

D6.3

INTERMEDIATE
INTEROPERABILITY
REPORT

Author(s): Iztok Kosem, Roberto
Navigli, John McCrae, Miloš Jakubíček

Date: 31. 1. 2021

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

6.3. INTERMEDIATE INTEROPERABILITY REPORT

Deliverable Number: 6.3

Dissemination Level: Public

Delivery Date: 31. 1. 2021

Version: 1.0

Author(s): Iztok Kosem

Roberto Navigli

Miloš Jakubiček

John McCrae

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

Deliverable/Document Information

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

Document History

Version Date	Changes/Approval	Author(s)/Approved by
08/01/2021	First draft	Iztok Kosem
21/01/2021	Specific sections added	Federico Martinelli, Ondrej Matuska, Anna Woldrich, Toma Tasovac
27/01/2021	Feedback	Roberto Navigli, Miloš Jakubiček, John McCrae
31/01/2021	Final version	Iztok Kosem, Simon Krek

D6.3 Intermediate interoperability report.

Table of Contents

1	Executive summary	3
2	Tools and services	4
2.1	Sketch Engine	4
2.2	One-click dictionary	6
2.3	Lexonomy	7
2.4	Elexifier	12
2.4.1	Case study 1 (XML): Converting and aligning Wordnet lexical databases between languages	13
2.4.2	Case study 2 (PDF): Converting an 18th century Spanish-Basque-Latin Trilingual Dictionary.....	13
2.5	VerbAtlas.....	14
2.6	SyntagNet.....	15
2.7	NAISC.....	16
2.8	Game of words.....	17
2.9	TEI Lex-0	17
2.10	Elexifinder	19
2.10.1	Usage statistics.....	20
2.11	ELEXIS Lexicographic Newsfeed.....	22



D6.3 Intermediate interoperability report.

List of Tables

Table 1: Institutions per country.....	5
Table 2: ELEXIS partner and observer lexical resources in Lexonomy.....	12
Table 3: Most frequently searched concepts in Elexifinder (July 2019 - January 2021).....	21
Table 4: Most frequently searched keywords in Elexifinder (July 2019 - January 2021).....	22
Table 5: Top 10 ranking of most popular elex.is pages.....	23
Table 6: lexicographic Newsfeed - page views over time period (Aug 2019 - incl. Jan 2021).	23
Table 7: Where Newsfeed users come from: source/medium.....	24
Table 8: Newsfeed users per country (top 10) over time period (Aug 2019 - incl. Jan 2021).	25

List of Figures

Figure 1: Graphic Guide to ELEXIS Dictionary Tools.....	3
Figure 2: Institutions per country (in percentages).	5
Figure 3: Commits on Lexonomy GitHub (excluding merges).	7
Figure 4: VerbAtlas user statistics.....	15
Figure 5: SyntagNet user statistics.....	16
Figure 6: First author locations by country (Elexifinder).	19
Figure 7: Newsfeed postings per Social Media platform over time period.....	24



1 Executive summary

Since the last interoperability report, several new tools have been added to the ELEXIS infrastructure ecosystem, while existing ones have been upgraded and improved. As is evident from the report, the main part of the user activity has been focussed on ELEXIS dictionary tools, specifically on the conversion (Elexifier), creation (Sketch Engine, One-Click Dictionary, Lexonomy), editing (Lexonomy) and publication (Lexonomy) of dictionary content (see also Figure 1). These four tools exhibit a high level of interoperability, which is also evident from the number of users and resources produced with them. In addition to dictionary tools, other resources for lexicographers and other users, e.g. VerbAtlas and SyntagNet, have been developed and made available. The third aspect crucial to ensuring interoperability is a standard for encoding dictionaries - TEI Lex-0, to which the ELEXIS project is making a considerable contribution, has received an important acknowledgement from the community.

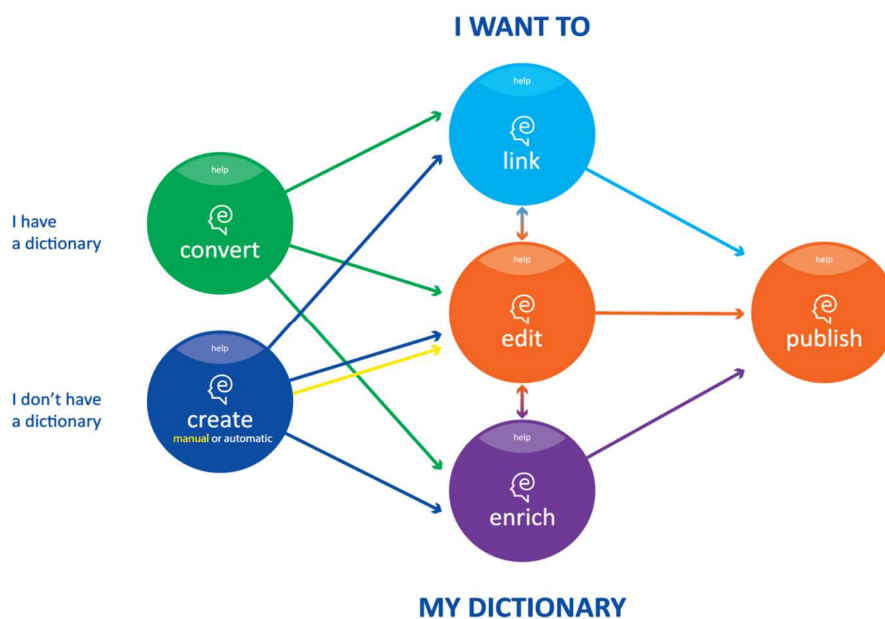


Figure 1: Graphic Guide to ELEXIS Dictionary Tools.

Another important aspect of ELEXIS is providing necessary information to lexicographic community about the research literature and related developments. This is offered via Elexifinder, and relatedly LexBib bibliographic library, and ELEXIS News Feed, with interoperability of these services being improved constantly.

D6.3 Intermediate interoperability report.

2 Tools and services

2.1 Sketch Engine

The Sketch Engine tool has continued to gain institutional and individual users, and was also facilitated by supporting institutional Single-Sign-On, which can be also used for using other ELEXIS tools such as Lexonomy and Elexifier. Institutions from 31 different countries access the Sketch Engine through ELEXIS. Table 1 and Figure 2 present the number of institutions and percentage of all institutions per country respectively, with United Kingdom, Germany, France, Spain and Italy leading the way.

Austria	9
Belgium	9
Bulgaria	2
Canada	1
Croatia	28
Czech Republic	13
Denmark	8
Estonia	3
Finland	11
France	42
Germany	58
Greece	9
Hungary	6
Ireland	11
Italy	37
Latvia	2
Lithuania	5
Luxembourg	1
Malta	1
Netherlands	8
Norway *)	1



D6.3 Intermediate interoperability report.

Poland	10
Portugal	9
Romania	2
Serbia *)	1
Slovakia	3
Slovenia	8
Spain	40
Sweden	12
United Kingdom *)	69
United States *)	1
Grand Total	420

Table 1: Institutions per country.

*) access for the UK extended following the Brexit agreement, other non-EU countries only include institutions which gained the ELEXIS observer status

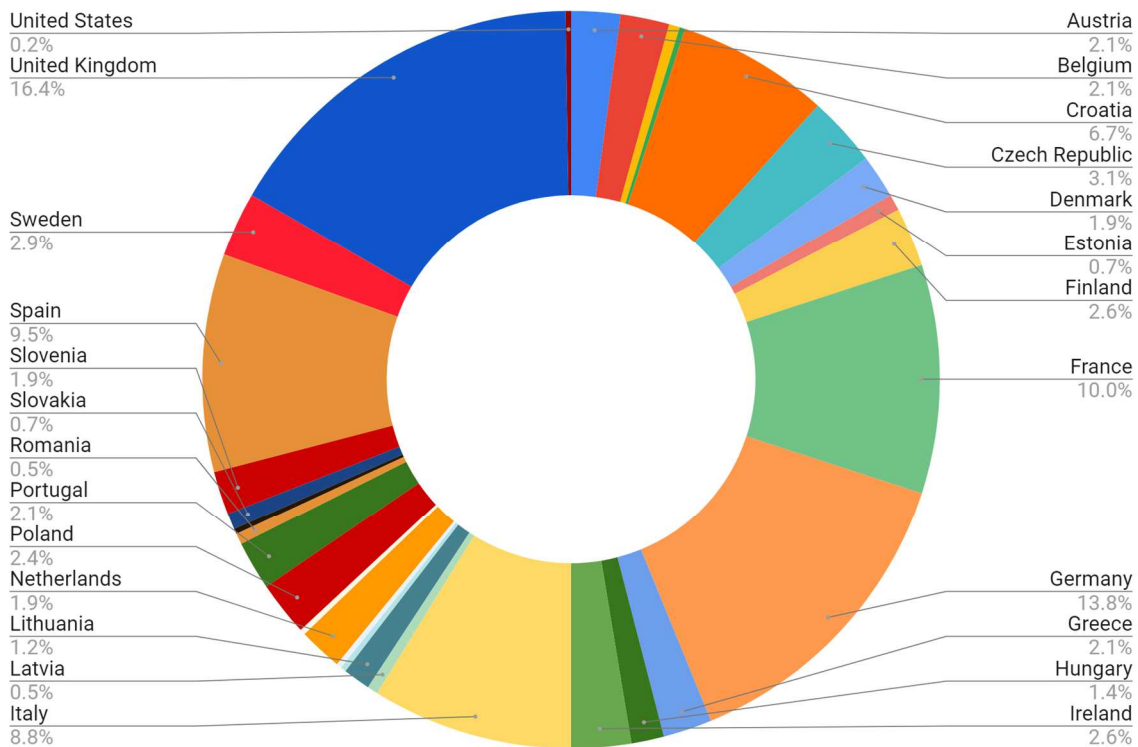


Figure 2: Institutions per country (in percentages).



2.2 One-click dictionary

OneClick Dictionary (OCD) is a dictionary drafting module. It interconnects a corpus management system (in our case SketchEngine) with our dictionary writing and online dictionary publishing system Lexonomy and provides an automatically created dictionary draft (e.g. headwords, wordforms, collocations, examples), to be post-edited in Lexonomy by the lexicographer. OneClick Dictionary enables lexicographers to shift all lexicographers work and intellectual input into the post-editing phase instead of manually analyzing the input data before creating a dictionary draft.

The users have created 798 dictionaries using One-click dictionary module, with most dictionaries being specialised in nature focussed on terms from sports, medicine, computational sciences and other disciplines. 69 of the dictionaries contained more than 50 manual edits, indicating the contents were post-edited and curated. Selected public examples:

- Climbing - <https://www.lexonomy.eu/86bythpd>. A dictionary containing climbing terminology in English, with entries including sense division, collocations, examples and synonyms.
- Neural Networks - <https://www.lexonomy.eu/8fknkq99>. A dictionary containing English terms related to the field of neural networks.
- Koronavirus - <https://www.lexonomy.eu/8i5djr4>. A dictionary containing Czech terms on the topic of coronavirus.
- Zodpovědnost za své zdraví - <https://www.lexonomy.eu/gg3f5qq7>. A dictionary containing medical terms in Czech.
- Ploutvové plavání - <https://www.lexonomy.eu/i4webhqy/>. A dictionary of terms related to finswimming.
- RICETTE - <https://www.lexonomy.eu/sbr3kbws>. An Italian dictionary of words and combinations used in recipes.
- Vazna hudba - <https://www.lexonomy.eu/rv6ajjuw>. A dictionary of Czech terms related to classical music.
- IB047 - Machine Learning terminology - <https://www.lexonomy.eu/uv2auigh>. A dictionary of English terms related to machine learning.



D6.3 Intermediate interoperability report.

- Business - <https://www.lexonomy.eu/wuhm2fp7>. A dictionary of English business terms. This dictionary also includes automatically extracted definitions or descriptions.

2.3 Lexonomy

The Lexonomy dictionary writing tool has been further developed, among the most important improvements, both in terms of interoperability and user experience, were:

- support for manual linking, both in the backend database and in the user interface, which will be crucial for linking activities and interoperability with NAISC and Babelnet Linker;
- support for TEI Lex0 import, and export of dictionary data in Ontolex format (still experimental)
- integration with new Sketch Engine user interface (login via SkE, corpora queries...)
- rewriting the source code of backend to Python which resulted in faster dictionary search and better extensibility of Lexonomy;
- a lot of small bugfixes and new features for better usability, faster editing and searching, more user-friendly interface.

There were 158 commit in the GitHub repository in total, the activity is shown in Figure 3.

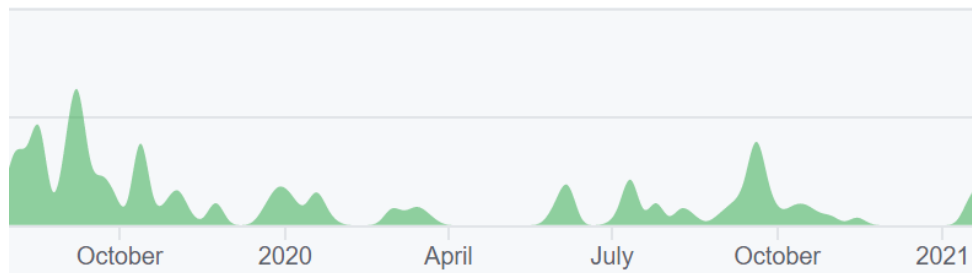


Figure 3: Commits on Lexonomy GitHub (excluding merges).

At the time of writing this report, Lexonomy has had a total of 2,760 users who created 5,416 dictionaries, 798 of them using the One-Click dictionary module. The dictionaries contain 34,889,601 headwords in total.

We have obtained 75 lexical resources from ELEXIS partners and observers, 56 of them (coming from 25 different institutions) have already been uploaded to Lexonomy. The lexical resources range from



D6.3 Intermediate interoperability report.

different types of dictionaries, e.g. large general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), to lemma lists. We list them in Table 2, along with the information on the copyright status and number of entries.

Lexical resource	Institution	Licence	Number of entries
Dictionary RAE 22	Real Academia Espanyola	restricted	88,455
Schranka 1905 - Wiener Dialekt-Lexikon	The Austrian Academy of Sciences	CC BY-NC 3.0	1,334
Dictionary of Bavarian Dialects of Austria	The Austrian Academy of Sciences	unknown	9,139
Jakob 1929 - Wörterbuch des Wiener Dialektes	The Austrian Academy of Sciences	open access	8,479
DEU LORITZA - Neues Idioticon Viennense	The Austrian Academy of Sciences	CC BY-NC 3.0	4,171
A machine-readable Persian-English Dictionary	The Austrian Academy of Sciences	CC BY-NC-SA 3.0	10,628
A machine-readable Dictionary of Dagaare	The Austrian Academy of Sciences	CC BY-NC-SA 3.0	1,252
A digital dictionary of Damascus Arabic	The Austrian Academy of Sciences	CC BY-NC-SA 3.0	2,600
A Digital Dictionary of Tunis Arabic	The Austrian Academy of Sciences	CC BY-NC-SA 3.0	7,661
Orthography Dictionary 2001	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY-NC-ND 4.0	92,617
Dictionary of Contemporary Dutch	The Instituut voor de Nederlandse Taal	restricted	57,953
Bulgarian Explanatory Dictionary	Institute for Bulgarian Language Prof Lyubomir	academic	59,622



D6.3 Intermediate interoperability report.

	Andreychin		
Bulgarian Dictionary of Synonyms	Institute for Bulgarian Language Prof Lyubomir Andreychin	academic	29,998
Bulgarian Dictionary of New Words	Institute for Bulgarian Language Prof Lyubomir Andreychin	unknown	2,658
Bulgarian Dictionary of Antonyms	Institute for Bulgarian Language Prof Lyubomir Andreychin	unknown	10,390
The Dictionary of Standard Estonian 2013	Institute of the Estonian Language	academic	425,766
Kalkars Dictionary - Danish from 1300 to 1700	The Society for Danish Language and Literature	unknown	76,430
Dictionary of the Danish Language - ODS lemmas	The Society for Danish Language and Literature	restricted	163,012
Moth's Dictionary	The Society for Danish Language and Literature	restricted	93,832
The Danish Dictionary - DDO lemmas	The Society for Danish Language and Literature	restricted	99,286
Czech lemma lists	Institute of the Czech National Corpus	CC BY-SA 4.0	169,934
Monier-Williams Sanskrit-English Dictionary	Cologne Center for Humanities	CC BY 3.0	398,412
Svenska Akademiens Ordlista	Swedish Academy	open access license, for non-commercial use	984,823
Swedish Academy Dictionary	Swedish Academy	open access license, for non-commercial use	550,424
The lemma list of the	Leibniz Institute for the	open access	275,756



D6.3 Intermediate interoperability report.

German dictionary "elexiko"	German Language		
the lemma list from the 2001 Dicionario do Lingua Portuguesa Contemporanea	Lisbon Academy of Sciences	unknown	69,360
CJVT Thesaurus 1.0	Centre for Language Resources and Technologies, University of Ljubljana	CC BY-SA 4.0	105,473
Tezaurs Latvian	Institute of Mathematics and Computer Science, University of Latvia	CC BY-SA 4.0	320,869
Sarasola 1996 Basque monolingual Dictionary	UPV/EHU University of the Basque Country	unknown	
"Hiztegi Batua" Basque Dictionary	UPV/EHU University of the Basque Country	CC BY-SA	
Iberian integrated Wordnets' Multilingual Central Repository "MCR 3.0"	UPV/EHU University of the Basque Country	CC BY	30,263
Dizionar F LAD-DEU	Institute for Applied Linguistics, Eurac Research	restricted	872
Trilingual Legal Terminology Bistro	Institute for Applied Linguistics, Eurac Research	restricted	11,323
Schweizerisehes Idiotikon	Schweizerisehes Idiotikon	CC BY-SA	160,254
Icelandic lemma list plus related translations: Danish, Swedish and Norwegian (its two standards)	The Árni Magnússon Institute for Icelandic Studies	CC BY-NC-ND	48,480
Dictionary of Slovenian Phrasemes	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC-BY-NC 4.0	3,002



D6.3 Intermediate interoperability report.

Dictionary of Slovenian Particles	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	429
Dictionary of Newer Words	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	5,382
PleTERSNIK Dictionary	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	103,185
OdZADNJI slovar	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	114,341
Nova beseda frequency lexicon	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	2,251,151
JSV	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	8,461
Dictionary of lesser used Slovenian words	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	178,457
Systematic Dictionary of the Lithuanian Language	The Institute of the Lithuanian Language	unknown	1,614
Comparative Dictionary	The Institute of the Lithuanian Language	unknown	7,093
Dictionary of Modern Lithuanian	The Institute of the Lithuanian Language	unknown	48,075
Nords Ordbank - Nynorsk	University of Bergen Library	CC-BY	121,112
Nords Ordbank - Bokmal	University of Bergen Library	CC-BY	153,939
Bulgarian Valency Lexicon	Institute of Information and Communication Technologies at the Bulgarian Academy of	CC BY 3.0	9,574



D6.3 Intermediate interoperability report.

	Sciences (IICT-BAS)		
Historical Finnish Dictionary	Institute for the Languages of Finland	CC BY 4.0	42,341
Finnish dialect dictionary	Institute for the Languages of Finland	CC BY 4.0	161,148
Karelian dictionary	Institute for the Languages of Finland	CC BY 4.0	88,575
TermiMining	Univesity of Belgrade, Faculty of Mining and Geology	CC-BY-NC-SA	47
GeologyTerm	Univesity of Belgrade, Faculty of Mining and Geology	CC-BY-NC-SA	312
ESMIND	Univesity of Belgrade, Faculty of Mining and Geology	CC-BY-NC-SA	
Latin Vallex	University Cattolica del Sacro Cuore - CIRCSE Research Centre	CC-BY-NC-SA 3.0 Unported	1,379

Table 2: ELEXIS partner and observer lexical resources in Lexonomy.

2.4 Elexifier

Elexifier is an app that allows you to transform XML and PDF dictionaries into an Elexis Data Model compliant format. It was first launched in spring 2020. In total, 98 users have registered in the app (either through Elexifier registration process or via the Sketch Engine Single Sign-On login facility) and they have uploaded a total of 177 dictionaries, or 3.67 per user.

Of the 177 dictionaries, 69 were in the XML format where users must define a transformation into the Elexis Data Model. The total number of transformations is 89, or 1.43 per XML dictionary. The remaining 108 uploaded dictionaries were in the PDF format where users have to annotated a small portion of the PDF text to provide training data for a machine learning algorithm which is then used to annotate the entire dictionary.



D6.3 Intermediate interoperability report.

2.4.1 Case study 1 (XML): Converting and aligning Wordnet lexical databases between languages

Elexifier's XML module is used in the process of converting Wordnet lexical databases into a standardized TEI format which can then be used to align word senses between different languages.

2.4.2 Case study 2 (PDF): Converting an 18th century Spanish-Basque-Latin Trilingual Dictionary

Elexifier was used in the process of converting an 18th century Spanish-Basque-Latin Trilingual Dictionary from scanned images into a TEI-compliant XML format. The optical character recognition was performed outside of Elexifier (<http://kraken.re>) and a customization was developed to support the OCR format (ALTO-XML). The annotation has been carried out for twenty columns (ten dictionary pages), and then used as training set for the segmentation algorithm, which structures the content of the whole dictionary according to what it has been given as training set. A first evaluation of the information extraction results suggests that Spanish headwords and Basque translation equivalents have been recognized by the software with high precision. Latin equivalents, in turn, the third category we have looked at, has been recognized with much lower precision. Headwords seem to be recognized seamlessly, which should be due to the fact that headwords are positioned in the entry layout in a first negatively indented line, and followed by a comma. This has been the case in all annotated entries, and thus is a very straightforward criterion for the ML algorithm. On the other hand, also items that do not describe headwords are placed in a negatively indented line, and subsequently, have been identified as headwords. Latin equivalents will appear for headwords, but also for translating examples; here we have a contradictory evidence that makes the algorithm unable to predict the correct annotation for Latin items in many cases. Basque equivalents are not straightforwardly identifiable, since their layout feature (italics font style) is also present in examples, as well as in Spanish to Spanish cross-references.

This example clearly shows Elexifier's potential as well as some of the pitfalls it may encounter. The machine learning algorithm used for PDF conversion depends on high quality training data which is somewhat difficult to obtain with scanned images of 18th century texts. But despite the difficulties, Elexifier produced good results for some elements and provided a useful alternative to manual extraction of information from scanned texts.



D6.3 Intermediate interoperability report.

2.5 VerbAtlas

VerbAtlas (<http://verbatlas.org/>) is a novel large-scale manually-crafted semantic resource for wide-coverage, intelligible & scalable Semantic Role Labeling. The goal of VerbAtlas is to manually cluster WordNet synsets that share similar semantics into sets of semantically-coherent frames. The main features are:

- 466 semantically-coherent frames using 26 cross-frame VerbNet-inspired semantic roles for their argument structure.
- Available both for download and via RESTful API.
- Full coverage of WordNet 3.0 verb synsets (13,000+).
- Complete linkage to BabelNet 4.0, which supports 280+ languages (new version to come later this year!).
- Manual mapping to PropBank of all CoNLL-2009 and CoNLL-2012 dataset occurrences (5000+ mappings).
- Selectional preferences: the superconcept most probably associated with a semantic role in a frame (e.g. food for the patient role of the EAT frame).
- Default/shadow arguments: arguments logically implied or already incorporated into a verb.
- Implicit arguments: arguments that are implicit in the argument structure of a verb.

As the user statistics in Figure 4 shows, the resource has been visited 62,422 times by 2,068 different users who have conducted 5,143 sessions (2.49 sessions per user on average). The most often identified languages of users were English (American and British), Italian, Chinese, Spanish, French, and German. 79 % of the users were new visitors, whereas 21 % were returning visitors, i.e. they visited the resource more than once.



D6.3 Intermediate interoperability report.

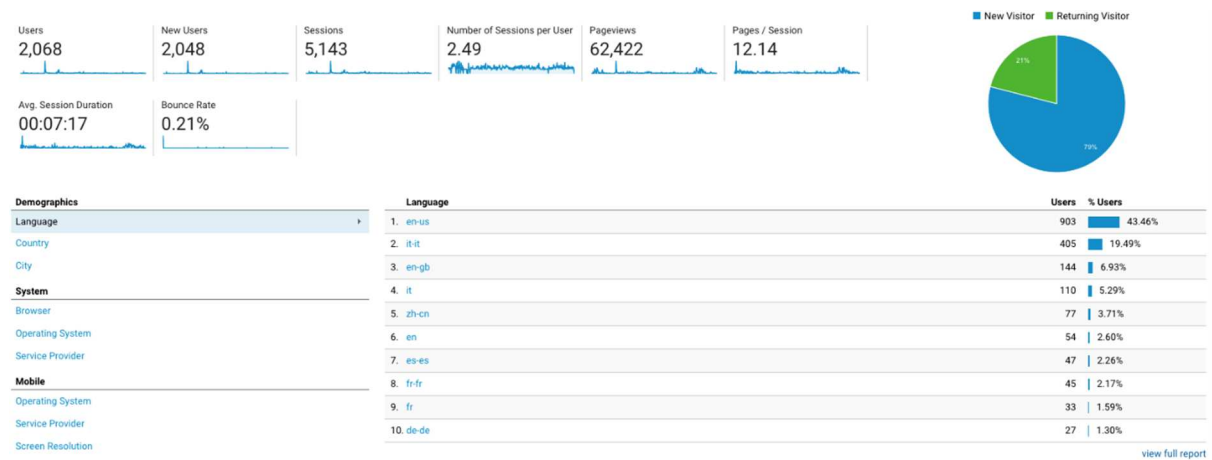


Figure 4: VerbatAtlas user statistics.

2.6 SyntagNet

SyntagNet (<http://syntag.net/>) is a manually-curated large-scale lexical-semantic combination database which associates pairs of concepts with pairs of co-occurring words. The goal of SyntagNet is to capture sense distinctions evoked by syntagmatic relations, hence providing information which complements the essentially paradigmatic knowledge shared by currently available Lexical Knowledge Bases such as WordNet. Its main features are:

- Wide coverage, with 78,000 noun-verb and noun-noun lexical combinations extracted from the English Wikipedia and the British National Corpus.
- High-quality, fully manual disambiguation for all of the lexical combinations, according to the WordNet 3.0 sense inventory.
- A resulting Lexical Knowledge Base made up of 88,019 semantic combinations linking 20,626 WordNet 3.0 unique synsets with a relation edge.
- A user-friendly web interface for looking up terms and their lexical-semantic combinations, with complete linkage to BabelNet 4.0.

The user statistics in Figure 5 shows that SyntagNet has been visited 18,100 times by 1,115 different users who have conducted 2,036 sessions (1.83 sessions per user on average). The most often identified languages of users were English (American and British), Italian, Chinese, French, Spanish, and German. 80 % of the users were new visitors, whereas 20 % were returning visitors, i.e. they visited the resource more than once.



D6.3 Intermediate interoperability report.

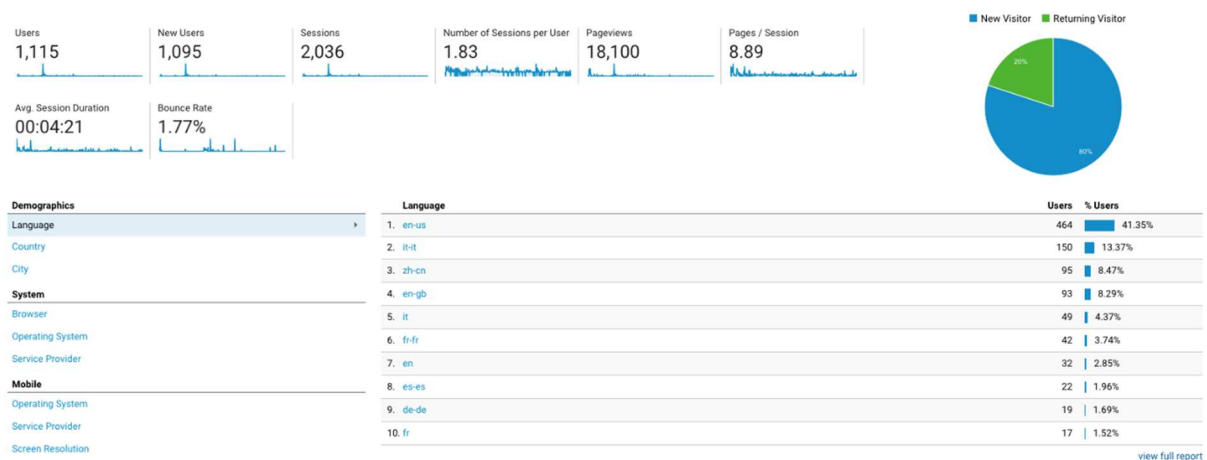


Figure 5: SyntagNet user statistics.

2.7 NAISC

The Naisc tool is available as open source on GitHub at <https://github.com/insight-centre/naisc>. It has also been practically applied in a number of case studies in collaborative projects between NUIG and commercial partners. The tool was used in the context of an engagement between NUIG and Fidelity Investments, where it was used to link terminologies derived in the project (see this paper¹) with openly available financial ontologies. A second case study was conducted with Translators without Borders (TWB), a non-profit organization working to develop language resources to assist aid workers in crisis situations. In this project, the tool was used to link the lexical resources developed by TWB with an open lexical resource, English WordNet, and a semantic resource, Wikidata. This linking was then used to enrich the resources further with extra translations, related terms and images. The focus of this project was on extending the resources for highly under-resourced languages, namely, Bengali, Burmese, Chittagonian and Rohingya. Finally, the Naisc tool has been used in collaboration with Oxford University Press in the Prêt-à-LLOD project, where we are comparing its effectiveness with in-house sense linking tools developed by OUP.

¹ Taxonomy Extraction for Customer Service Knowledge Base Construction. Bianca Pereira, Cécile Robin, Tobias Daudert, John P. McCrae, Paul Buitelaar and Pranab Mohanty, Proceedings of the SEMANTiCS 2019, (2019)



D6.3 Intermediate interoperability report.

2.8 Game of words

Game of words, a mobile app for iOS and Android devices, which was designed to be used for crowdsourcing purposes in lexicographic context, has been initially made available in four languages, Slovenian, English, Estonian and Dutch. For Android devices, we only have cumulative statistics available which show that there have been 318 installations in total, with the users being located in Slovenia, Italy, Estonia, Austria, Serbia, the Netherlands, Portugal, Argentina, Spain, India and other countries.

For iOS devices, a more detailed statistics is available. The Slovenian version of the app has had 974 visits, 137 installations and 457 sessions in total. The users were located in Slovenia, Italy, United States, Germany and China. The English version of the app has had 231 visits, 30 installations and 226 sessions in total. The users were located in from the United States, Spain, Saudi Arabi and Vietnam. The Estonian version of the app has had 357 visits, 76 installations and 236 sessions in total. The users were located in Estonia and Finland.

2.9 TEI Lex-0

TEI Lex-0 is a technical specification and a set of community-based recommendations for encoding machine-readable dictionaries. It establishes a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers.

The activities of the TEI Lex-0 team within ELEXIS take part in collaboration with the DARIAH Working Group “Lexical Resources”. The work is conducted and recorded through a GitHub [repository](#). The Guidelines are available for [online consultation](#). TEI Lex-0 activities and outputs have already been seminal in achieving more consistency in the adoption and implementation of the TEI Guidelines; propelling the scholarly debate around modeling lexicographic data; championing the importance of open standards to a new generation of scholars; and establishing TEI Lex-0 as an internationally recognized data interchange format.



D6.3 Intermediate interoperability report.

The TEI Lex-0 project was awarded the [2020 Rahtz Prize for TEI Ingenuity](#), which is awarded to an individual or team judged to have made a significant contribution to the TEI Consortium’s mission of “developing and maintaining a set of high-quality guidelines for the encoding of humanities texts”.

In addition to creating customized TEI Lex-0 Guidelines, the team has contributed to the development of TEI itself by submitting and getting accepted a number of tickets proposing changes to the TEI Guidelines. See for instance tickets [1791](#), [1809](#), [1512](#), [1702](#), [1510](#), [1734](#), [1819](#) etc.

The team has published a number of papers and presented at a number of conferences. For a full bibliography, see https://www.zotero.org/groups/2711819/tei_lex-0/items/6WKACI9B/library.

During the ELEXIS grant period, TEI Lex-0 and best practices in lexical data modeling using TEI have been introduced to more than 90 young scholars from across Europe at a number of training events including:

- [Lexical Data Masterclass 2018](#) (Berlin, 3-7 Decmeber 2018). For an overview of student projects, check out [From Àbèsàbèsì to XPath](#) on DigiLex.
- [From Print to Screen: The Theory and Practice of Digitizing Dictionaries](#). Lisbon Summer School in Linguistics (2-6 July 2018).
- [Encoding Dictionaries with TEI: A Masterclass](#). Lisbon Summer School in Linguistics (1-5 July 2019).
- [DH Training Workshop: Digital Methods for Linguistic Investigation](#) (Berlin, 13-15 November 2019).

In addition, TEI Lex-0 will be used as the native encoding format in two recently funded projects:

- [Electronic lexical database of Indo-Iranian languages](#) funded by The Technology Agency of the Czech Republic.
- [MORDigital – The Digitization of the Dicionario da Lingua Portuguesa by António de Morais Silva](#) funded by the Portuguese Foundation for Science and Technology.



D6.3 Intermediate interoperability report.

2.10 Elexifinder

The search tool Elexifinder (<http://finder.elex.is/>) is dedicated to helping lexicographers and other researchers find scientific output in lexicography and related fields. The users can search by keywords, concepts (words or set of words with a Wikipedia page), authors, type of publication (paper, video), date and other conditions. The contents of Elexifinder have been significantly increased since the last report: 4,584 publications (mainly conference papers, journal papers and book chapters) have been added to existing 1,755 publications, making 6,339 publications in total. The number of different languages in which the publications are written has increased from 11 to 20. Leading authors of publications come from 76 different countries, top 25 countries are listed in Figure 6.

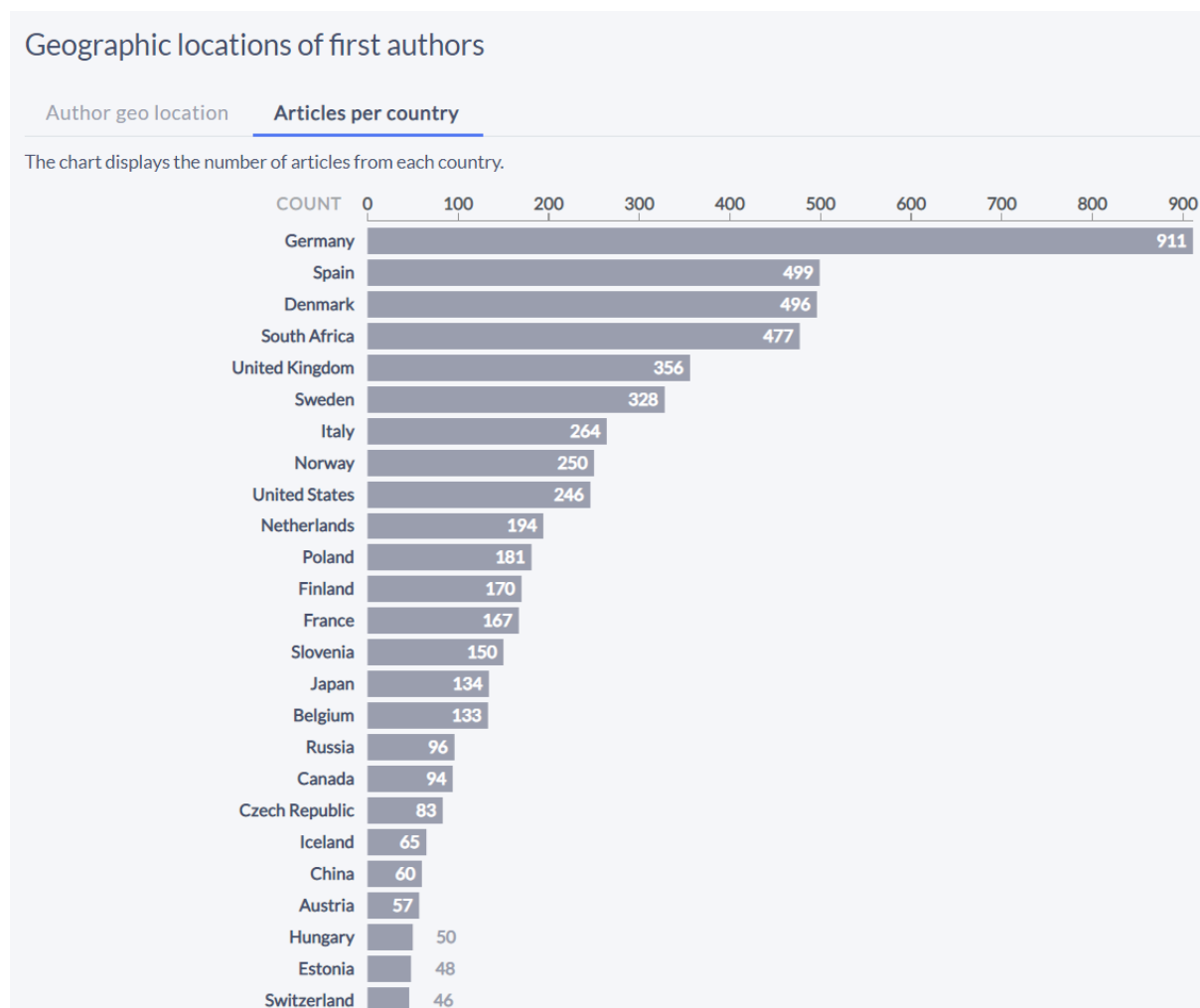


Figure 6: First author locations by country (Elexifinder).



D6.3 Intermediate interoperability report.

Every publication in Elexifinder is linked to the document (PDF or abstract), source (journal issue or conference page) and its bibliographic entry with additional information (e.g. pages, ISSN) in the LexBib Zotero collection (<https://www.zotero.org/groups/1892855/lexbib>). The Zotero library enables the users, among other things, to export the bibliographic information on the publications in a wide number of different citations formats.

One of important updates to Elexifinder data is author name unification. Certain authors use(d) different versions of their names, e.g. sometimes with a middle name, other times without, which means that when searching for a certain author, the users needed to select all the variants found in the database. We have addressed this problem by selecting one name variant as the main one that links together all the variants. For example, B. T. Sue Atkins (now the main name) has variants Beryl T. Sue Atkins, Sue Atkins, B. T. S. Atkins, and Beryl T. Atkins in the database. This solution not only facilitates searching in Elexifinder but also enables easier interoperability, e.g. we intend to provide Elexifinder links leading to all publications of registered members on soon to be launched LexMeet (meet.elex.is) platform.

2.10.1 Usage statistics

Since the last report, the Elexifinder service was used by 395 different users, who made 6,850 queries in total or just over 17 requests per user. Most frequently searched concepts and keywords are shown in Tables 3 and 4 respectively.

Concept	Number of queries
Circular definition	1806
Collocation	106
Semantics	74
Dictionary	60
Polysemy	44
WordNet	24
Application programming interface	22
Multimedia	21



D6.3 Intermediate interoperability report.

Lexical semantics	19
Usability	14
Software	13
Beryl Atkins	12
Cognitive Synonymy	11
Adjective	11
Lexicography	11
Patrick Hanks	11
User research	10
Robot	10
Part-of-speech tagging	8
Terminology extraction	7

Table 3: Most frequently searched concepts in Elexifinder (July 2019 - January 2021)

While most searches are still done in English, the analysis of keyword-based searches reveals that the users are also conducting searches in other languages (we have identified Slovenian, Croatian, Russian, German and Spanish, to name just a few), which is in line with the Elexifinder functionalities that support cross-language searches.

Keyword	Number of queries
REST interface	260
collocation	126
valency	14
semantics	13
paella	11
language planning	10
krek	10



D6.3 Intermediate interoperability report.

abusive	10
asialex	10
patterns	9
Terminological dictionary of the pharmaceutical sciences	8
collocations	6
paronym	6
eye-tracking	6
GDEX	6
user	6
lexonomy	6
slovar	6
academic writing	6
pandemia	6

Table 4: Most frequently searched keywords in Elexifinder (July 2019 - January 2021)

2.11 ELEXIS Lexicographic Newsfeed

Lexicographic newsfeed is an ELEXIS service that uses the Event Registry API to extract latest news articles identified to be related to lexicography. News articles are extracted from 30,000 news sources, and over 35 languages are currently supported. The Newsfeed has become a very popular service, as is evident from the number of visits on the ELEXIS website in Table 5 (a slightly more detailed breakdown of the page visits is provided in Table 6).²

² Please note that the actual number is probably higher as the most current news is also displayed on the front page.



D6.3 Intermediate interoperability report.

Which pages do elex.is users visit (top 10)?
(Aug. '19 - incl. Jan. '21)

page	visits
/	11.277
/tools-and-services/	2.715
/grants-for-research-visits/	1.812
/observers/	1.732
/join-as-observer/	1.477
/objectives/	1.040
/tools-and-services/lexicographic-news/	1.016
/all-events/	957
/partners/	785
/travelgrants/	667

Table 5: Top 10 ranking of most popular elex.is pages.

page	page views	unique page views	avg. time spent
https://elex.is/tools-and-services/lexicographic-news/	1,016	922	00:04:41

Table 6: lexicographic Newsfeed - page views over time period (Aug 2019 - incl. Jan 2021).

The breakdown of user source, i.e. where the users come from (Table 7), shows that they mostly come via Google or directly (presumably via bookmarked link), but a significant number of users also come to the newsfeed via Twitter and Facebook. This reflects our constant efforts of promoting selected Newsfeed news on social media, which is shown in Figure 7.



D6.3 Intermediate interoperability report.

	page	source/medium	page views	unique page views	avg. time spent	access
			1.016 % des Gesamtwerts: 2,41 % (42.103)	922 % des Gesamtwerts: 2,77 % (33.341)	00:04:41 Durchn. für Datenansicht: 00:01:56 (142,12 %)	264 % des Gesamtwerts: 1,36 % (19.457)
1.	/tools-and-services/lexicographic-news/	google / organic	525 (51,67 %)	497 (53,90 %)	00:03:52	107 (40,53 %)
2.	/tools-and-services/lexicographic-news/	(direct) / (none)	202 (19,88 %)	165 (17,90 %)	00:02:47	71 (26,89 %)
3.	/tools-and-services/lexicographic-news/	t.co / referral	138 (13,58 %)	125 (13,56 %)	00:09:09	61 (23,11 %)
4.	/tools-and-services/lexicographic-news/	facebook.com / referral	65 (6,40 %)	59 (6,40 %)	00:05:55	4 (1,52 %)
5.	/tools-and-services/lexicographic-news/	l.facebook.com / referral	20 (1,97 %)	20 (2,17 %)	00:09:52	3 (1,14 %)
6.	/tools-and-services/lexicographic-news/	m.facebook.com / referral	18 (1,77 %)	13 (1,41 %)	00:00:44	12 (4,55 %)
7.	/tools-and-services/lexicographic-news/	lexonomy.eu / referral	5 (0,49 %)	4 (0,43 %)	00:05:26	0 (0,00 %)
8.	/tools-and-services/lexicographic-news/	mailchi.mp / referral	5 (0,49 %)	5 (0,54 %)	00:03:00	2 (0,76 %)
9.	/tools-and-services/lexicographic-news/	sketchengine.eu / referral	5 (0,49 %)	4 (0,43 %)	00:00:23	1 (0,38 %)
10.	/tools-and-services/lexicographic-news/	github.com / referral	4 (0,39 %)	3 (0,33 %)	00:01:41	1 (0,38 %)

Table 7: Where Newsfeed users come from: source/medium.

ELEXIS lexicographic Newsfeed

Number of postings per Social Media platform (August 2019 - incl. January 2021)

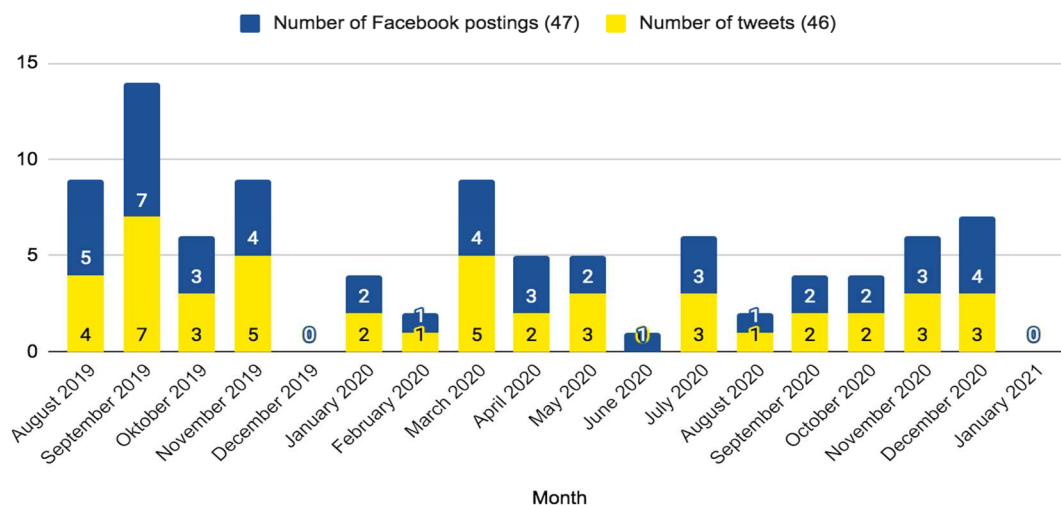


Figure 7: Newsfeed postings per Social Media platform over time period



D6.3 Intermediate interoperability report.

Finally, the analysis of users' country of origin shows that majority of lexicographic newsfeed users come from Europe, with Slovenian users leading the way, followed by the users from Austria, France, Spain and Czechia.











country	page views	unique page views	avg.time spent	access
	1.016 % des Gesamtwerts: 2,41 % (42.103)	922 % des Gesamtwerts: 2,77 % (33.341)	00:04:41 Durchn. für Datenansicht: 00:01:56 (142,12 %)	264 % des Gesamtwerts: 1,36 % (19.457)
1.  Slovenia	431 (42,42 %)	408 (44,25 %)	00:03:38	40 (15,15 %)
2.  Austria	209 (20,57 %)	193 (20,93 %)	00:08:15	26 (9,85 %)
3.  France	66 (6,50 %)	58 (6,29 %)	00:08:26	52 (19,70 %)
4.  Spain	49 (4,82 %)	47 (5,10 %)	00:06:37	38 (14,39 %)
5.  Czechia	41 (4,04 %)	16 (1,74 %)	00:03:04	7 (2,65 %)
6.  Croatia	22 (2,17 %)	17 (1,84 %)	00:01:05	7 (2,65 %)
7.  Portugal	19 (1,87 %)	15 (1,63 %)	00:00:10	7 (2,65 %)
8.  United Kingdom	16 (1,57 %)	16 (1,74 %)	00:00:00	14 (5,30 %)
9.  Germany	15 (1,48 %)	15 (1,63 %)	00:06:26	7 (2,65 %)
10.  Italy	14 (1,38 %)	12 (1,30 %)	00:01:07	8 (3,03 %)

Table 8: Newsfeed users per country (top 10) over time period (Aug 2019 - incl. Jan 2021).

