editors and IT people, updating user guides, taking care of their translation, testing new editions, negotiating with the publisher about forthcoming editions
- organizing dictionary updates / updating the web site, designing new GUIs
- presenting and discussing the updates in the media (including social media channel), e.g. Word of the day, weekly language question, news items
- contact with users via help desk questions
- analysis of feedback from the users (proposals, corrections). The feedback is regularly given via mail or Web feedback form
- provide expertise on different formats, exporting data as XML/XHTML
- typesetting with LaTeX for book publication.

The respondents were asked if they are involved in user research for their dictionary, and if so what kind of user research they do. The options proposed were: 1. analysing user logs; 2. interviewing end users.

62.5% (55 respondents) revealed that they are not involved in user research. 59% of those lexicographers who do user research conduct analyses of user logs, 33.2% also conduct interviews with end users (mostly before and during the conceptual phase).

Other tasks mentioned include:

- analysis of data from language-related advisory services (such as recommendations on word usage, grammatical constructions) and from Google Analytics
- analysis of user feedback, mostly proposals and corrections (the feedback is gathered through mail or online feedback forms)
- conceiving and supervising user studies carried out by others
- informal consultation.

3.1.5 Retrodigitisation

The aim of this part of the survey was to reveal the involvement of the lexicographers in different phases of the retrodigitisation process (i.e. the process of converting a dictionary published in paper into a digital, computer-readable format, which involves not only scanning and OCRing but also data encoding and enrichment), to get an overview of the software used in this process and to provide an insight in the lexicographers' opinion on which dictionaries should be retrodigitised.

The number of respondents in this part is rather small compared to the total number of respondents (10.06%). This corresponds with the fact that some parts of the retrodigitising activities (image and

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.

text capturing) are not related directly with the lexicographic work, while other parts (data encoding and enrichment) require additional technical support.

3.1.5.1 Involvement of lexicographers in retrodigitising (N=15, Q34)

This question allowed multiple answers selected from the following options: not involved in retrodigitisation; image capture; text capture; data encoding; data enrichment; Other (requiring a specification). The results are presented in Figure 19.

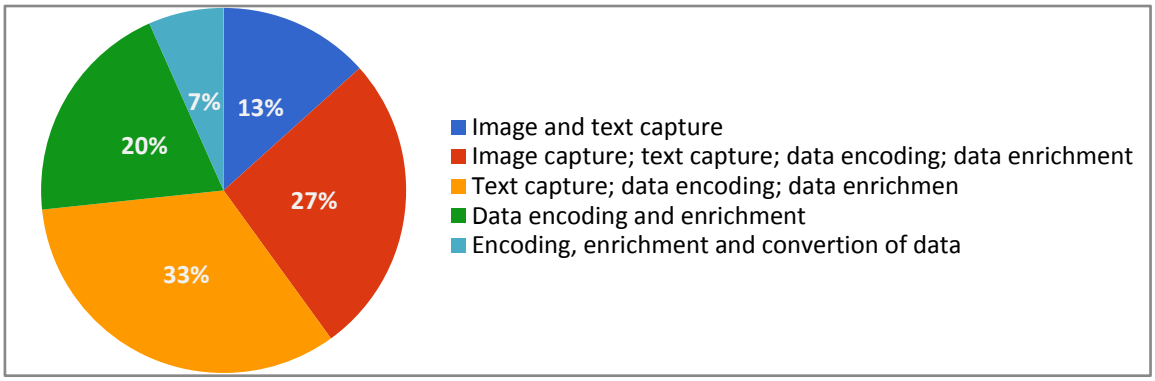

Figure 19: Involvement of lexicographers in different phases of retrodigitising

All respondents have selected multiple answers which shows that the lexicographers have been involved in more than one phase of retrodigitisation (the option Other was specified as conversion of data by one respondent).

However, if we look at the individual phases of retrodigitisation, we see that the lexicographers take part mostly in the activities which require lexicographic competence such as data encoding (15 responses) and data enrichment (13 responses in Figure 20).

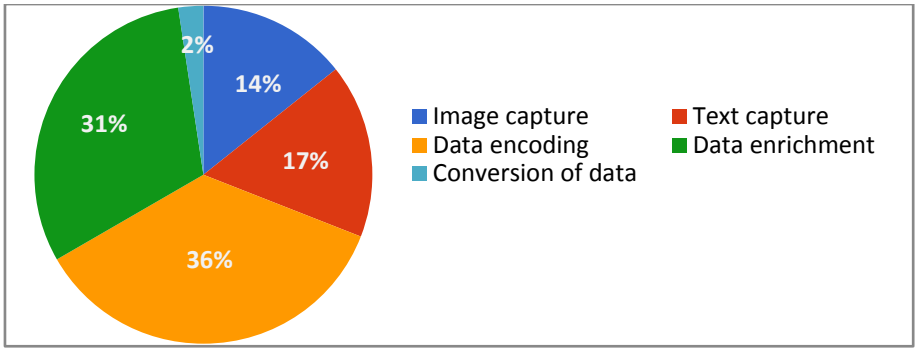

Figure 20: Involvement of lexicographers in different phases of retrodigitising (separately)

3.1.5.2 Image capture: procedures and software (N=2, Q35)

The question is open-ended requiring a short description of what has been done for image capturing (scanning print dictionaries) and what software has been used (name, Internet address (url) or other reference). There are two responses: a response pointing out that the image capturing was performed by an external company and a response noting that a particular software was used but not specifying which. The responses show that not many lexicographers among those completing the survey are involved in the technical part of scanning.

3.1.5.3 Text capture: procedures and software (N=5, Q36)

The question is open-ended requiring a short description of what has been done for text capturing (OCRing and post-editing) and what software has been used (name, Internet address (url) or other reference). Regarding the procedures for text capturing, one respondent pointed out double keying and another described a chain of word processing, dictionary compilation and publishing. Regarding the software used for text capturing, three responses referred to ABBYY FineReader

3.1.5.4 Data encoding: procedures and software (N=10, Q37)

The question is open-ended requiring a short description of what has been done for data encoding (structural and semantic markup) and what software has been used (name, Internet address (url) or other reference). There are 10 responses, presented in Table 4:

PROCEDURE	USERS	SOFTWARE
Conversion from plain text to XML	2	not specified
Conversion from plain text to XML	1	Oxygen
Conversion from plain text to TEI XML	1	Oxygen
Conversion from plain text to TEI XML	1	Oxygen, XSLT, Xpath

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.

not specified	1	Oxygen, XSLT, PERL scripts, Excel
Cleaning, proofreading, tagging and parsing	1	Emacs, for tagging and parsing N/S
Preprocessing the markup of Toolbox data	1	PERL scripts
Preprocessing	1	not specified

Table 4: Data encoding: procedures and software

The entries in paper-born dictionaries are usually paragraphs of text with surface formatting like bold and italics, but very little explicit structure beyond that. That is why a conversion from plain text to XML (5 responses) is performed to obtain an explicit structure comparable with the structure of born-digital dictionaries.

The most widely used tool for data encoding is the Oxygen XML Editor. Also, Perl scripts and XML-based technologies such as XSLT and Xpath are used.

3.1.5.5 Data enrichment: procedures and software (N=9, Q38)

The question is open-ended requiring a short description of what has been done for data enrichment (adding additional language and/or linguistic information) and what software has been used (name, Internet address (url) or other reference). There are 9 responses, presented in Table 5:

PROCEDURE	USERS	SOFTWARE
Text normalisation	1	XSLT in Oxygen; BaseX
Enrichment lexical data with audio documentation	1	Lame, MP3DirectCut
Internal and external linking, adding superordinate grammar information	1	not specified
Mapping pos-tags; expanding abbreviated forms	1	not specified
Transforming TEI into LOD (Ontolex-Lemon) and linking to existing resources and vocabularies	1	not specified
Producing indexes of grammatical and semantic information	1	not specified
not specified	1	Oxygen
not specified	1	http://lkiis.lki.lt/
not specified	1	http://гизаурус.рф

Table 5: Data enrichment: procedures and software

Some responses point out the tools but do not describe the data enrichment itself and vice versa - there are responses describing the data enrichment but not specifying the tools used, i.e. the cells containing the text not specified in the above table. Data enrichment is an important procedure because it concerns not only retrodigitised dictionaries but also born-digital dictionaries and affects interconnection of data and options for querying and presenting the information. That is why the obtained information might mean that: a) the data enrichment (and linking) is still not widely used; b) there are no standardised procedures for data enrichment.

3.1.5.6 List of dictionaries for retrodigitisation (N=13, Q39)

The aim of this question was to create a list of dictionaries which should be considered for retrodigitisation, and to find out why these dictionaries are important. The results are summarised in Figure 21.

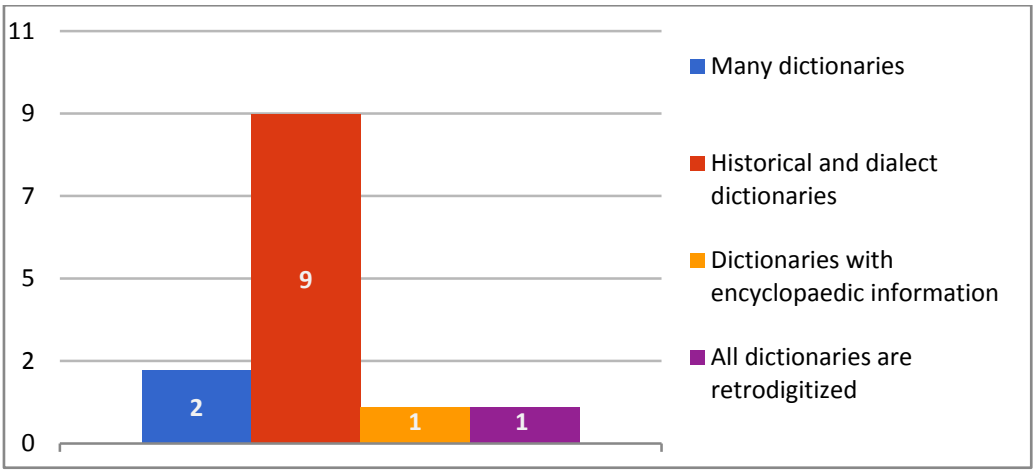

Figure 21: Dictionaries for retrodigitisation

The responses show that there are dictionaries that are not digitised yet. Of great interest are historical and dialect dictionaries (9 responses). The respondents note that retrodigitisation has a big potential for research in linguistics - dialectal, historical, etymology; philology; digital humanities.

There are some conclusions from this part of the survey. Although activities such as scanning, OCRing and proofreading require special attention, they do not need the involvement of qualified lexicographers. Instead, this step of retrodigitising could be outsourced if there is dedicated financing.

The added value of retrodigitised dictionaries might be two-fold: a) as a source for online references; b) as building blocks for developing new dictionaries. In both cases efforts for structuring (data encoding) the retrodigitised dictionary data are needed while in the re-use case the linking to the explicit and (possibly) complex structure of born-digital dictionaries is also required (data enrichment). Linking, enrichment and reusing data does not affect only retrodigitised dictionaries, so the efforts in this direction might be consolidated and if further investigation of lexicographers'

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.

practices are performed questions and answers for data encoding and data enrichment might be more detailed.

On the other hand, the lexicographers value the data described in the old dictionaries as they pointed out many dictionaries (mainly historical and dialect) as a possible target for digitisation. These conclusions might suggest a development of a common infrastructure for retrodigitising ensuring and possibly unifying the technical part of the process leaving the lexicographers room for creative work.

3.1.6 Past and future (N=116, Q40-41)

The main positive changes noted by the respondents during the last 10-15 years are connected mostly with the digitisation and automation of lexicographic work, online publishing (moving from paper to online) and with the beginning of corpus era together with access to better data (corpora, internet) and better tools (e.g. Sketch Engine). New type of systems (Corpus Query Systems, Dictionary Writing Systems) and tools (not only for editing and analysing, but also for semi-automatic extraction) were developed. The same process took place in different European countries. As a result a numerous amount of in-house and commercial CQSs and DWSs were created. As the biggest advantage of an online platform, the possibility for regular updates was mentioned, as well as more effective collaboration via internet.

The second challenge noted by the respondents is the change in the attitude towards dictionary users (user needs are considered as one of the more important tasks to be taken into consideration) and attempts to involve the public (to implement crowdsourcing) into dictionary content compilation. Also the interaction between the users and the dictionary has improved, since users can directly contact lexicographers online about words they are looking for, technical issues etc.

The last biggest challenge that was mentioned is connected with the use of mobile devices. It was noted that the impact of mobile phones is immense as a distribution method, and a mobile-first approach has to be adopted.

On the other hand, the survey reveals that the community is very heterogeneous, some issues that are favourably mentioned by some lexicographers can be considered as negative by others, e.g. moving from paper to online would not be good as “paper is more durable than web”. Some respondents reported moving from typewriting or handwriting to using the computer as the major change during the past 10-15 years.

One of the main concerns is connected with rapid technology development. Software constantly changes, lexicographers need to heavily rely on IT support. Some respondents find that an overestimation of the presentational/technological component of dictionary, especially focusing only on smartphone-view, may result in neglecting the aspect of the quality and reliability of lexicographic data. Moreover, some respondents raised a point about the publication of printed dictionaries online, and the fact that the new format does not encode information that was

presented in the original work. Another concern is connected with information overload caused by a fascination with the endless possibilities offered by the electronic medium.

It is also important to mention that some respondents noted the low status of lexicography in their countries: there is not enough money to keep lexicographers working.

As for the future of lexicography, the main change is expected in relation to lexicographic data modelling and publishing policy. The turn towards unified data is expected, with respondents mentioning that publishers will produce a single resource containing all the data that the publisher has about the language, including data traditionally not considered part of a dictionary.

Respondents were also asked to to identify their wishes and needs in the next 10-15 years. Below the most frequently mentioned topics are listed:

- better tools for extraction and automatic processing of data from corpora (incl. clustering corpus occurrences by sense, semantic analysis, detection of new senses and language changes, detection of conceptual relations, definition extraction, extraction of syntactic patterns, terms etc.)
- semantic web technologies, publishing as Linked Data; more use of AI and Deep Learning
- the need for common standard for the development of lexicographic resources; the need for central repositorium; tools for harmonisation of dictionary formats
- better corpus analysis tools for spoken language
- better support for retrodigitisation
- better infrastructure for online publishing and tools for visualisation
- tools for crowdsourcing; tools for the analysis of forum data and other user-generated content
- speech to text tools/audio dictionaries
- more use of Google corpora (Books (including NGrams), Scholar, News, UseNet) and Google analytics
- support for API access
- dictionary app builders
- empirical dictionary user research, more communication with users, incl. more teaching of dictionary use
- more writing tools for text production
- the need for publishing policies and licensing regulations.

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.

3.2 Survey for Institutions

The *ELEXIS Survey of Lexicographers' Needs for Institutions* was targeted specifically at the lexicographic partner institutions within the project. One survey had to be completed per institution. This survey was more elaborate than the survey targeted at individual lexicographers and the expertise of a computational linguist or IT specialist was most likely required to answer some of the more technical questions.

The *ELEXIS Survey of Lexicographers' Needs for Institutions* contained 86 questions divided into 6 sections, i.e. (1) General information; (2) Types of lexicographic resources, software and tools supporting the workflow; (3) Publication and access. Crowdsourcing and gamification; (4) Retrodigitised dictionaries; (5) Data formats. Metadata. Availability; (6) Past and Future. Of those 86 questions, there were 17 "yes/no" questions, 34 multiple choice questions (for 24 of those more than one answer could be selected and for 10 only one answer could be given), and 40 open-ended questions.

Below, we present the results of our analysis following the structure of the sections in the survey. As the survey was quite long (the estimated time to complete it was 45 minutes to an hour) respondents were offered the opportunity to save the survey at the end of each section and to continue later. Note that these questions are not included in the total number of 86 questions. Next to each (sub)heading we provide the number of the question in the survey (e.g. Q3, Q25-27). These numbers relate to the survey questions which can be found in Appendix 2.

3.2.1 General information (Q1-17)

The *ELEXIS Survey of Lexicographers' Needs for Institutions* was completed by the 11 lexicographic partner institutions in the project, i.e:

NAME OF INSTITUTION	SHORT NAME	COUNTRY
Austrian Academy of Sciences: Centre for Digital Humanities	OEAW	Austria
Institute for Bulgarian Language	IBL	Bulgaria
Society for Danish Language and Literature	DSL	Denmark
Institute of the Estonian Language	EKI	Estonia
Trier University, Trier Center for Digital Humanities	TCDH	Germany
Hungarian Academy of Sciences, Research Institute for Linguistics	RILMTA	Hungary
K Dictionaries	KD	Israel
Instituut voor de Nederlandse Taal	INT	Netherlands
Belgrade Center for Digital Humanities	BCDH	Serbia
"Jožef Stefan" Institute	JSI	Slovenia
Real Academia Española	RAE	Spain

Table 6: ELEXIS lexicographic partner institutions

3.2.1.1 General information about the respondents (Q5-8)

Only one questionnaire had to be completed per institution. The first set of questions collected general information about the person who completed the survey on behalf of the institution. The results show that the survey was primarily completed by lexicographers/terminologists in a senior position (i.e. being a member of the board/council or a project manager), with more than 20 years experience in lexicography. Most of them have a PhD and the majority has a degree in language/linguistics.

The respondents were also asked to characterise themselves with regard to traditional lexicography vs. modern e-lexicography. Their responses show that e-lexicography is clearly growing. None of the respondents indicated that they feel more comfortable with traditional lexicography (paper slips, writing in Word, paper dictionaries) or that they are used to work electronically, but think that dictionaries should be printed (in addition to e-dictionary). As the diagram below shows, about half answered that they feel comfortable with both, traditional and e-lexicography, and the other half indicated a clear preference for e-lexicography (corpora, dictionary writing systems, born-digital dictionaries, e-publishing).

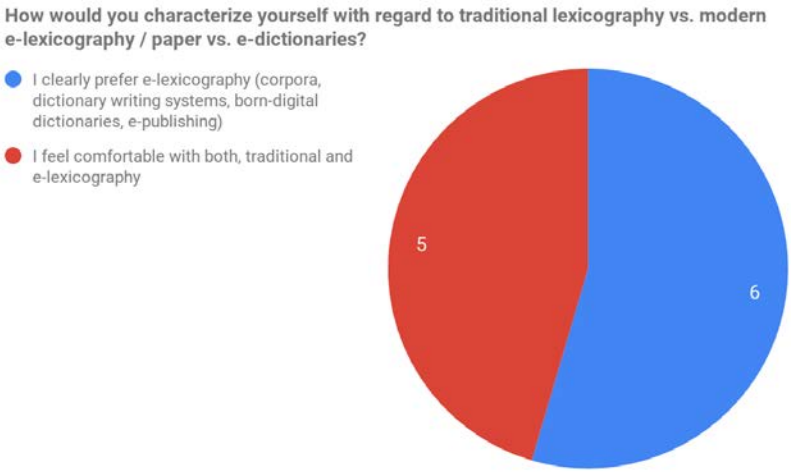

Figure 22: Respondents' characterisation with regard to traditional lexicography vs. modern e-lexicography

3.2.1.2 General information about the institutions (Q10-17)

Most of the 11 lexicographic partner institutions are public institutions or non-profit organisations. Only one of the lexicographic partner institutions is a commercial company. The majority of the public/non-profit institutions receive, in full or as part of their income, funding on a regular basis (eg. stable funding by government/ministry/academy), which can be complemented by project funding.

Between 1 and 10 lexicographers (summed up into full-time employment) are employed by each of the partner institutions, except for one which employs around 26 lexicographers.

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.

Most partner institutions provide some sort of training for their lexicographers. In-house training is most common, but some institutions also offer their lexicographers the opportunity to go to external courses, workshops or summer schools. Only one institution indicated that it does not offer any kind of training to its lexicographers.

Lexicographers who are employed by the partner institutions mainly work on lexicographical projects (especially in the case of third-party funded projects), but not exclusively. Common other tasks that lexicographers are involved in are teaching, management, and dissemination.[8]

All 11 partner institutions have IT support. Although the software engineers are often not working full-time on lexicographic projects, half of the institutions have answered that they do not outsource their work. It should be noted though that the definition of outsourcing was not clear to all respondents (i.e. digitisation in the sense of converting from printed to digital format was not counted as outsourcing by one of the respondents).

Development of a user-interface is the task which is most commonly outsourced, followed by the development of a CQS, DWS, or database. In the case of retrodigitisation, scanning and typing and converting audio files are typical tasks that are outsourced.

3.2.2 Types of lexicographic resources, software and tools supporting the workflow

3.2.2.1 Lexicographic resources and expertise (Q19-22)

3.2.2.1.1 Lexicographic expertise of the institutions (Q19)

The diagram shows the lexicographic expertise of the partner institutions. It shows that the partner institutions have a 'varied' lexicographic expertise ranging from general dictionaries to specialised dictionaries, dialect dictionaries to terminological dictionaries, both monolingual as well as multilingual and synchronic and historical.

[8] The survey did not ask how much time lexicographers generally spend on these other tasks.

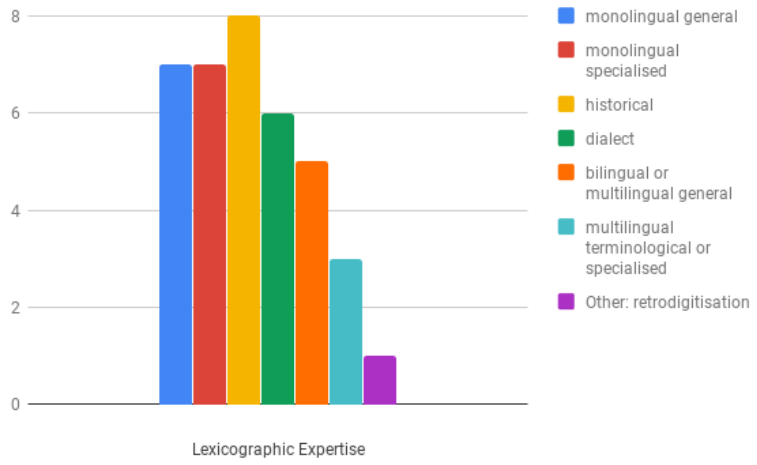

Figure 23: Lexicographic expertise

Eight out of the eleven institutions have expertise in historical lexicography. Five of those work both on historical and contemporary lexicography. Bilingual and multilingual expertise is slightly less represented within the lexicographic partner institutions.

3.2.2.1.2 Amount of lexicographic resources per institution (Q20)

As the diagram shows, most of the partner institutions have between 10-50 lexicographic resources at their institution.

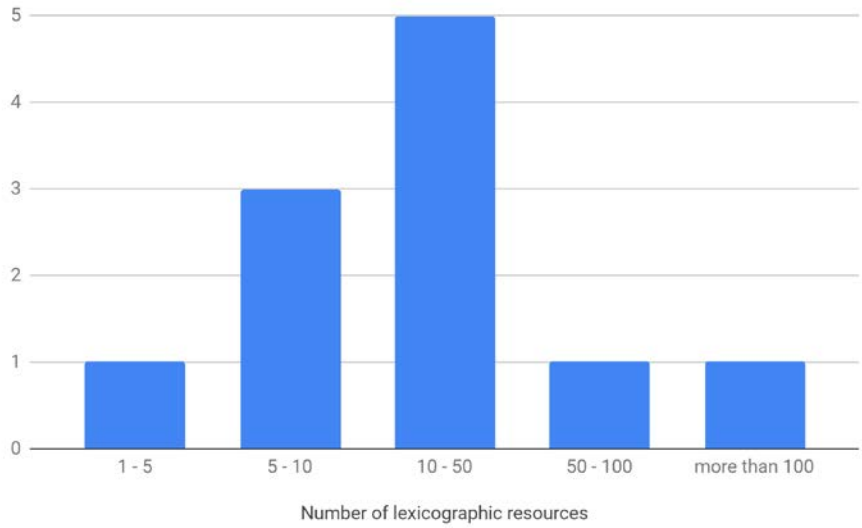

Figure 24: Number of lexicographic resources per institution

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.

3.2.2.1.3 Projects per institution (Q21,Q22)

Table 7 gives an overview of the main lexicographic projects that have recently started per institution (2014-2021) **(Q21).**

PROJECTS THAT HAVE RECENTLY STARTED	
JSI	1. Thesaurus (https://viri.cjvt.si/sopomenke/eng/) 2. Slovene Lexical Database (http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza) 3. Morphological lexicon Sloleks (http://eng.slovenscina.eu/sloleks)
RAE	1. Diccionario de la lengua española, 23th ed., annual update (http://dle.rae.es) 2. Diccionario de la lengua española, 24th ed. (new design) 3. Diccionario del estudiante, 3rd ed. (http://enclave.rae.es, Android/iOS) 4. Diccionario de español jurídico (http://dej.rae.es) 5. Diccionario panhispánico del español jurídico (http://www.rae.es/obras-academicas/diccionarios/diccionario-panhispanico-del-espanol-juridico) 6. Nuevo diccionario histórico del español, updates (http://web.frl.es/DH) 7. Diccionario fraseológico panhispánico (http://www.rae.es/noticias/francisco-javier-perez-el-dfp-sera-un-proyecto-hermanado-y-paralelo-al-nuevo-dle)
KD	1. German/Arabic bilingual bidirectional dictionary 2. Danish-English-Korean trilingual dictionary 3. Revision of English learning dictionary
IBL	1. Dictionary of active Polish and Bulgarian phraseology (2018-2020). 2. Dictionary of economy terms (2015-2018). 3. Dictionary of new words in Bulgarian (2018-2019).
TCDH	1. Trier Dictionary Net (www.woerterbuchnetz.de) 2. (Retro)Digitisation and web publication of the 2DWB (revised edition of the „Grimm“) (https://www.kompetenzzentrum.uni-trier.de/en/projects/projects/digitising-the-revision-of-the-german-dictionary-by-jacob-and-wi/) 3. ZHistLex (http://zhistlex.de/) three year project aimed at the building of an

	eHumanities Centre for Historical Lexicography, which is funded by the German Federal Ministry of Education and Research.
EKI	1. Collocations Dictionary 2. Dictionary of Place Names 3. Small dialect dictionaries 4. Synonym Dictionary (to be started in 2018)
OEAW	1. WBÖ (https://www.oeaw.ac.at/acdh/projects/wboe/) 2. Dictionary of Loanwords in the Midrash Genesis Rabbah
INT	1. Neologism dictionary (not yet available on the internet) 2. Dictionary of Word Combinations (pilot) (not yet available on the internet)
DSL	1. Constant development and expansion of The Danish Dictionary (DDO) (https://ordnet.dk/ddo) 2. The Danish Thesaurus (only in print) (https://dsl.dk/publication?id=430) 3. Development of a dictionary portal including a series of retrodigitised vocabularies from the 16th century as well as some newer dictionaries describing Danish Language in that period. The project is part of a larger project focusing on Danish hymn books of the 16th century. (https://dsl.dk/projekter/musik-og-sprog-i-reformationstidens-danske-salmesang) 4. Retrodigitisation of several dictionaries, among these: Dictionary of the Danish Language, ODS, (Danish 1700-1950) (https://ordnet.dk/ods) 5. Kalkar's Dictionary (Danish 1300-1700) (https://kalkarsordbog.dk/) 6. Swedish-Danish (https://ordnet.dk/sdo) 7. Latin-Danish (https://latinskordbog.dk/)
BCDH	A platform for Serbian dictionaries (http://raskovnik.org)
RILMTA	New Etymological Dictionary of Hungarian. (1st part commenced in 2011 and the 2nd part in 2017. It will end in 2021.)

Table 7: Main projects per institution

Table 8 gives an overview of the projects that will be published in the near future (2018-2021) **(Q22)**

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.

PROJECTS TO BE PUBLISHED SOON	
JSI	1. Collocations Dictionary 2. Multiword Expressions Database
RAE	1. Diccionario de la lengua española, 23th ed., annual updates 2. Diccionario de la lengua española, 24th ed., advances 3. Nuevo diccionario histórico del español, updates
KD	1. German/Arabic bilingual bidirectional dictionary 2. Danish-English-Korean trilingual dictionary
IBL	1. Dictionary of Bulgarian language, vol. 16. (http://ibl.bas.bg/rbe/) 2. Dictionary of ecological terms 3. Dictionary of new words in Bulgarian
TCDH	1. revised edition of the Trier dictionary net (see above) 2. internet publication of the revised edition of the „Grimm“ (see above)
EKI	1. Explanatory Dictionary of Estonian (130,000 headwords) 2. Dictionary of Standard Estonian 3. Associations Dictionary
OEAW	1. WBÖ (https://www.oeaw.ac.at/acdh/projects/wboe/) 2. Dictionary of Loanwords in the Midrash Genesis Rabbah
INT	1. Algemeen Nederlands Woordenboek; dictionary of contemporary Dutch (from 1975 onwards; http://anw.ivdnt.org/search (daily updates) 2. Neologism dictionary (not yet available on the internet) 3. Dictionary of Word Combinations (pilot); not yet available on the internet
DSL	1. 2-3 yearly updates of The Danish Dictionary (DDO).
BCDH	A platform for Serbian dictionaries (http://raskovnik.org)
RILMTA	Comprehensive Dictionary of Hungarian, volume VII. (in 2018 autumn) (http://nagyszotar.nytud.hu/index.html)

Table 8: Projects to be published soon

3.2.2.2 Software and tools supporting the workflow

3.2.2.2.1 Dictionary Writing Systems (Q23- 29)

All lexicographic partner institutions use one or more dictionary writing system, except for one institution which currently does not use a DWS, but mentions that they have developed and used one in the past. This institution is specialised in the retrodigitisation and online-publication of printed dictionaries.

As was the case in the 2014 COST ENeL survey, it still seems quite common for lexicographic institutions to develop their own DWS (five institutions indicated that they use an in-house DWS). It is also not uncommon for the partner institutions to use more than one DWS (four institutions selected this answer). The following reasons are given for using more than one system:

- moving from commercial or in-house to open-source
- different project needs or needs of lexicographers (e.g. one for retrodigitised dictionaries, one for born-digital dictionaries; one for word-based, one for concept-based lexicography.)

The following systems are mentioned:

KIND OF DWS	NAME AND URL OR REFERENCE
commercial	IDM, iLex
open-source	Lexonomy
in-house	Hydra for Web (http://dcl.bas.bg/bulnet/)
	LexIt (http://dcl.bas.bg/LexIt/)
	EELex (https://eelex.eki.ee), since 2019 Ekilex (https://ekilex.eki.ee)
	TAReS (https://www.kompetenzzentrum.uni-trier.de/en/projects/projects/tares-webbased-system-editing-producing-publishig-dictionaries/)
	INT-DWS (previously known as INL-DWS) Tiberius, Carole, Jan Niestadt and Tanneke Schoonheim (2014): 'The INL Dictionary Writing System'. In: Iztok Kosem and Michael Rundell (eds) *Slovenščina 2.0: Lexicography*, 2 (2): 72–93
	VLE (https://clarin.oeaw.ac.at/lrp/dict-gate/vle_docu/vle_docu__v001.html)
general-purpose XML editor	oXygen, Xmetal

Table 9: Dictionary Writing Systems used

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.

About half of the partner institutions have indicated that they did not adapt or customise an off-the-shelf DWS to make it more suitable for their project(s). Those who did mention the following customisations:

- customisation of schemas, DTDs and menus
- customisation of view options (i.e for getting an overview of the entry)
- customisation of search and extraction options

Most partner institutions are quite satisfied with the DWS they use at the moment. How satisfied they are with a DWS seems to depend on factors such as the availability of support; available functionalities; possibility to adapt and add functionalities; the ability to work with multiple users and real-time updating of the database.

3.2.2.2.2 Corpus Query Systems (Q30-33)

Only two institutions have indicated that they do not use a CQS. All other institutions use one or more CQS, often combining a commercial system with an in-house or open-source system (five institutions selected this answer). The following systems are mentioned:

KIND OF CQS	NAME
commercial	Sketch Engine, Folio Views
open-source	BlackLab, Korp, noSketchEngine
in-house	https://korpus.dsl.dk; http://dcl.bas.bg/bulnc/ ; Jordi Porta (2014). From several hundred million to some billion words: Scaling up a corpus indexer and a search engine with MapReduce. *Workshop on challenges in the management of large corpora (CMLC-2)*, At LREC-2014, Reykjavik

Table 10: Corpus Query Systems used

Of the various systems, Sketch Engine is the most mentioned CQS.

Overall, the institutions are quite satisfied with the CQSs they use. Features which are not yet integrated are expected to be integrated soon as most of the systems are continuously being developed. However, most institutions do have some additional wishes for their CQS. The following important functionalities are mentioned:

- sense clustering; clustering concordances against senses (Note: this was suggested in the question.)
- implementation of syntactic and semantic annotation

- detection of neologisms
- automatic acquisition of translation equivalents
- diachronic analysis
- lexical-semantic relations
- more corpora in more languages, including more parallel corpora
- ergonomy and flexibility of the user interface, dictionary drafting, data visualisation
- easy access to metadata (i.e. author, title, etc. of a citation)
- the possibility to collect, process and query texts in different scripts (e.g. Cyrillic and Latin) in one corpus.

3.2.2.2.3 Integration of data from the Corpus Query System directly into the Dictionary Writing System (Q34)

Most CQS that are used do not allow the lexicographers to integrate data directly into their DWS. Only two institutions can integrate data (concordances or example sentences together with the metadata (source information)) from Sketch Engine directly into the DWS that they use, and one institution can integrate this kind of data directly from their in-house CQS into the DWS they use.

3.2.2.2.4 Integration of DWS and CQS into one piece of software (Q35-36)

Although it is possible to integrate data from CQS directly into DWS, most systems are not integrated into one piece of software. Most partner institutions do, however, feel that the integration of DWS and CQS would be beneficial, especially for the linking, selection and retrieval of examples, collocations, etc. It was suggested that the integration could be realised via tickboxes or something similar.

3.2.2.2.5 Automatic data extraction/Automatic knowledge extraction (Q37-40)

Most institutions use some kind of automatic data extraction. Automatic extraction of headword lists is most common (cf. results of COST ENeL 2014 survey on automatic knowledge extraction). Extraction of frequency information, collocations and dictionary examples are fairly common too. None of the institutions has indicated that they use automatic extraction methods for audio data from speech corpora, knowledge rich contexts, register information, or the diachronic distribution of senses (although one institution mentions that it has semi-automatic extraction tools for several of the tasks mentioned in the answers, including diachronic distribution).

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.

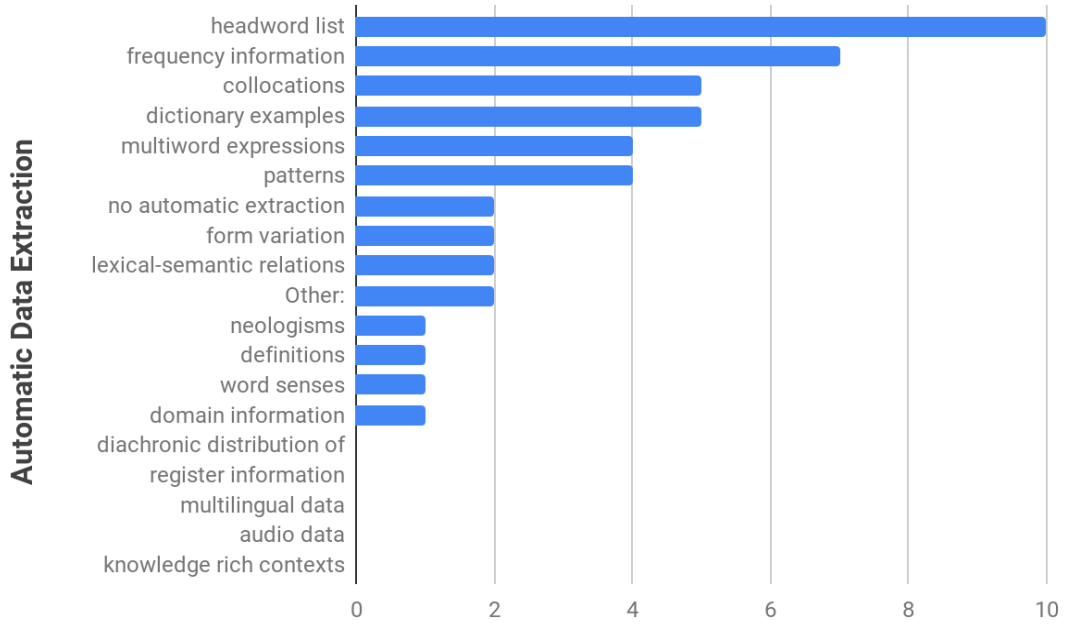

Figure 25: Automatic data extraction types

Opinions are divided on whether more is needed for automatic data extraction: 6 institutions do not have additional wishes for automatic data extraction and 5 do, i.e.

- all possible types of information extracted directly from the corpus
- extraction of knowledge rich contexts
- extraction of definitions
- extraction of word senses
- extraction of collocations
- sense-clustering

Three partner institutions indicated that they have lexicographic projects which are based on post-editing of automatically extracted data. In two of these, all raw material is or has been extracted from the corpus (both projects on collocations), and in the other, some data is or has been extracted from the corpus.

3.2.2.2.6 Reuse of existing lexicographic data within the institution in new projects (Q41, 42)

Most partner institutions reuse or integrate lexicographic data from other lexicographic projects within the institution in a new project or have done this in the past. Generally, it concerns lexicographical information from the published/existing dictionaries which is reused/integrated in another project, e.g.

- multiword expressions, including collocations
- information about senses

- synonyms
- dialect words
- etymologies
- neologisms
- headword lists
- definitions
- morphological information
- content of a monolingual dictionary for bilingual dictionaries.

3.2.3 Publication and access. Crowdsourcing and gamification

3.2.3.1 Publication of lexicographic data (Q44-47)

3.2.3.1.1 Publishing medium for lexicographic data (Q44)

The results show that online dictionaries are the most used publication medium for lexicographic data since 2010. This is also the case for projects which will be published in the near future. A reason for publishing in print is tradition; the dictionary is part of a larger project and previous volumes have appeared in print.

PUBLICATION MEDIUM SINCE 2010	RESPONSES
scanned or photographed electronic dictionary (pdf or jpg)	1
online dictionary, looking like a paper dictionary	6
online dictionary, much more dynamic than a paper dictionary	9
desktop web page without responsive design for mobile devices	4
desktop web page with responsive design for mobile devices	5
App	3

Table 11: Publication medium for lexicographic data

Three institutions also provide an app. These apps are all available on both Android and iOS. They were developed using native software development with in-house engines or using specialised software (native apps accessing central webservice). (Q45)

3.2.3.1.2 DWS and the functionality of dictionary publishing (Q46)

Table 12 shows the results for the question whether the DWS used by the institution, offers the functionality of dictionary publishing. Multiple answers could be selected, and the answers were predefined.

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.

DOES YOUR SOFTWARE (DWS OR OTHER) OFFER THE FUNCTIONALITY OF DICTIONARY PUBLISHING? SELECT ALL THAT APPLY.	RESPONSES
we do not use special software	2
export for printing (pdf, Indesign etc.)	4
export for publishing online (e.g. 'click-to-publish')	4
export for saving	5
automatic creation of metadata	3

Table 12: DWS dictionary publishing functionality

The results show that export functionalities, when available, are used by the partner institutions.

3.2.3.1.3 Access to the lexicographic data (Q48-49)

All institutions make their lexicographic data available, either through a website or a portal, and there seems to be a slight preference to make dictionaries available through their own website. More than one answer could be selected. (Q49)

HOW WOULD YOU DESCRIBE ACCESS TO THE DATA IN YOUR WEBSITE OR PORTAL?	RESPONSES
no, we do not have a website	0
each dictionary has its own website	5
dictionary collection (i.e. only external access by means of hyperlinks to the individual dictionaries, e.g. Slang Portal)	2
dictionary search engine (i.e. access to articles in the individual dictionaries, e.g. OneLook)	4
dictionary net (i.e. access to elements within the articles of the individual dictionaries, e.g. Owid, Canoonet)	4

Table 13: Access to lexicographic data

3.2.3.1.4 Customisation of the interface and the metalanguage by the user (Q48)

The answers in Table 14 show that dictionary websites often cannot be customised. When customisation is possible, it is generally limited to the interface (e.g. changing from L1 to L2).

CAN THE METALANGUAGE OF THE INTERFACE BE CUSTOMISED?	RESPONSES
no, customisation is not possible	4
interface customisation (e.g. changing from L1 to L2, according to the user's language)	5
Other: the user can choose between print and web layout, hide/show examples	1
no, customisation is not possible; Other: some kinds of customisation are possible, e.g. collapse/expand for specific information types.[9]	1

Table 14: Customisation of the interface and the metalanguage by the user

3.2.3.1.5 Access options (Q50-55)

The access options differ per institution. Most institutions provide the option of free text search on their website. Faceted browsing and API access are also quite common. SPARQL querying is currently not offered by any of the partner institutions.

ACCESS OPTIONS	YES	NO
Free text search	7	4
Filtering/faceted browsing	4	7
API access	5	6
SPARQL querying + endpoint	1	10

Table 15: Access options

3.2.3.1.6 Search options (Q56)

When we look at the search options, we see that the traditional search option of searching for a lemma (and inflected forms) is still the most common search option offered. However, more and more dictionary websites seem to offer the possibility to search for other information as well. In particular, searching for senses and definitions, syntactic information and usage notes are also offered. The combined answers (see Table 16 below) show that different institutions do different things and that there is not really a trend to be observed in the search options that institutions offer

[9] Note that more than one answer could be selected for this question.

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.

on their website(s). It may well be that the different search options that are offered correlate with the target users of the individual dictionary sites, but the survey did not ask about the target users.

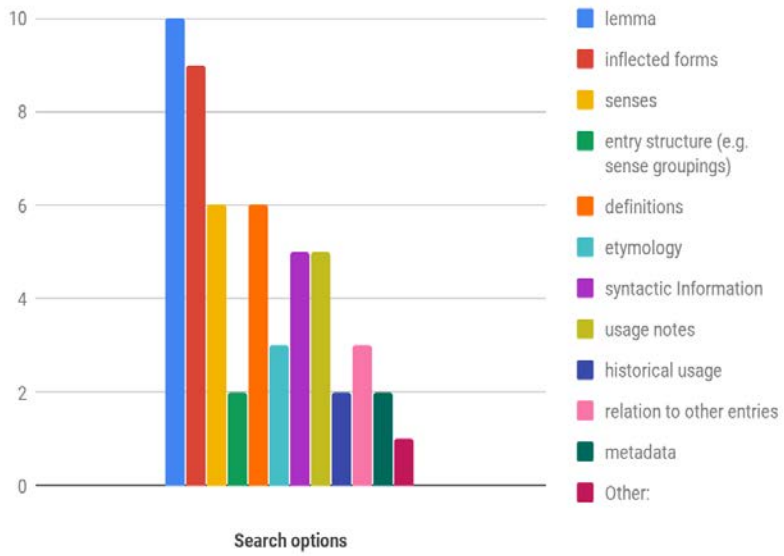

Figure 26: Search options

More than one answer could be selected and the answers were combined as follows:

COMBINED ANSWERS FOR SEARCH OPTIONS ON WEBSITE	RESPONSES
lemma, inflected forms	4
lemma	2
lemma, inflected forms, entry structure (e.g. sense groupings), definitions, syntactic Information (e.g. part-of-speech, gender), usage notes	1
lemma, inflected forms, senses, definitions, etymology, syntactic Information (e.g. part-of-speech, gender), usage notes, relation to other entries (e.g. synonyms, hypernyms, antonyms), metadata	1
lemma, inflected forms, senses, etymology, usage notes	1
lemma, inflected forms, senses, entry structure (e.g. sense groupings), definitions, etymology, syntactic Information (e.g. part-of-speech, gender), usage notes, historical usage information, relation to other entries (e.g. synonyms, hypernyms, antonyms), metadata	1
lemma, senses, definitions, another filter we have is for searching only within examples	1

Table 16: Combined answers for search options on website

3.2.3.1.7 Link to corpus data on dictionary website (Q57-58)

Most dictionary websites of the institutions involved in the survey do not offer a link to corpus data. If links are offered, this is generally implemented in a way that the entries contain an automatic URL pointing to the CQS for the given headword (four institutions). One institution also offers direct links from DWS clients into their online CQS to access corpus data (query, collocations, idioms).

If a link is offered, the user can generally not specify which elements he/she wants to retrieve from the corpus (e.g. example sentences with metadata/without metadata). Only after the user has entered the CQS, he/she can change the query.

3.2.3.2 Crowdsourcing and Gamification (Q59-62)

3.2.3.2.1 Crowdsourcing (Q59-60)

Four partner institutions currently use or have used crowdsourcing in the past. The crowdsourcing projects deal/dealt with synonyms, word associations, neologisms (in particular blends) and the transcription of a particular dialect.

3.2.3.2.2 Gamification (Q61-62)

Only one institution uses or has used gamification in a lexicographic project related to collocations. No information was given on the software used in the project.

3.2.3.2.3 Enrichment of lexicographic data with multi-modal data (images, videos) (Q63)

Only two partner institutions have indicated that they use multi-modal data from publicly available resources to enrich their lexicographic data. One institution uses images, mainly in blog posts and in some historical dictionaries, not in contemporary dictionary entries. The other institution uses both video material and images amongst others in a contemporary monolingual dictionary.

we do not use multi-modal data from the web	9
images (e.g. from Flickr, Wikimedia Commons, Europeana)	2
video material (e.g. from Videolectures.net)	1
Other:	0

Table 17: Enrichment of lexicographic data with multi-modal data

3.2.4 Retrodigitised dictionaries

A special section was dedicated to retrodigitisation. Institutions that are not or have not been involved in retrodigitisation could skip this section, except for the last question in which we asked for names of dictionaries that should definitely be retrodigitised. Four institutions have not been involved in retrodigitisation.

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.

3.2.4.1 Phases of Retrodigitisation (Q65-70)

The following phases of retrodigitisation have been considered:

- image capture: capturing images using scanners or cameras
- text capture: OCR, or keying (i.e. typing), proofreading etc.
- data enrichment: e.g. normalizing values, geo-locating, expanding content etc.
- data encoding: adding structural, i.e. semantic markup, using XML, whether TEI or not

The chart diagram below shows that the institutions which have been involved in retrodigitisation, have mostly been involved in text capture.

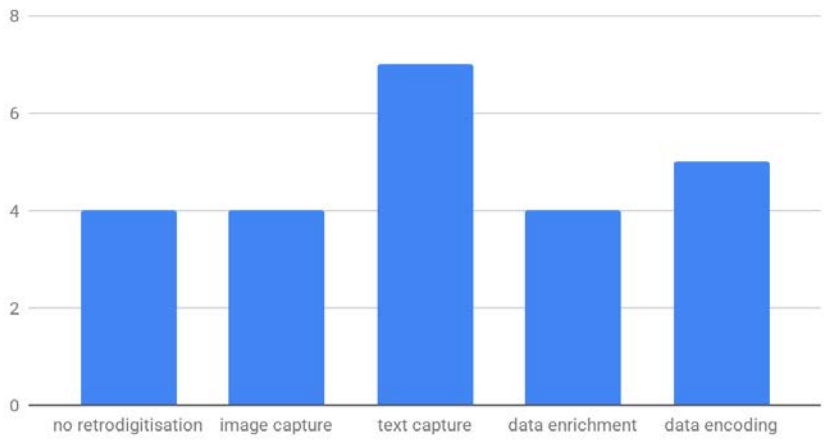

Figure 27: Retrodigitisation involvement

Seven institutions have been or are still involved in text capture. Three institutions indicate that this task is (sometimes) performed by external companies (even abroad). The double keying method is mentioned. ABBYY FineReader is mentioned as software for the OCR of scanned texts.

Five institutions have been involved in data encoding, using the following software:

- oXygen
- scripts (XSLT, Python, Perl)
- TUSTEP (TUebinger System von Textverarbeitungsprogrammen http://www.tustep.uni-tuebingen.de/tustep_eng.html)
- Access
- Excel.

Four institutions have been involved in image capture. Image capturing has been used for scanning print dictionaries and lexicographic slips. ABBYY FineReader is mentioned as software that has been used. One institution indicated that this task was performed by external companies. One institution mentions having experience with image capture in the past in the context of corpus creation, but not for retrodigitising dictionaries.

Four institutions have been involved in data enrichment (such as normalizing values, geo-locating and expanding content). The following forms of data enrichment are mentioned: geo-locating data in dialect dictionaries, recognition and completion of abbreviations and inflected forms, lemmatisation, adding modern equivalents to historical dictionary lemmas.

The following software was mentioned:

- TUSTEP (TUebinger System von Textverarbeitungsprogrammen http://www.tustep.uni-tuebingen.de/tustep_eng.html)
- Geonames
- XSLT
- Python

3.2.4.2 Access to the retrodigitised dictionaries

Access to the retrodigitised dictionaries is realised in different ways. Two institutions have kept them as standalone dictionaries. The other four have integrated them in different ways: 1) the content is integrated into an aggregator with access to data within entries; 2) the retrodigitised dictionaries are a group of dictionaries in a set with access to the dictionary via a hyperlink; 3) the retrodigitised dictionaries have been integrated, and they are one of the dictionaries in a set with access to entries within the dictionary. Two institutions use a combination of these last two approaches.

Five institutions offer access to their retrodigitised resources through an institutional portal. Two of those also offer access through an API (note however, that in one instance the API is functional and used internally by the institution, but it is not yet open to the public). One of these two partners offers a third access option and also allows users to download the full text. One institution publishes the retrodigitised material as separate websites with cross-query links and one institution allows users to download the image files.

3.2.4.3 Sharing the full text of retrodigitised dictionaries with users (Q72)

Most of the partner institutions which are involved in retrodigitisation do not share the full text of the the dictionaries with their users. Copyright is given as the main reason for not offering this functionality.

3.2.4.4 Dictionaries which should be retrodigitised (Q73)

The following dictionaries are mentioned as dictionaries which should definitely be retrodigitised:

- Estonian-German Dictionary (Wiedemann 1872)
- The dictionary of the Royal Danish Academy of Sciences and Letters (1793-1905)
- Stallaert, Rechtskundig Glossarium (Dutch legal glossary describing a specific medieval language domain)

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.

- Some 'German' dialectal dictionaries to be interlinked with already retrodigitised dialectal dictionaries to cover the entire dialectal space
- Some multilingual historical dictionaries (including Dutch)

3.2.5 Data formats. Metadata. Availability

In this section, we asked for information about technical matters at the institution. We asked about 1) data formats; 2) metadata; 3) availability. By metadata we mean data about data: information describing properties of linguistic resources, for instance, the size of a corpus, the recording date of a specific file, the purpose for which annotations were created (https://www.clarin.eu/faq-page/273#t273n2850). It was noted that the expertise of an IT person or a software developer could be required to answer these questions.

3.2.5.1 Data format(s) used for lexicographic projects (Q75)

This question intended to collect information about data formats used for lexicographic projects at different institutions. Multiple choices were provided with the following options and more than one answer could be selected: non-structured data format / text format (e.g. Word); table format (e.g. CSV, TSV, XLS); database (e.g. relational database); XML; Resource Description Framework (RDF); and other. The responses are divided between the following formats: XML, database, table format, non-

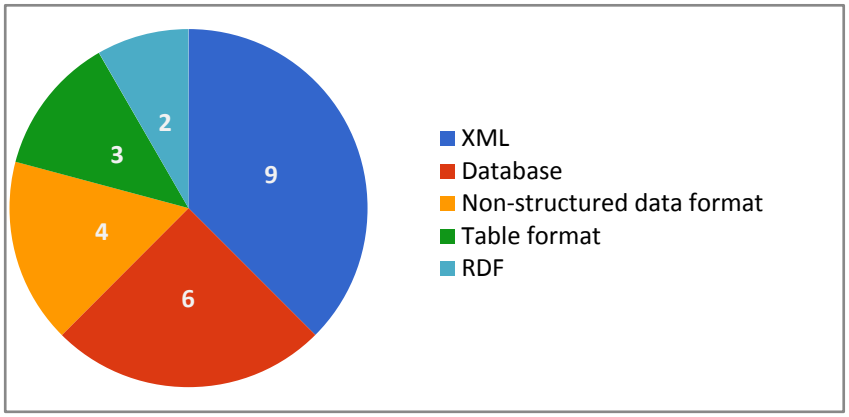

Figure 28: Data formats

structured data format and RDF, as shown in Figure 28:

The results show that many lexicographic projects use XML (9) or databases (6), but there are still projects working with non-structured data and text format (4). The re-use, linking, interchange and online publishing of the lexicographic data requires standardised and structured data formats such as XML, database, RDF, which can be used simultaneously for the collaborative production of dictionaries. Pointing out the use of non-structured data and text format simultaneously with other options shows different practice at one and the same institution. Overall, two tendencies might be outlined: a) a transition from non-structured data or text format to structured data format; b) still

insufficient use of (standardised) structured formats enabling reliable re-use and linking of dictionary data.

3.2.5.2 XML and TEI versions (Q76-77)

In this section, we show the results for the next two questions of the survey: a) *which version of XML is used*; only one answer had to be chosen among the following options: custom XML; LMF; TEI; TEI-lex and Other; and only if the response to the previous question was "TEI" b) *which version of TEI is used*, which was an open-ended question. The results presented in Figure 29 show that the custom XML (5) and the TEI (4) are the most popular XML formats, while P5 (4) is the most popular version of TEI (one institution has used P2 before P5; and one institution is using P5 but intends to move to TEI Lex-0).

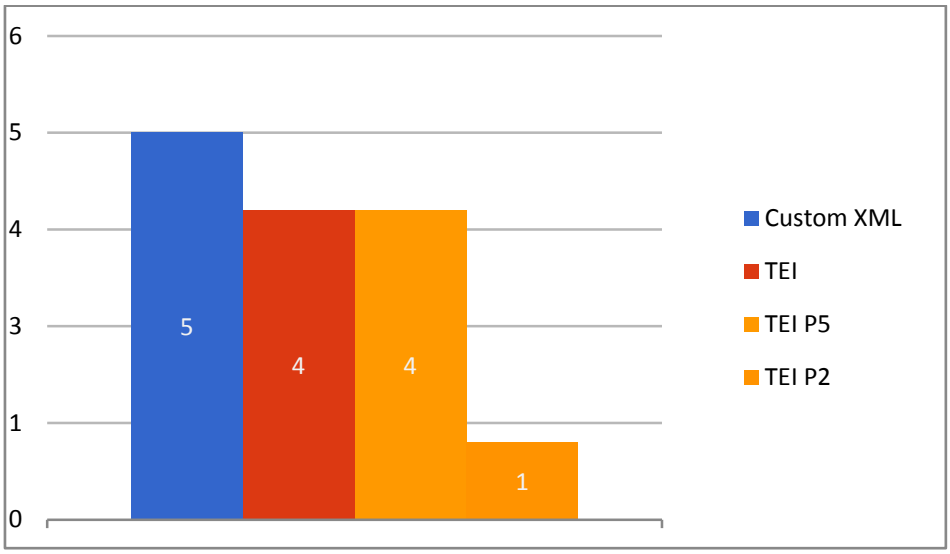

Figure 29: XML and TEI versions

3.2.5.3 Availability of tools for automatic conversion and alignment of different dictionary data formats (Q78)

The question intends to collect information for the availability of tools that allow automatic conversion and alignment of different dictionary data formats (e.g. from database format to XML). The results show that the lexicographic projects that do not have at their disposal tools for conversion and alignment of different dictionary data formats prevail. The collected information corresponds with the information about the dictionary data formats used by lexicographic projects and supports the conclusion that there is insufficient use of (standardised) structured formats enabling reliable re-use and linking of dictionary data.

3.2.5.4 Use of standard vocabularies for encoding lexicographic data (Q79)

The question aims at receiving information about the use of existing standard vocabularies for encoding lexicographic data and the respondents could select more than one answer from the

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.

following options: no, we don't; IsoCat; Clarin Concept Registry; Lemon-Ontolex; Lexinfo; GOLD; TEI and Other. Most of the responses (7) show that the lexicographic projects do not use existing standard vocabularies for encoding lexicographic data. Two institutions pointed out TEI as the standard vocabulary used for their projects and, one institution uses IsoCat, GOLD, TEI (most likely for different projects).

3.2.5.5 Use of metadata schema (Q80)

To answer the question whether a special metadata schema is used the respondents could select multiple answers from the following options: no, we do not have metadata; no, but we try to move towards a standard metadata schema; META-SHARE metadata schema v3.0 (in the CLARIN Component Registry); CMDI; Dublin Core; OLAC; TEI-header; Other. Under the option *Other* in-house developed metadata schemas are mentioned. An overview of the results is presented in Figure 30:

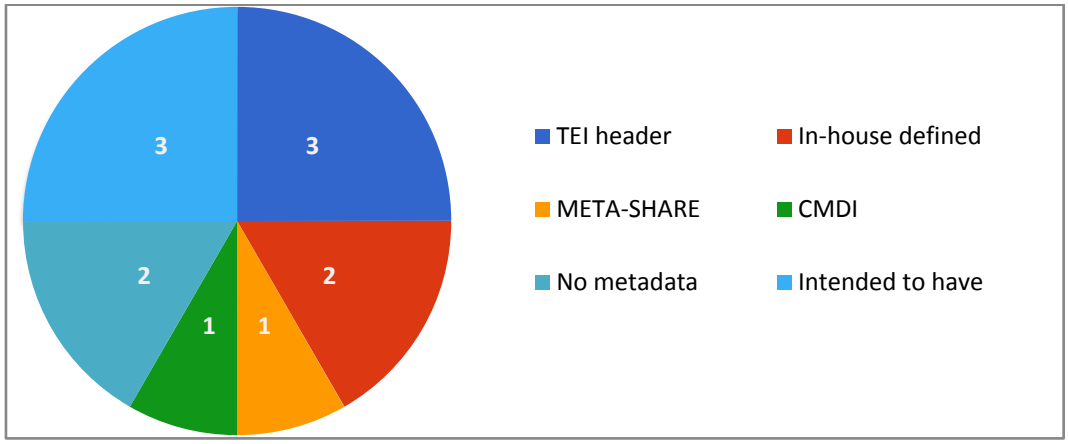

Figure 30: Use of metadata schema

Many lexicographic projects (5 responses) still do not use special metadata schema. Among the rest (6 responses) - TEI is the most preferred one (4 responses). Two institutions use different metadata schemas for different lexicographic projects.

3.2.5.6 Tools for metadata creation and editing (Q80-82)

In this section we grouped the results of the next three questions of the survey: a) if a special tool for metadata creation and editing is used; only one answer could be chosen among the yes-no options; b) if yes, specification of the tools used for metadata creation and editing; c) if no, brief explanation on whether the institution felt that it would be necessary / easier to use a special tool for metadata creation and editing.

Most of the institutions (9 responses) do not use a specific tool for metadata creation and description, while only two do. In these cases the specific tools for metadata creation and editing are DWSs. Three of the institutions that do not use a special tool for metadata creation and editing at the moment, do not envisage using such a tool in the future, while two institutions consider such a

tool necessary. It was pointed out by one institution that a tool for metadata creation and description would reduce manual work and version management.

3.2.5.7 Ways of distribution of dictionaries (Q83)

The question of how dictionary data are distributed was a question allowing multiple answers (the options are presented in the left column of Table 18):

WAYS OF DISTRIBUTION	RESPONSES
Free online	8
Restricted online / for usage fee	1
Both (some for free, others restricted or for usage fee)	1
Paper dictionary (paid)	3
(Paid) paper dictionary first, later online for free (e.g. after 1 year)	3
(Paid) paper dictionary first, later online for usage fee (e.g. after 1 year)	1

Table 18: Ways of distribution of dictionaries

The ways of dictionary distribution vary. The combination between a free online access and paid paper dictionary prevails (6 responses), however, it is not clear if this combination is used for one and the same dictionary. The number of reported free online distributions (not combined with any other way of distribution) is relatively high (4 responses). Overall, free online distribution is preferred by the academic institutions which can be explained by two main factors: a) the opportunities that the online dictionaries provide; b) the paper dictionaries developed by academic institutions often are published and distributed by third parties: publishing houses.

3.2.5.8 Access by other applications (Q84)

The question of how other applications can access the dictionary content was a multi-answer question with the following options: free API access (e.g. retrieve list of words, retrieve dictionary information for a given word); paid API access (e.g. retrieve list of words, retrieve dictionary information for a given word); free download and using under certain licence; paid download and using under certain licence; Other. There were six responses at the Other option: web interface; access on request; paid API developed but not yet supported; free API access in the near future; not sure about the legal ramifications; no access. The other responses are presented in Table 19.

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.

TYPES OF ACCESS	RESPONSES
Free API access (e.g. retrieve list of words, retrieve dictionary information for a given word)	2
Paid API access (e.g. retrieve list of words, retrieve dictionary information for a given word)	2
Free download and using under certain licence	2
Paid download and using under certain licence	1

Table 19: Types of access by other applications

3.2.5.9 Standard licensing schema (Q85)

The question on the use of a standard licensing schema was another multi-answer question. Answers could be selected from the following options: no; yes, CLARIN licensing framework; yes, Creative Commons; yes, Open Data Commons; Others. The results are presented in Figure 31.

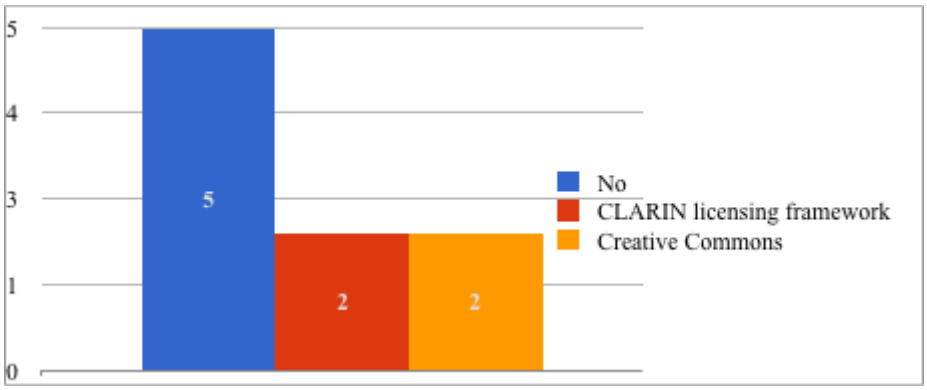

Figure 31: Use of standard licensing schema

The results show that the institutions are fairly familiar with the standard licensing schemes and some use them.

3.2.5.10 Not-supported but useful for users forms of access (Q86)

The question was open-ended and aimed to elicit a short expression of opinion on forms of access that are currently not supported but would be useful for users. The respondents suggested full text search, access via API, and the availability of free download of data.

3.2.5.11 Version control (Q87)

The question was an open-ended question that aimed to elicit a short description on how version control and archiving of different versions of the dictionary is managed. The respondents mentioned approaches such as (local) repositories or archives (GitHub, SVN), entry versioning, regular backups of databases. Overall, a tendency for systematic control is observed.

3.2.6 Past and Future

In this last section, we asked the respondents about their views for the past and the future in lexicography. Note that the answers tend to reflect the personal points of view of those who completed the survey on behalf of their institution, and that this is not necessarily the point of view of the institution.

The respondents observed the following major changes in lexicographic projects in the past 10-15 years:

- the (availability of) software, clever algorithms and tools, e.g. DWS and CQS
- better integration of dictionary and corpora
- new methods for corpus creation
- new possibilities of working with massive amounts of (corpus) data (including data from the internet)
- automatic data extraction
- online publishing and free online access to dictionaries
- the possibility to link lexicographic data to other resources and to use online resources to create new possibilities
- the radical move towards digital media.

These are considered as positive changes. Another point that was mentioned is that the task has changed from creating a dictionary to maintaining and expanding a dictionary. It should, however, be noted that this is not something trivial because of the relation between synchronic versus diachronic description within one dictionary.

Also some less appreciated changes were mentioned, such as the diminishing lexicographical competence in some "digital projects" and the fact that lexicographic resources are now also made by computer scientists without proper linguistic and lexicographic knowledge.

For lexicographic projects in the next 10-15 years, the following wishes and needs were expressed:

- integration of CQS and DWS
- more freely available data
- sharing and reusing data
- standardisation
- new technologies, automatic compilation, post-editing

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.

- different presentation modes, including mobile applications
- combining a synchronic and a diachronic approach in one resource
- a network of all etymological dictionaries in Europe
- increasing (cross-linguistic) interlinking of dictionaries, sources and bibliographic databases
- preservation of lexicographical expertise.

4 Summary

In this deliverable we have presented the results of the two ELEXIS surveys that were carried out in the context of WP1, Task 1.1, one targeted at individual lexicographers and one targeted at lexicographic institutions. The survey for institutions was much longer than the survey for lexicographers (86 versus 44 questions) and also contained a number of more technical questions which a lexicographer would not necessarily be able to answer without the help of a computational linguistic or IT person. The results give us a fairly detailed overview of lexicographic practices across Europe (and beyond) both for born-digital and retrodigitised resources. They also show what is currently needed by lexicographers and lexicographic institutions in terms of tools, functionalities and training.

Overall, the number of responses was quite high (159 for the survey for lexicographers and 11 for the survey for institutions). We obtained answers from a rather heterogeneous group of respondents, in terms of their experience, employment status, projects they are involved in (types of dictionaries, language etc.), and the country in which they are based. This to some extent ensures that the results can be generalised to the lexicographic community as a whole.

Most respondents came from public institutions and non-governmental organisations. Only a small number of respondents came from commercial companies. Most of the 11 partner institutions are also public institutions or non-profit organisations. These results seem to suggest that lexicographic work in Europe is mainly done in public institutions and non-profit organisations. This is in line with the findings of the European survey on dictionary use and culture (Kosem et al. 2018: 5)[10] conducted in 26 countries, where it was reported that in the majority of the countries participating in the survey, monolingual dictionaries are published solely or mainly by public institutions funded by the government, which is especially the case for the countries/languages with a small number of native speakers. On the other hand, commercial publishers tend to dominate in countries with a large number of speakers.

Most respondents have been working in lexicography for a long time (more than 20 years) and have a background in language and linguistics (often complemented with a PhD). However, specific lexicographic training is often received on the job. In-house training is most common, usually by a tutor or a senior lexicographer, followed by external courses, workshops or summer schools. Only a small number of respondents reported studying lexicography at the university, either as part of an MA course on lexicography or as a special course.

[10] Iztok Kosem, Robert Lew, Carolin Müller-Spitzer, Maria Ribeiro Silveira, Sascha Wolfer et al. 2018. The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture, *International Journal of Lexicography*, Advanced access: https://doi.org/10.1093/ijl/ecy022.

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.

Most respondents in the lexicographers' survey indicated to work in a team consisting of up to 10 people. This corresponds to the number of lexicographers employed by the partner institutions, which ranges from 1-10 (summed up into full-time employment). A special case are freelancers who in the majority of cases reported not working in a team.

As expected, IT support is an important part of lexicographer's job. Over 80% of the respondents answered this question and reported to have either basic or good IT support. We did not look into the dynamics between lexicographers and IT staff into more detail in this survey, but it definitely deserves more attention, particularly the way IT staff are perceived by lexicographers, and whether there are differences in the way the lexicographers perceive IT staff and computational linguists and NLP experts. IT tasks are also the only tasks that seem to be outsourced in dictionary projects, ranging from designing the online interface of the dictionary to developing and/or offering support in the use of DWS or CQS. Both positive and negative experiences with using outsourcing were mentioned by the respondents, mainly indicating the need for close(r) collaboration between the two parties involved. Still, due to a low number of responses we cannot draw any general conclusions.

In terms of software and tools, the responses to both surveys show that a large number of different tools are used to support lexicographic work (see Tables 3, 9, and 10). 15 different tools for dictionary editing and 22 different tools for corpus querying were mentioned by the respondents. Especially the combination of an in-house DWS and a commercial CQS is commonly used by the respondents. This suggests that the situation has not really changed since 2014. As a matter of fact, the COST ENeL 2014 survey on DWS and CQS also observed that it was quite common for lexicographic institutions to develop their own DWS. Consequently, the resulting lexicographic resources are typically encoded in incompatible data structures, which prevents sharing the data across different projects and applications. It also hinders linking the individual lexicographic resources to other (lexicographic and NLP) resources, which forms a significant obstacle for reusing the data in other fields, e.g. Linked Open Data, AI and NLP and the Semantic Web. As such, it is not surprising that the respondents mentioned interoperability and customisability as key requirements for DWS and CQS.

In both surveys, general monolingual dictionary projects were mentioned most often. Bilingual or multilingual projects, and dialectal projects were mentioned by a small portion of the respondents. Most of these projects are or will be published online. This applies to both the results from the survey for lexicographers and the survey for institutions. The most popular option is that dictionaries are published online only, followed by the option of publication both online and in print. The smallest number of projects will appear in print only. These results are also in line with what was reported by Kosem et al. (2018) on the status of lexicography in the 26 countries involved in their study. It should be noted though that still 24 projects out of the 124 projects mentioned in the survey for lexicographers will appear in print only. A reason for publishing in print (given by the lexicographic partner institutions) is tradition; the dictionary is part of a larger project and previous

volumes have appeared in print. This means that although fewer and fewer projects are being published as print dictionaries, the software should still cater for this option.

The online medium also brings new opportunities, such as crowdsourcing and gamification. The surveys show that crowdsourcing and gamification are not yet common practice in the lexicographic projects that our respondents are involved in. Only three projects were mentioned in the survey for institutions and the wish for tools for crowdsourcing was put down by several respondents in the survey for lexicographers. These results are not that surprising as crowdsourcing has become a hot topic in lexicography only in the last 5 years, so it is understandable that many projects are still cautious about using the wisdom of the crowd.

Most of the ELEXIS lexicographic partner institutions have expertise in historical lexicography (8 out of 11) and most of them have also been or are still involved in retrodigitisation (7). This is a relatively high number compared to the number of respondents answering the questions on retrodigitisation in the survey for lexicographers (only 15). The low number of respondents in the survey for lexicographers may suggest that lexicographers are not necessarily involved in all parts of the retrodigitisation process, either because these tasks are not directly related to their core business of editing dictionary entries, or because they require additional technical support.

In both surveys similar procedures and software tools were mentioned for the different phases of retrodigitisation (image capture, text capture, data encoding and data enrichment). This is reassuring and suggests that there are already some best practices in place for the retrodigitisation workflow, which may be the effect of the ENeL COST Action[11]. For instance, in both surveys, the use of ABBYY FineReader was mentioned for text capture, the oXygen XML editor in relation to data encoding, and outsourcing was mentioned as an option for text capture and image capture.

Of particular interest are the results concerning data enrichment, which means adding additional linguistic and non-linguistic information to the data such as normalizing values, geo-locating, expanding content etc. Different forms of data enrichment were mentioned by the respondents in both surveys, e.g. text normalization, expanding abbreviations, adding grammatical information as well as adding internal and external links. This makes data enrichment a broad and an important task which does not only concern retrodigitsed dictionaries, but also born-digital dictionaries which can be enriched with various types of information. The survey for institutions shows that in contemporary lexicographic projects within the consortium, data enrichment is not yet very common. Only two institutions indicated that they include images and/or videos in their dictionaries.

[11] Within the ENeL COST Action, a lot of attention was paid to retrodigitisation in Working Group 2 “Retro-digitized Dictionaries”, advancing research in the development of a standard workflow for retrodigitisation as well as standards for the encoding and description of information in retrodigitised dictionaries.

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 separate section about more technical matters, such as data formats, metadata and availability was included in the survey for institutions. Overall, we can conclude from this section that dictionary makers started shifting from non-structured data or text format to structured data formats (especially TEI or custom XML). However, the collected information for the low number of the lexicographic projects that use tools for conversion and alignment of different dictionary data formats corresponds with the information about the use of non-structured data formats and shows that the shift from non-structured to structured data formats is still not common practice. Furthermore, it can be noted that the use of standard vocabularies for encoding lexicographic data (e.g. IsoCat), the use of a special metadata schema (e.g. CMDI) and the use of a standard licensing schema (e.g. Creative Commons) are not yet widespread among the lexicographic partner institutions.

The respondents to both surveys noted many positive changes that took place in the field in the last 10-15 years. Most of these changes are connected with the digitisation and automation of lexicographic work, online publishing (moving from paper to online) and with the beginning of corpus era together with access to better and more data (corpora, internet) and better tools (e.g. Sketch Engine). It was pointed out that the task has changed from creating a dictionary to maintaining and expanding a dictionary. Some concerns were also expressed, especially about the quality and reliability of lexicographic data in state-of-the-art lexicography, information overload, and the potentially reduced value of lexicographic skills in digitally oriented projects.

4.1 Some caveats about the surveys and suggestions for future research

We knew from the start that the questionnaire method has its drawbacks, and that the results would also point out aspects where a different type of question, or a different method might have been more appropriate. In this section, we thus point to certain shortcomings of our method, and discuss potential avenues for future research.

The surveys were conducted in Google Forms as the tool was easy to use and administer (for non-technical people), and it covered the majority of our needs, such as easy sharing with people, user-friendly interface, possibility of saving the survey and returning to it at a later point, and the familiarity of the research team with the tool. The only real downside in our case was that Google Forms does not support nesting of questions. This meant that we could not restrict subquestion to subsets of the respondents depending on their answer to the parent question. As a result, some questions were answered by respondents who should not have answered them, which led to some unexpected results requiring further analysis.

In addition, we noted during the analysis of the results that some questions were not clear to the respondents, e.g. one respondent commented that it was not clear what was meant exactly by “outsourcing”. Furthermore, terms such as “born-digital” and “IT support” seem to have been interpreted in different ways by different respondents, even although a definition of “born-digital” was provided. For example, the share of respondents who answered the question whether they

work on born-digital dictionaries affirmatively was unusually high, especially considering the information they provided at related questions about the types of projects, compilation methods and the format of publication, which suggest a different interpretation of the term “born-digital”. This experience shows not only that care needs to be taken in future surveys, but also that there is a need for a better definition of the term in the lexicographic community, something that the ELEXIS project should also pay attention to.

However, overall the decision to include many open-ended questions proved to be correct, even though that this meant a lot of coding. The answers were often detailed and have provided us with information we wanted, sometimes even beyond what was needed/expected. In fact, we would have used even more open-ended questions but we wanted to keep the survey length manageable and not overwhelming for the respondents. It can also be said that in certain cases, an interview would be a better method as it would allow further clarifications from both parties; therefore, we are aiming to combine our results with the results of the interviews conducted as part of WP5 to get an even better insight into lexicographic practices and needs of lexicographers.

The survey for institutions remains open as we expect to extend it to the observers as one of the steps for obtaining information about their projects, workflows and infrastructures. Of course, we intend to resolve the above mentioned shortcomings of the survey first. Moreover, we will add a few additional questions that were identified as helpful when analysing the data. For instance it would be interesting to know what type of personnel is involved in retrodigitisation within an institution, plus additional information about data encoding and data enrichment would be helpful. Also, it would be worth investigating job changes in the field, for example what jobs people had before becoming lexicographers, or have between working on different lexicographic projects. It would be interesting to learn whether they have been doing something lexicography- or language-related before becoming lexicographers. A separate study on freelancers would also be useful, just to understand the difference in working conditions they encounter. Although job changes and the working conditions of freelancers are perhaps more suited for the survey for lexicographers, the findings would definitely be of interest to institutions which makes is worth considering including these questions in the survey for institutions as well.

4.2 Implications for ELEXIS

The two surveys have given the ELEXIS project a detailed insight into lexicographic practices and as such the results provide a valuable input for a number of tasks that will be completed within the project in the next three years. The survey results are particularly relevant for “T1.3 Best Practices for Lexicography”, “T2.1: Common models and protocols for lexicon access”, “WP4 NLP for lexicography” and “WP5 Training and Education” as a whole.

Cooperation on a larger European scale has been limited and standardization efforts have not been particularly successful before the arrival of the digital age. Since linking and online publishing of

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.

lexicographic data require standardised and structured data formats, there has recently been an increase in awareness among dictionary developers of the need to achieve a higher degree of convergence, consistency and adaptability in data formats and data encoding. This tendency is confirmed by the survey results which show a clear need for common standards and solutions. One of the ELEXIS' aims is to establish such common standards and solutions for the development of lexicographic resources. A set of common protocols will be defined (WP2) to improve the interoperability of lexicographic resources and robust documentation, guidelines and collections of best practices will be created (WP1) in order to promote clearly defined workflows for producing, describing and annotating lexicographic resources (both synchronic and diachronic) in accordance with international standards and interoperability formats.

Another ELEXIS objective is to promote an open access culture in lexicography, in line with the European Commission Recommendation on access to and preservation of scientific information. General (open) access to lexicographic data is traditionally very limited and within the ELEXIS infrastructure (in particular WP6) serious efforts will be dedicated to solve IPR issues related to lexicographic data and to enable their integration as linked data. The results of the survey of institutions show that free online distribution is preferred by the academic partner institutions, which is a positive result in the context of promoting open access.

The results from the sections on software and tools provide important input for WP4: NLP for lexicography. ELEXIS will support novel lexicography by providing lexicographers with tools and methods that help them create new resources. As specified in the proposal, two complementary sets of tools will be provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first will include a user-friendly open-source online dictionary writing system, with the aim to provide the central dictionary writing platform for new lexicography which also includes new possibilities of online collaboration. The other will provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification). As such, the ELEXIS project will already fulfill a large number of the wishes and needs that have been expressed in the survey in relation to tools supporting the lexicographic workflow (e.g. user-friendly and intuitive, online, open-source, support for collaborative input, tools for crowdsourcing). Other important features that were frequently mentioned by the survey respondents were that the tools should be interoperable (e.g. it should be possible to integrate data from CQS and other resources into DWS), customisable, browser-independent, fast and should support API access. The availability of support also helps to increase customer satisfaction. This information, together with the full list of wishes and needs (see section 3.1.3.1), will help to fine-tune the development and will ensure that the end product will be embraced by the lexicographic community.

The survey results are also particularly relevant in the context of WP5: Training and Education. Within the infrastructure, (online) tutorials and instruction manuals for ELEXIS services will be created, assessed, revised and disseminated, partly in cooperation with #dariahTeach. In addition, a series of workshops and summer schools will be organised to develop methodological and

technological skills needed for the productive use of and contribution to ELEXIS. This approach to training and education is consistent with current practice in lexicography. The survey results show that specific lexicographic training is often received on the job. In-house training is most common, usually given by a tutor or a senior lexicographer, but external courses, workshops or summer schools are also a popular means of training lexicographers.

In this way ELEXIS will educate a new generation of researchers who understand the full potential of digital research infrastructures to transform their research; who optimally exploit the existing state-of-the art tools; and who are able to create open, standards-compliant lexical datasets that can be fed back into the infrastructures and shared with other researchers. This is particularly important as the role of lexicographers and the tasks they do are changing rapidly. These days, lexicographers have to be technically skilled. The survey shows that quite a lot of lexicographers are actively involved in project management and communication with IT specialists, including user experience and interface designers, carry out user research, conduct interface evaluation, create add-on materials, present and discuss the updates in the media (including social media channels) etc. In sum, the needs of a modern lexicographer extend beyond linguistic knowledge, meaning that continuous training and development in various areas should become a regular part of a lexicographer's job.

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.

Appendix I: A Survey of Lexicographers' Needs

Lexicographic practices: A Survey of Lexicographers' Needs

Welcome to the ELEXIS (European Lexicographic Infrastructure, http://elex.is) survey about lexicographers' needs.

Please help us understand which new software and tools to support the lexicographic workflow should be developed by the ELEXIS project. The aim of the survey is to get an overview of lexicographic practices across Europe and worldwide both for born-digital and retrodigitized resources, and to make an inventory of the needs of lexicographers.

Filling out the survey will take approximately half an hour. There are some open questions which may take longer to answer.

The survey is split in 6 sections: (1) general information; (2) ongoing work; (3) software and tools; (4) publication; (5) retrodigitization; (6) past and future.

Declaration of consent:
Your participation in this research project is completely voluntary. Your responses will remain confidential. Data from this research will be kept secure and reported only as a collective combined total. No one other than the researchers will know your individual answers to this questionnaire. The results of the survey will be published as a deliverable on the ELEXIS website (https://elex.is).
If you agree to participate, please answer the questions in the survey as best you can. If you have any questions about this survey, feel free to contact Jelena Kallas at the Institute of the Estonian Language (jelena.kallas@eki.ee).

By clicking on the „Next" button at the bottom of this page you agree that you have read and understood this declaration of consent.

We thank you for your cooperation.

Institute of the Estonian Language (Jelena Kallas, Margit Langemets)
Dutch Language Institute (Lut Colman, Carole Tiberius)
Bulgarian Language Institute (Svetla Koeva)
Jožef Stefan Institute (Iztok Kosem)

This survey is part of the European project ELEXIS (European Lexicographic Infrastructure, http://elex.is). The ELEXIS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

* Required

1. **Email address** *

Skip to question 33.

(1) General information

Please provide information about yourself.

2. **Country** *

3. **What is your educational background? If other, please specify.** *
Mark only one oval.

- ◯ PhD degree in language/linguistics
- ◯ PhD degree in other humanities (e.g. literature, history, sociology)
- ◯ PhD degree in non-linguistics/non-humanities
- ◯ doctoral student in language/linguistics
- ◯ doctoral student in other humanities (e.g. literature, history, sociology)
- ◯ doctoral student in non-linguistics/non-humanities
- ◯ MA degree in language/linguistics
- ◯ MA degree in other humanities (e.g. literature, history, sociology)
- ◯ MA degree in non-linguistics/non-humanities
- ◯ BA degree
- ◯ Other: _____

4. **How long have you worked as a lexicographer?** *
Mark only one oval.

- ◯ 1-3 years
- ◯ 3-5 years
- ◯ 5-10 years
- ◯ 10-20 years
- ◯ more than 20 years

(2) Ongoing work

Please provide information about ongoing work at your institution.

5. **Do you work as an in-house employee or as a freelancer/self-employed?** *
Mark only one oval.

- ◯ full-time in-house employee
- ◯ part-time in-house employee
- ◯ freelancer/self-employed
- ◯ Other: _____

6. **If you work as an in-house employee, please specify the name of the institution or company you work for.**

_____

7. **Did you receive any specific training as a lexicographer to prepare you for your current work? Select all that apply.** *
Check all that apply.

- ☐ no, I just started to work
- ☐ yes, at the university (e.g. MA in lexicography, special course)
- ☐ yes, within my institute (e.g. by a tutor / senior lexicographer)
- ☐ yes, I have attended special courses (e.g. Lexicom)
- ☐ Other: _____

8. **If you work as part of a team, how big is your team? Please include only lexicographers. Select the best option.** *
Mark only one oval.

- ◯ no, I do not work in a team
- ◯ 2-3 employees/freelancers
- ◯ 3-6 employees/freelancers
- ◯ 5-10 employees/freelancers
- ◯ more than 10 employees/freelancers
- ◯ more than 50 employees/freelancers
- ◯ Other: _____

9. **If you work as part of a team, does it include people from different institutions/countries?**
Mark only one oval.

- ◯ no
- ◯ yes

10. **Please fill in the title of your ongoing project.** *
If you are involved in several projects, please select only one as the questions coming up will be related to that specific project.

_____

11. **When did the project start?** *
Mark only one oval.

- ◯ before 2010
- ◯ 2011-2014
- ◯ 2015-2018

12. **When will the project end?** *
Mark only one oval.

- ◯ 2018-2019
- ◯ 2020-2023
- ◯ 2024-
- ◯ permanent updating / permanent development

13. **What type of project is it? If other, please specify.** *
Mark only one oval.

- ◯ monolingual
- ◯ bilingual
- ◯ multilingual
- ◯ Other: _____

14. **What kind of data does your project cover? If other, please specify.** *
Mark only one oval.

- ◯ general language (general dictionary)
- ◯ specific area of language (e.g. dictionary of collocations, phrasal verbs, synonyms, rhyming)
- ◯ encyclopedias (encyclopedic and cultural material)
- ◯ terminology (e.g. dictionary of legal terms, accounting)
- ◯ Other: _____

15. **How is the database of your project organised: from word to meaning (word-based) or from meaning to word (concept-based). If other, please specify.** *
Mark only one oval.

- ◯ from word to meaning (word-based), i.e. this word/term has several meanings/senses
- ◯ from meaning to word (concept-based), i.e. this meaning/sense has several words/terms
- ◯ Other: _____

16. **Is your project a born-digital dictionary (i.e. a dictionary conceptualized for the electronic medium, offering radically different options for organisation and presentation of lexical information)? If other, please specify.** *
Mark only one oval.

- ◯ yes
- ◯ no
- ◯ Other: _____

17. **How did you compile your dictionary? If other, please specify.** *
Mark only one oval.

- ◯ manually
- ◯ semi-automatically with manual post-editing
- ◯ fully-automatic with manual post-editing
- ◯ fully-automatic with no post-editing
- ◯ Other: _____

18. **Do you have IT support in your work?** *
Mark only one oval.

- ◯ yes
- ◯ no

19. **If you have IT support in your work, are you satisfied with the amount of IT support? Please describe the situation in a few words.**

20. **Are there any IT companies outside your institution involved in your project?** *
Mark only one oval.

- ◯ yes
- ◯ no
- ◯ I do not know

21. **If your project uses IT companies outside your institution, does it affect your workflow? Please describe how it affects your workflow in a few words, mentioning the pros and cons.**

(3) Software and Tools

Please provide information about software and tools you are using for your work.

22. **What is your technical expertise? Do you use special software for your work? Select all that apply.** *
Check all that apply.

- ☐ no, I do not use any special software (i.e. I use Word/Excel etc.)
- ☐ yes, I use a Dictionary Writing System (DWS) (e.g. in-house, TLex, IDM, Lexonomy, Multiterm)
- ☐ yes, I use a Corpus Querying software (e.g. Sketch Engine, KORP)
- ☐ yes, I use software for retrodigitizing the dictionaries (e.g. OCR, ABBYY FineReader, oXygen XML)
- ☐ Other:

23. **If you use a Dictionary Writing System (DWS) for your work, please provide its name and Internet address (url) or other reference.**
For a widely known software (e.g. Tlex, IDM, Lexonomy), the name will do.

24. **Please provide information about the functionalities of your DWS/dictionary editing software which you particularly like and/or dislike or you think are important to mention.**

25. **What are your wishes concerning DWS/dictionary editing software to make your work more effective?**

26. **If you use a Corpus Querying System for your work, please provide its name and Internet address (url) or other reference.**
For a widely known software (e.g. Sketch Engine), the name will do

27. **Please provide information about the functionalities of your Corpus Query System which you particularly like and/or dislike or you think are important to mention.**

28. **What are your wishes concerning Corpus Querying software to make your work more effective?**

29. **If you use a Corpus Query System in your work, what kind of data do you obtain from your Corpus Query System. Select all that apply. If other, please specify.** *
Check all that apply.

- [ ] I do not use special corpus query software
- [ ] lemmas for headword list (e.g. based on frequencies)
- [ ] neologisms
- [ ] form variants (e.g. irregular morphology, orthographic variants)
- [ ] multiword expressions
- [ ] frequency information (e.g. for lemmas, for morphological forms)
- [ ] collocations
- [ ] patterns (e.g. syntactic patterns, valency patterns)
- [ ] definitions
- [ ] word senses
- [ ] diachronic distribution of senses
- [ ] lexical-semantic relations (e.g. synonyms, antonyms, hypernyms)
- [ ] example sentences
- [ ] information on register (e.g. colloquial, formal, slang, offensive terms)
- [ ] domain information (e.g. legal terms, accounting)
- [ ] multilingual data from parallel/comparable corpora (for bilingual/multilingual dictionaries)
- [ ] audio data from speech corpora
- [ ] knowledge rich contexts (a hybrid of a good dictionary example and a definition)
- [ ] Other:

30. **Does your project include data which is automatically extracted by a software program and post-edited by a lexicographer? If yes, which information categories are automatically extracted and post-edited? Select all that apply.** *

Check all that apply.

- [ ] no, the project does not include automatically extracted data
- [ ] extraction of headword list (e.g. based on frequencies)
- [ ] detection of neologisms
- [ ] extraction of form variation (e.g. irregular morphology, orthographic variants)
- [ ] extraction of multiword expressions
- [ ] extraction of frequency information (e.g. for lemmas, for morphological forms)
- [ ] extraction of collocations
- [ ] extraction of patterns (e.g. syntactic patterns, valency patterns)
- [ ] extraction of definitions / definition finding
- [ ] extraction of word senses
- [ ] diachronic distribution of senses
- [ ] extraction of lexical-semantic relations (e.g. synonyms, antonyms, hypernyms)
- [ ] extraction of dictionary examples (e.g. GDEX in SketchEngine)
- [ ] extraction of register information (e.g. colloquial, formal, slang, offensive terms)
- [ ] extraction of domain information (e.g. legal terms, accounting)
- [ ] extraction of multilingual data from parallel/comparable corpora (for bilingual/multilingual dictionaries)
- [ ] extraction of audio data from speech corpora
- [ ] extraction of knowledge rich contexts (a hybrid of a good dictionary example and a definition)
- [ ] Other: _____

(4) Publication

Please provide information about publishing your work.

31. **What is the publishing medium for your work? Select all that apply. If other, please specify.** *

Check all that apply.

- [ ] only in print
- [ ] only online / on the web
- [ ] both, in print and on the web
- [ ] Other: _____

32. **If you are involved in the online publication of your dictionary, what kind of work do you do? Select all that apply.** *

Check all that apply.

- [ ] no, I am not involved in the online publication of my dictionary
- [ ] evaluating the user interface and giving new ideas
- [ ] creating add-on materials (e.g. blogs, slideshows, videos, quizzes, word games)
- [ ] communicating with IT persons / User Experience designer (UX)/Interface designer (IX)
- [ ] Other: _____

33. **If you are involved in any kind of user research (e.g. log files, questioning) for your dictionary, what kind of work do you do? Select all that apply.** *

Check all that apply.

- [ ] no, I am not involved in user research for my dictionary
- [ ] analysing user logs
- [ ] interviewing end users
- [ ] Other: _____

(5) Retrodigitization

This section is about retrodigitization. If you have not been involved in retrodigitization, you can skip this section, except for the last question, and you can continue with the final section.

34. **Do or did you participate in retrodigitizing dictionaries? Please, select all that apply. If other, please specify.** *
Check all that apply.

- [ ] no, I have not been involved in retrodigitization
- [ ] yes, image capture (using scanners or cameras)
- [ ] yes, text capture (OCR, or keying (i.e. typing), proofreading etc.)
- [ ] yes, data encoding (structural, i.e. semantic markup, using XML, whether TEI or not)
- [ ] yes, data enrichment (such as normalizing values, geo-locating, expanding content etc.)
- [ ] Other: _____

35. **If you have been involved in image capture, provide a short description of what you have done and the software you have used (name, Internet address (url) or other reference).**

_____

_____

_____

_____

_____

36. **If you have been involved in text capture, provide a short description of what you have done and the software you have used (name, Internet address (url) or other reference).**

_____

_____

_____

_____

_____

37. **If you have been involved in data encoding, provide a short description of what you have done and the software you have used (name, Internet address (url) or other reference).**

_____

_____

_____

_____

_____

38. **If you have been involved in data enrichment, provide a short description of what you have done and the software you have used (name, Internet address (url) or other reference).**

_____

_____

_____

_____

_____

39. **Can you name a dictionary in your institution/country that should definitely be retrodigitized? Why?**

_____

_____

_____

_____

_____

(6) Past and future

The last two questions are about the past and future. Please feel free to answer these questions in your own language if you feel more comfortable writing in your own language.

40. **Think back for about 10-15 years. What are the major changes in your work, if any. What do you like better now? Or what do you dislike?**

41. **Think forward for about 10-15 years. What might be the major changes in your work, if any. Could you identify some of your wishes and needs? Feel free to write in your own language, if it helps.**

42. **May we contact you for follow-up information?** *
Mark only one oval.
- yes
- no

43. **Would you like to subscribe to the ELEXIS newsletter?** *
Mark only one oval.
- yes
- no

44. **If we may contact you or if you would like to subscribe to the ELEXIS newsletter, please provide your name and e-mail address**

Thank you!

We appreciate that you have taken the time to complete this survey. For those who are attending the Euralex conference in Ljubljana from 17 July til 21 July, we will be there to answer any questions you may have about the survey. Please come and find us at the ELEXIS booth.

Appendix II: A Survey of Lexicographers' Needs for Institutions

Lexicographic practices: A Survey of Lexicographers' Needs for Institutions

Welcome to the ELEXIS survey about lexicographers' needs.

The aim of this survey is to get an overview of lexicographic practices both for born-digital and retrodigitized resources, and to make an inventory of the needs of lexicographers. The results of this survey will feed back into the ELEXIS project and new software and tools will be developed to support the lexicographic workflow.

This survey is addressed at institutions. One survey needs to be completed per institution. The institution should determine the representative who should complete the survey. He/she might be a senior lexicographer or a computational lexicographer/linguist. The expertise of a computational linguist or IT specialist will most likely be required to answer some of the questions.

The survey consists of 6 sections:
(1) General information; (2) Types of lexicographic resources, software and tools supporting the workflow; (3) Publication and access. Crowdsourcing and gamification; (4) Retrodigitized dictionaries; 5) Data formats. Metadata. Availability; 6) Past and Future.

Your responses will remain confidential. Data from this research will be kept secure and reported only as a collective combined total. No one other than the researchers will know the individual answers to this questionnaire. The results of the survey will be published as a deliverable on the ELEXIS website (https://elex.is).

If you have any questions about this survey, feel free to contact Jelena Kallas at the Institute of the Estonian Language (jelena.kallas@eki.ee).

The survey will take approximately 45 mins to 1 hour to complete. Please take your time to answer the questions as best as you can. It is possible to stop at the end of each section and to continue later.

We thank you for your cooperation.

Institute of the Estonian Language (Jelena Kallas, Margit Langemets)
Dutch Language Institute (Lut Colman, Carole Tiberius)
Bulgarian Language Institute (Svetla Koeva)
Jožef Stefan Institute (Iztok Kosem)

This survey is part of the European project ELEXIS (European Lexicographic Infrastructure, http://elex.is). The ELEXIS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

* Required

1. **Email address** *

Skip to question 1.

(1) General information

Please provide general information about your institution and your background.
NOTE: You will answer the survey as the representative of your institution.

2. **Your name** *

3. **Name of your institution** *

4. **Country** *

5. **What is your position within the institution? Select all that apply. If other, please specify.** *

Check all that apply.

- [ ] lexicographer / terminologist
- [ ] member of the board / council of the institution
- [ ] corpus linguist / computational lexicographer / computational linguist
- [ ] IT person / software developer
- [ ] Other: _____

6. **What is your educational background?** *
Mark only one oval.

- ◯ PhD degree in language/linguistics
- ◯ PhD degree in other humanities (e.g. literature, history, sociology)
- ◯ PhD degree in non-linguistics/non-humanities
- ◯ doctoral student in language/linguistics
- ◯ doctoral student in other humanities (e.g. literature, history, sociology)
- ◯ doctoral student in non-linguistics/non-humanities
- ◯ MA degree in language/linguistics
- ◯ MA degree in other humanities (e.g. literature, history, sociology)
- ◯ MA degree in non-linguistics/non-humanities
- ◯ BA degree
- ◯ Other: _____

7. **How long is your experience with dictionaries, lexical databases, different lexicographic/terminological projects?** *
Mark only one oval.

- ◯ 1-3 years
- ◯ 3-5 years
- ◯ 5-10 years
- ◯ 10-20 years
- ◯ more than 20 years

8. **How would you characterize yourself with regard to traditional lexicography vs. modern e-lexicography / paper vs. e-dictionaries? Select one option.** *
Mark only one oval.

- ◯ I feel more comfortable with traditional lexicography (paper slips, writing in Word, paper dictionaries)
- ◯ I clearly prefer e-lexicography (corpora, dictionary writing systems, born-digital dictionaries, e-publishing)
- ◯ I feel comfortable with both, traditional and e-lexicography
- ◯ I am used to work electronically, but I think dictionaries should be printed (in addition to e-dictionary)
- ◯ Other: _____

9. **Is your institution partner or observer in ELEXIS? Select your option.** *
Mark only one oval.

- ◯ partner of ELEXIS
- ◯ observer in ELEXIS
- ◯ no, we are not involved in ELEXIS

10. **Type of your organisation. Select your option. If other, please specify.** *
Mark only one oval.

- ◯ public institution, non-profit organisation (NGO) (eg. National Institute, National Center or Society)
- ◯ private / commercial company
- ◯ department of the University (usually legal person in public law)
- ◯ mixture of public and private (public-private partnership, PPP)
- ◯ Other: _____

11. **Funding of the lexicographic work at your institution. Select your (best) option. If other, please specify.** *
Mark only one oval.

- ◯ funded on a regular basis (eg. stable funding by government/ministry/academy)
- ◯ funded on an irregular bases (eg. via different/occasional projects)
- ◯ partly on a regular basis, partly on an irregular basis
- ◯ private equity (commercial businesses)
- ◯ funded by a private person or fund
- ◯ Other: _____

12. **How many lexicographers work at your institution? Please sum up into full-time employment. Select your option.** *
Mark only one oval.

- ◯ 1-10
- ◯ 11-25
- ◯ 26-50
- ◯ more than 50

13. **Do lexicographers work on lexicographic projects only or do they also have other tasks? Please specify briefly what other tasks they do (e.g. teaching, management, public relations)?** *

_____
_____
_____
_____
_____

14. **Do lexicographers receive any specialized lexicographic training as part of their job? Where (e.g. in-house or university course), how often, what kind of training do they receive?** *

_____
_____
_____
_____
_____

15. **How many software developers / IT persons work at your institution? Please sum up into full-time employment and specify their availability for lexicographic work?** *

_____

16. **Do you outsource parts of your lexicographic work to an IT company or language technology company? If other, please specify.** *
Check all that apply.

- ☐ no, we do not outsource
- ☐ yes, we are outsourcing
- ☐ yes, we have outsourced (in the past)
- ☐ Other: _____

17. **If you do outsource, what kind of lexicographic work is or has been outsourced to an IT company or language technology company? Select all that apply. If other, please specify.**
Check all that apply.

- ☐ development of a dictionary writing system (DWS)
- ☐ development of a corpus query system (CQS)
- ☐ database preparation/development
- ☐ development of a user interface
- ☐ Other: _____

18. **Do you want to save your result and quit for now?**
IMPORTANT: You do need to save the link 'Edit this form' which will appear on the next page on your computer and then you will be able to return to the survey at a later time using this link.
Mark only one oval.

- ◯ Yes *Stop filling out this form.*
- ◯ No

(2) Ongoing work. Types of lexicographic resources, Software and Tools

NOTE: You will answer the survey as the representative of your institution. The expertise of a computational linguist or IT person may be required to answer some of the questions.

19. **What do you consider as the lexicographic expertise of your institution? Select all that apply. If other, please specify.** *
Check all that apply.

- ☐ monolingual general dictionaries (modern, synchronic)
- ☐ monolingual specialized dictionaries (e.g. dictionary of collocations, phrasal verbs, synonyms, rhyming)
- ☐ historical dictionaries (e.g. diachronic, etymological, old literary language)
- ☐ dialect dictionaries
- ☐ bilingual or multilingual general dictionaries
- ☐ multilingual terminological or specialized dictionaries(e.g. dictionary of legal terms, accounting)
- ☐ Other: _____

20. **Total amount of lexicographic resources at your institution.** *
Mark only one oval.

- ◯ 1-5
- ◯ 5-10
- ◯ 10-50
- ◯ 50-100
- ◯ more than 100

21. **Please list lexicographic projects (max. 3-4) that have been started recently (2014-2018) or will start in the near future (2018-2021). Please provide name, short description, Internet address (url) or other reference.** *

_____
_____
_____
_____
_____

22. **Please list lexicographic projects (max. 3-4) that will be published in the near future (2018-2021). Please provide name, short description, Internet address (url) or other reference.** *

_____
_____
_____
_____
_____

23. **Do you use a dictionary writing system (DWS) or any specialized editing software to produce dictionary(-like) products at your institution? Select all that apply.** *

Check all that apply.

- [ ] no
- [ ] no, but we feel we need one urgently
- [ ] yes, commercial DWS (e.g. TLex, IDM)
- [ ] yes, open-source DWS (e.g. Lexonomy)
- [ ] yes, in-house DWS
- [ ] yes, we use 2 or more DWSs (e.g. one for lexicographers, one for terminologists)
- [ ] Other: _____

24. **If you use a DWS or any specialized editing software, what do you use? Please provide its name, Internet address (url) or other reference.**

For a widely known software (e.g. Tlex, IDM, Lexonomy), the name will do.

_____

_____

_____

_____

_____

25. **If you use 2 or more DWSs, please specify the reason for using more than one?**

_____

_____

_____

_____

_____

26. **If you use an off-the-shelf DWS (commercial or open-source), did you make any adaptations/customizations (within or outside your institution) to make it more suitable for your lexicographic project(s)?** *

Mark only one oval.

- ◯ yes
- ◯ no

27. **If you made any adaptations/customizations to the DWS, please describe them in a few words.**

_____

_____

_____

_____

_____

28. **If you use a a DWS or any other specialized editing software, how satisfied are you with the software you use at the moment? Why?**

_____

_____

_____

_____

_____

29. **If you do not use a DWS or any specialized dictionary editing software at your institution, please specify why not? Please describe your workflow in a few words.**

_____

_____

_____

_____

_____

30. **Do you use a Corpus Query System (CQS) or any specialized software to work with corpora at your institution? Select all that apply. If other, please specify.** *
Check all that apply.

- [ ] no
- [ ] no, but we feel we need one urgently
- [ ] yes, commercial CQS (e.g. Sketch Engine)
- [ ] yes, open-source CQS (e.g. KORP, BlackLab)
- [ ] yes, in-house CQS
- [ ] Other: _____

31. **If you use a CQS or any specialized corpus software, what do you use? Please provide its name, Internet address (url) or other reference.**

_____

32. **If you use a CQS, how satisfied are you with the system you use at the moment? Why?**

_____
_____
_____
_____
_____

33. **Do you have any additional wishes for a CQS? What would be the most important function to be added? (e.g. clustering the concordances against senses, additional views on the data).** *

_____
_____
_____
_____
_____

34. **Can you integrate data from your Corpus Query System directly into your Dictionary Writing System?** *
Mark only one oval.

- ◯ yes
- ◯ no

35. **If you use a DWS and CQS, are they integrated into one piece of software?**
Mark only one oval.

- ◯ yes
- ◯ no

36. **If your DWS and CQS are not integrated, would you like them to be integrated? What kind of functionalities would you be interested in?**

_____
_____
_____
_____
_____

37. **If you use automatic data extraction / automatic knowledge extraction at your institution, please mark the kind of data that is / has been automatically extracted for different projects. Select all that apply.** *

Check all that apply.

- [ ] we do not use special extraction software
- [ ] extraction of headword list (e.g. based on frequencies)
- [ ] detection of neologisms
- [ ] extraction of form variation (e.g. irregular morphology, orthographic variants)
- [ ] extraction of multiword expressions
- [ ] extraction of frequency information (e.g. for lemmas, for morphological forms)
- [ ] extraction of collocations
- [ ] extraction of patterns (e.g. syntactic patterns, valency patterns)
- [ ] extraction of definitions / definition finding
- [ ] extraction of word senses
- [ ] diachronic distribution of senses
- [ ] extraction of lexical-semantic relations (e.g. synonyms, antonyms, hypernyms)
- [ ] extraction of dictionary examples (e.g. GDEX in Sketch Engine)
- [ ] extraction of register information (e.g. colloquial, formal, slang, offensive terms)
- [ ] extraction of domain information (e.g. legal terms, accounting)
- [ ] extraction of multilingual data from parallel/comparable corpora (for bilingual/multilingual dictionaries)
- [ ] extraction of audio data from speech corpora
- [ ] extraction of knowledge rich contexts (a hybrid of a good dictionary example and a definition)
- [ ] Other: _____

38. **Are there any lexicographic projects at your institution based totally on post-editing of automatically extracted data? If other, please specify.** *

Check all that apply.

- [ ] no, there are not
- [ ] yes, all the raw material is or has been extracted from the corpus
- [ ] yes, some data is or has been extracted from the corpus (e.g. examples, frequency information)
- [ ] Other: _____

39. **If there are lexicographic projects at your institution based totally on post-editing of automatically extracted data, please list them (max 3-4) and provide references/urls, if possible.**

_____

_____

_____

_____

_____

40. **Have you got any additional wishes for automatic data extraction / automatic knowledge extraction for lexicography?** *

_____

_____

_____

_____

_____

41. **Do you reuse existing lexicographic data within your institution in new projects (e.g. integrate spelling/etymological information in a new dictionary project)?** *

Mark only one oval.

- ◯ no, we do not (e.g. all projects are standalone projects)
- ◯ yes, we do reuse existing lexicographic data in new projects

42. **If you reuse/integrate lexicographic data from other lexicographic projects within your institution in a new project or have done so in the past, please specify in a few words.**

43. **Do you want to save your result and quit for now?**
You do need to save the link 'Edit this form' which will appear on the next page on your computer and then you will be able to return to the survey at a later time using this link.
Mark only one oval.

- ◯ Yes *Stop filling out this form.*
- ◯ No

(3) Ongoing work. Publication and access. Crowdsourcing and gamification

Please provide information about 1) publication and access, 2) crowdsourcing and gamification of lexicographic resources at your institution.
NOTE: You will answer the survey as the representative of your institution. The expertise of a computational linguist or IT person may be required to answer some of the questions.

44. **What kind of publishing medium have you used since 2010 for lexicographic data at your institution? Select all that apply. If other, please specify.** *
Check all that apply.

- ☐ scanned or photographed electronic dictionary (pdf or jpg)
- ☐ online dictionary, looking like paper dictionary
- ☐ online dictionary, much more dynamic than paper dictionary
- ☐ desktop web page without responsive design for mobile devices
- ☐ desktop web page with responsive design for mobile devices
- ☐ App
- ☐ Other:

45. **If you make your lexicographic data available as an App, please provide name, on which platforms the App is available (e.g. Android, iOS) and give information on the technology/software you use/have used.**

46. **Does your software (DWS or other) offer the functionality of dictionary publishing? Select all that apply. If other, please specify.** *
Check all that apply.

- ☐ we do not use special software
- ☐ export for printing (pdf, Indesign etc.)
- ☐ export for publishing online (e.g. 'click-to-publish')
- ☐ export for saving
- ☐ automatic creation of metadata
- ☐ Other:

47. **Consider the main ongoing/new projects (max 3-4) that will be published in the near future (2018-2021). How will they be published? If other, please specify.** *
Check all that apply.

- ☐ online
- ☐ print
- ☐ Other:

48. **If you publish online, can the user customize the interface and the metalanguage while using the online dictionary / web site? Select all that apply. If other, please specify.** *
Check all that apply.

- [ ] no, customization is not possible
- [ ] interface customization (e.g. changing from L1 to L2, according to the user's language)
- [ ] meta-language customization (e.g. labels within entries from L1 to L2, according to the user's language)
- [ ] Other: _____

49. **If you have a website/portal for your dictionaries, how would you describe the access to the data? Select the best option. If other, please specify.** *
Check all that apply.

- [ ] no, we do not have a website
- [ ] each dictionary has its own website
- [ ] dictionary collection (i.e. only external access by means of hyperlinks to the individual dictionaries, e.g.Slang Portal)
- [ ] dictionary search engine (i.e. access to articles in the individual dictionaries, e.g. OneLook)
- [ ] dictionary net (i.e. access to elements within the articles of the individual dictionaries, e.g. Owid, Canoonet)
- [ ] Other: _____

50. **If you have a website/portal, does it provide free text search?**
Mark only one oval.

- ◯ yes
- ◯ no

51. **If you have a website/portal, does your website provide filtering/faceted browsing (e.g. like booking.com or big e-stores)?**
Mark only one oval.

- ◯ yes
- ◯ no

52. **If you have a website/portal, does your website provide API access?**
Mark only one oval.

- ◯ yes
- ◯ no

53. **If your website/portal provides API access, please give the url of the API.**

_____

54. **If you have a website/portal, does your website provide SPARQL querying?**
Mark only one oval.

- ◯ yes
- ◯ no

55. **If your website provides SPARQL querying, please give the url of the SPARQL endpoint.**

_____

56. **Does your dictionary website/portal provide search options for the following features? Select all that apply.** *

Check all that apply.

- [ ] we do not have a dictionary website
- [ ] lemma
- [ ] inflected forms
- [ ] senses
- [ ] entry structure (e.g. sense groupings)
- [ ] definitions
- [ ] etymology
- [ ] syntactic Information (e.g. part-of-speech, gender)
- [ ] usage notes
- [ ] historical usage information
- [ ] relation to other entries (e.g. synonyms, hypernyms, antonyms)
- [ ] metadata
- [ ] Other: _____

57. **Does your online dictionary offer a link to corpus data? Select all that apply to your practice. If other, please specify** *

Check all that apply.

- [ ] no, we do not link the online dictionary to corpus data
- [ ] yes, through API
- [ ] direct links from DWS clients into CQS to access corpus data
- [ ] entries contain automatic URL pointing to CQS for the given headword
- [ ] Other: _____

58. **If you link your online dictionary to corpus data, can you customize this linking? Can the user specify which elements he/she wants to retrieve from the corpus or not (e.g. example sentences with metadata/without metadata)?**

_____

_____

_____

_____

_____

59. **Do you use crowdsourcing at your institution or have you used crowdsourcing in the past? Select your option.** *

Mark only one oval.

- ◯ no
- ◯ yes

60. **If you use crowdsourcing at your institution or have done so in the past, please give a short description of the project(s) and the software you use.**

_____

_____

_____

_____

_____

61. **Do you use gamification at your institution or have you used gamification in the past? Select your option.** *

Mark only one oval.

- ◯ no
- ◯ yes

62. **If you use gamification in your lexicographic projects or have done so in the past, please give a short description of the project(s) and the software you use.**

63. **What kind of multi-modal data (images, videos) from publicly available resources do you use or have you used to enrich lexicographic data? Select all that apply. If other, please specify.** *
Check all that apply.

- [ ] we do not use any multi-modal data from the web
- [ ] images (e.g. from Flickr, Wikimedia Commons, Europeana)
- [ ] video material (e.g. from Videolectures.net)
- [ ] Other:

64. **Do you want to save your result and quit for now?**
You do need to save the link 'Edit this form' which will appear on the next page on your computer and then you will be able to return to the survey at a later time using this link.
Mark only one oval.

- ◯ Yes *Stop filling out this form.*
- ◯ No

(4) Retrodigitized dictionaries

This section is about retrodigitization. If you have not been involved in retrodigitization, you can skip this section, except for the last question, and you can continue with the next section.

65. **Does or did your institute participate in retrodigitizing dictionaries? Please, select all that apply. If other, please specify.**
Check all that apply.

- [ ] no, my institute has not been involved in retrodigitization
- [ ] yes, image capture (using scanners or cameras)
- [ ] yes, text capture (OCR, or keying (i.e. typing), proofreading etc.)
- [ ] yes, data encoding (structural, i.e. semantic markup, using XML, whether TEI or not)
- [ ] yes, data enrichment (such as normalizing values, geo-locating, expanding content etc.)
- [ ] Other:

66. **If your institute has been involved in image capture, provide a short description of what you have done and the software you have used (name/url).**

67. **If your institute has been involved in text capture, provide a short description of what you have done and the software you have used (name/url).**

68. **If your institute has been involved in data encoding, provide a short description of what you have done and the software you have used (name/url).**

69. **If your institute has been involved in data enrichment, provide a short description of what you have done and the software you have used (name/url).**

70. **Have you managed to integrate the retrodigitized dictionary to your dictionary website/portal? Select the (best) option for your institution. If other, please specify.**
Mark only one oval.

- ◯ no, we have kept them standalone
- ◯ yes, it is one of the dictionaries in the set with access to the dictionary via a hyperlink
- ◯ yes, it is one of the dictionaries in the set with access to entries within the dictionary
- ◯ yes, its content is integrated into an aggregator with access to data within entries
- ◯ Other: _____

71. **How can users access your retrodigitized dictionaries? Check all that apply.**
Check all that apply.

- ☐ through an institutional portal
- ☐ through an API
- ☐ by downloading image files
- ☐ by downloading full text
- ☐ Other: _____

72. **If you do not share files containing the full text of your retrodigitized dictionaries with your users, what are the main reasons for that?**

73. **Can you name a dictionary in your institution/country that should definitely be retrodigitized? Why?**

74. **Do you want to save your result and quit for now?**
You do need to save the link 'Edit this form' which will appear on the next page on your computer and then you will be able to return to the survey at a later time using this link.
Mark only one oval.

- ◯ Yes *Stop filling out this form.*
- ◯ No

(5) Data formats. Metadata. Availability

Please provide information about technical matters at your institution. We will ask about 1) data formats; 2) metadata; 3) availability. By metadata we mean data about data: information describing properties of linguistic resources, for instance, the size of a corpus, the recording date of a specific file, the purpose for which annotations were created. (https://www.clarin.eu/faq-page/273#t273n2850)
NOTE: You will answer the survey as the representative of your institution. The expertise of an IT person or a software developer may be required to answer these questions.

75. **What data format(s) do you use for lexicographic projects at your institution? Select all that apply to your practice. If other, please specify** *
Check all that apply.

- [ ] non-structered data format / text format (e.g. Word)
- [ ] table format (e.g. CSV, TSV, XLS)
- [ ] database (e.g. relational database)
- [ ] XML
- [ ] Resource Description Framework (RDF)
- [ ] Other: _____

76. **If you use XML, do you use**
Mark only one oval.

- ◯ custom XML
- ◯ LMF
- ◯ TEI
- ◯ TEI-lex0
- ◯ Other: _____

77. **If you use TEI, which TEI-version do you use?**

_____

78. **Do you have tools that allow automatic conversion and alignment of different dictionary data formats (e.g. from database format to XML)?** *
Mark only one oval.

- ◯ yes
- ◯ no

79. **Do you use existing standard vocabularies for encoding your lexicographic data? Select all that apply. If other, please specify.** *
Check all that apply.

- [ ] no, we don't
- [ ] IsoCat
- [ ] Clarin Concept Registry
- [ ] Lemon-Ontolex
- [ ] Lexinfo
- [ ] GOLD
- [ ] TEI
- [ ] Other: _____

80. **Do you use a special metadata schema? Select all that apply.** *
Check all that apply.

- [ ] no, we do not have metadata
- [ ] no, but we try to move towards a standard metadata schema
- [ ] META-SHARE metadata schema v3.0 (in the CLARIN Component Registry)
- [ ] CMDI
- [ ] Dublin Core
- [ ] OLAC
- [ ] TEI-header
- [ ] Other: _____

81. **Do you use a special tool for metadata creation and editing? Select your option.** *
Mark only one oval.

- ◯ yes
- ◯ no

82. **If you use a special tool for metadata creation and editing, please specify (name/url).**

83. **If you don't use any tool, do you feel it would be necessary/easier to use a special tool for metadata creation and editing? Please explain briefly.**

84. **How do you make your dictionary data available? Select all that apply. If other, please specify** *
Check all that apply.

- ☐ free online
- ☐ restricted online / for usage fee
- ☐ both (some for free, others restricted or for usage fee)
- ☐ paper dictionary (paid)
- ☐ (paid) paper dictionary first, later online for free (e.g. after 1 year)
- ☐ (paid) paper dictionary first, later online for usage fee (e.g. after 1 year)
- ☐ Other:

85. **How can other applications access your dictionary content? Select all that apply. If other, please specify** *
Check all that apply.

- ☐ free API access (e.g. retrieve list of words, retrieve dictionary information for a given word)
- ☐ paid API access (e.g. retrieve list of words, retrieve dictionary information for a given word)
- ☐ free download and using under certain licence
- ☐ paid download and using under certain licence
- ☐ Other:

86. **Do you make use of a standard licensing schema for your lexicographic data? If other, please specify.** *
Check all that apply.

- ☐ no
- ☐ yes, CLARIN licensing framework
- ☐ yes, Creative Commons
- ☐ yes, Open Data Commons
- ☐ Other:

87. **Which forms of access that you do not support, do you think would be useful for your users?**

88. **How do you deal with version control and archiving of different versions of the dictionary?**

89. **Do you want to save your result and quit for now?**
You do need to save the link 'Edit this form' which will appear on the next page on your computer and then you will be able to return to the survey at a later time using this link.
Mark only one oval.

- ◯ Yes *Stop filling out this form.*
- ◯ No

(6) Past and Future

90. **Think back for about 10-15 years. What are the major changes in your lexicographic projects, if any. What do you like better now? Or what do you dislike?** *

91. **Think forward for about 10-15 years. What might be the major changes in your lexicographic projects, if any. Can you identify some of your wishes and needs?** *

Thank you

We appreciate that you have taken the time to complete this survey. For those who are attending the Euralex conference in Ljubljana from 17 July til 21 July, we will be there to answer any questions you may have about the survey. Please come and find us at the ELEXIS booth.

A copy of your responses will be emailed to the address you provided