EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

New developments in Elexifinder, a discovery portal for lexicographic literature

Kosem I., Lindemann D.

Jožef Stefan Institute, Slovenia
iztok.kosem@cjvt.si, david.lindemann.soraluze@gmail.com

Abstract

In this paper, we present ongoing work on Elexifinder (https://finder.elex.is), a lexicographic literature discovery portal developed in the framework of the ELEXIS (European Lexicographic Infrastructure) project. Since the first launch of the tool, the database behind Elexifinder has been enriched with publication metadata and full texts stemming from the LexBib project, and from other sources. We describe data curation and migration workflows, including the development of an RDF database, and the interaction between the database and Elexifinder. Several new features that have been added to the Elexifinder interface in version 2 are presented, such as a new Lexicography-focused category system for classifying article subjects called LexVoc, enhanced search options, and links to LexBib Zotero collection. Future tasks include getting lexicographic community more involved in the improvement of Elexifinder, e.g. in translation of LexVoc vocabulary, improving LexVoc classification, and suggesting new publications for inclusion.

Keywords: ELEXIS; bibliographical data; metalexicography; lexicographic research

1 Introduction

In 2019, a service called Elexifinder,[1] which enables search for lexicographic research, was made available to the lexicographic community. The tool was developed as part of the European Lexicographic Infrastructure (ELEXIS),[2] a H2020 project funded by the European Commission. Elexifinder has been built using some of the elements of the Event Registry system architecture (Leban et al. 2014; Leban, Fortuna & Grobelnik 2017). The first version included 1,755 publications and 78 videos in 11 different languages, with the majority of publications coming from EURALEX and eLex proceedings. A detailed presentation of this version, including the features available at the time, was made by Kosem and Krek (2019).

At the time of Elexifinder launch, there were already plans in motion to improve it even further. The planned improvements were related to the contents, interface and data preparation workflow. For example, an extensive list of publications to be added to Elexifinder was already compiled at the time, and has been regularly updated since. The way article full texts and publication metadata were collected and recorded was far from optimal, and did not utilize existing sources well enough.

In parallel, at University of Hildesheim, the LexBib project was planned, with the goal of creating a domain ontology and digital bibliography of Lexicography and Dictionary Research. First steps in that project are described in Lindemann, Kliche & Heid (2018); the LexBib metadata and full text collection was put together and made accessible using the Zotero platform,[3] and workflows for a conversion of publication metadata to RDF Linked Data were explored.

Despite slightly different aims (ELEXIS being focussed more on providing a tool to efficiently find the relevant publication, and the LexBib Zotero collection having a more bibliographic focus, that is, to provide validated metadata for the purpose of citations), both initiatives had a great deal of overlap. Therefore, it made perfect sense to merge the efforts and develop a workflow that would benefit both purposes and do away with unnecessary duplication of effort, especially in terms of obtaining publications and recording their metadata.

This paper presents various improvements, some considerable, that have been made to Elexifinder since 2019. We start by looking at the backend, presenting the development of an RDF database, and the interaction between the database and eLexifinder. Next, the new LexVoc SKOS vocabulary of subject headings is presented, including the linking of content-describing terms to publication metadata. Then, we present the new contents added, and the improvements made to the Elexifinder interface. The paper concludes by outlining future plans, such as the translation of LexVoc vocabulary, and the inclusion of more publications.

2 Bibliographical data

2.1 Sources and workflows

[1] Accessible at https://finder.elex.is .
[2] The project homepage is accessible at https://elex.is.
[3] Accessible at https://lexbib.org/zotero.

For the first version of Elexifinder, publication metadata were stored in spreadsheet files, along with manually revised plain text versions of the corresponding full texts. In addition, publication metadata was enhanced with the location of the first author, by manual annotation. We have merged that data with the data present at that time in the LexBib Zotero collection, defined priorities for the inclusion of further bibliographic items, and established a workflow for data migration from Zotero to Elexifinder (see workflow schema in Figure 1).

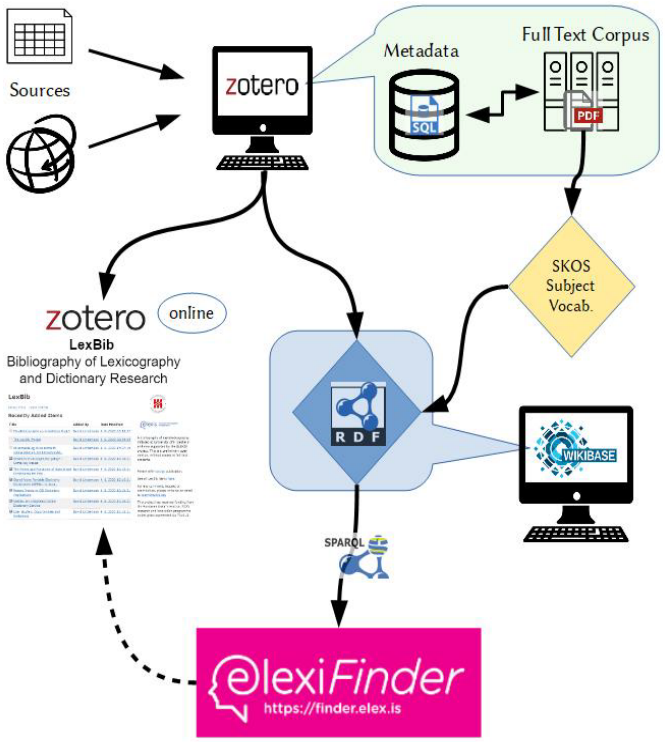

Figure 1: Workflow and data migration scheme

At present, all articles of major journals in the discipline of lexicography, and EURALEX and eLex conference series, a set of edited book volumes, and a set of presentation videos are part of our collection (see complete reference in section 4.1 below). We also have started to integrate bibliographical records stemming from the EURALEX bibliography, compiled by A. Dykstra with the support of individual community members,[4] and OBELEX-meta (Möhrs 2016).[5] Since the enhancement of publication metadata by computational processing of full texts is one of our goals, we have given priority to Open Access publications; we also recorded publications where the full text was accessible due to suitable license agreements. Our copies of full texts are not publicly accessible; they remain restricted for the described text mining purposes in the framework of the project. Through the Zotero platform, and in the Elexifinder tool, we provide download links that lead to the publisher site, which are restricted according to the applicable license.
All publication metadata records have been manually validated. In nearly all cases, the metadata sets resemble all categories necessary for citations, and in most cases, also article abstracts. The Zotero platform allows an export of single items or item batches in several citation styles, or as a structured dataset, e.g. in the bibtex format. Wherever possible, i.e. where that information has been available in the full texts themselves, or journal issue or edited volume back matters, we have manually included the location of the first author. That is now the case for around 98% of the approximately 7,000 bibliographic items included in the LexBib collection.

2.2 Author disambiguation

The most challenging task in the curation workflow of bibliographical data is, beyond any doubt, the disambiguation of author and editor name variants. In our collection, some persons appear with up to six different name variants, which is due to capitalization of first names, the inclusion of middle names, hyphenation of double first or last names, alternative spellings, last name ordering errors, spelling errors, etc. Around 18,000 author/editor statements in our data, which presented around 5,000 different names, have been reduced (disambiguated) to around 4,000 persons.
The conversion of literal author/editor values (i.e. name strings) to statements that point to disambiguated person items, where name variants of the same person come together, is the necessary step for allowing searches and search result displays

[4] Accessible at http://euralex.pbworks.com/w/page/7230036/FrontPage.
[5] Accessible at https://www.owid.de/obelex/meta.

that involve all the articles written by that person, regardless of the name variants stated in the bibliographical records, and for linking of bibliographical records to metadata concerning the authors and editors. In other words, that step is what converts metadata consisting of literal values (as in Zotero, or library records in MARC standard) to Linked Data.
A widely used standard for representing Linked Data is RDF.[6] Statements, such as e.g. the links from a bibliographic item to its creators, are represented as semantic triples. We have adapted an existing Zotero RDF export script for our use case, and migrated all publication metadata to an RDF triple store. We have clustered name variants belonging to identical persons, using the OpenRefine software application,[7] so that, for updates of the collection, the recorded name variants of a person are considered; for updates, we are using the same application (see a screenshot detail in Figure 2).

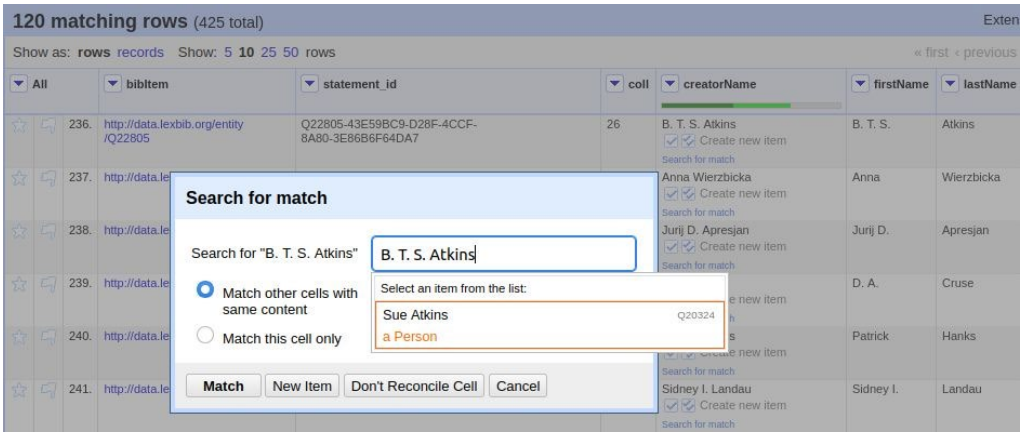

Figure 2: Auhor disambiguation in the OpenRefine application.

In addition to author disambiguation, generally speaking, the choice of a database that stores semantic triples as central data repository[8] allows us to extend the bibliographical database towards a knowledge graph that involves all relevant kinds of entities (see Lindemann, Klaes & Zumstein 2019). The knowledge graph includes an enrichment of entities with data harvested from other resources in the Linked Open Data Cloud, and/or the definition of links to entities available in these. For the time being, we concentrate on a disambiguation of natural persons, and on defining links between bibliographical records and content-describing terms (see next section). Furthermore, SPARQL database queries[9] provide a straightforward way to obtain structured datasets from RDF data in custom formats, such as the JSON export needed as ingest for the Elexifinder tool. At the same time, the SPARQL endpoint of our RDF database[10] allows the community to access the LexBib data for other research purposes.

3 Subject indexation of metalexicographical literature with LexVoc

We have started to develop LexVoc, a controlled vocabulary of Lexicography-related terms that shall be used as content descriptors, and linked to the corresponding bibliographical items. We have defined English preferred and alternative lexicalizations, and represented relations between terms according to the W3C SKOS standard,[11] that is, following a widely used practice,[12] and, at the same time, choosing a format that can be straightforwardly loaded into our RDF database. Sources for the controlled vocabulary have been the following:

1) An updated and extended version of the index of "Bibliografía Temática de la Lexicografía" (Córdoba Rodríguez 2003),[13] translated to English.
2) The typology of dictionaries by Engelberg and Storrer (2016).
3) The glossary of lexicographic terms by Kipfer (2013).
4) The index of the volume "Using Online Dictionaries" (Müller-Spitzer 2014).

We have defined relations between terms stemming from sources (1) to (4), so that terms can be represented as nodes in a graph, with SKOS relations as arcs. In the second step, we have extended the vocabulary with a manually revised subset of salient term candidates,[14] extracted from a corpus compiled using all English full texts present in the collection used for Elexifinder version 2 (see section 4.1 for full reference).
We are currently performing experiments for extending the vocabulary further, using term extraction results from subsets of our English full texts, e.g. of recent publications about electronic lexicography, in order to cover specialized state-of-

[6] See https://www.w3.org/RDF/.
[7] See https://openrefine.org/.
[8] We have employed Ontotext GraphDB and Wikibase database solutions.
[9] See https://www.w3.org/TR/rdf-sparql-query/.
[10] The SPARQL endpoint is available at https://lexbib.elex.is/query/sparql.
[11] See https://www.w3.org/2004/02/skos/.
[12] See, for example, the Library of Congress Subject Headings vocabulary, accessible at https://id.loc.gov/authorities/subjects.html.
[13] Accessible at https://www.udc.es/grupos/lexicografia/bibliografia/tematica.html.
[14] Salience is calculated according to a TF/IDF measure, in this case with EnTenTen18 as reference corpus. We have used the Sketch Engine for corpus compilation and processing (see https://www.sketchengine.eu/).

the-art terminology. The vocabulary can be explored in a constantly updated graph view,[15] and accessed using SPARQL. In addition to that vocabulary, we also use multilingual terms that denote natural languages as content-describing terms, i.e. to make the object language(s) of a research article explicit. We have obtained these terms from Wikidata.[16]
For automatic extraction of full text bodies from PDF documents, we have used the GROBID tool (Romary & Lopez 2015).[17] In subcollections where the GROBID default algorithm fails to isolate the text body from headers, footers, title, abstract, author affiliation data, and references, etc., we have resorted to manual cleaning of PDF plain text versions produced with standard 'pdf2txt' tools.[18]
LexVoc is now at a stage of development that allows first experiments of automatic indexation of articles. For this, we have performed a discovery of vocabulary terms in lemmatized full texts ("gazetteer approach"). Information on single terms, frequency data, and the bibliographic items they are associated to as content descriptors is available using SPARQL.[19] We are very interested in feedback regarding the structure of the vocabulary, and in suggestions for further sources to be included. We have set up a dedicated discussion group on the LexMeet platform.[20]
It is our goal to obtain a multilingual version of the LexVoc vocabulary in order to apply the indexation process to non-English text, on the one hand, and to provide to the users localized terms as search criteria on the other, that is, to enable users to search in their preferred language for articles indexed with certain terms, regardless of the text language. We want to cover as many languages as possible, but will be first focussing on the languages official in the countries of ELEXIS partners and observers.

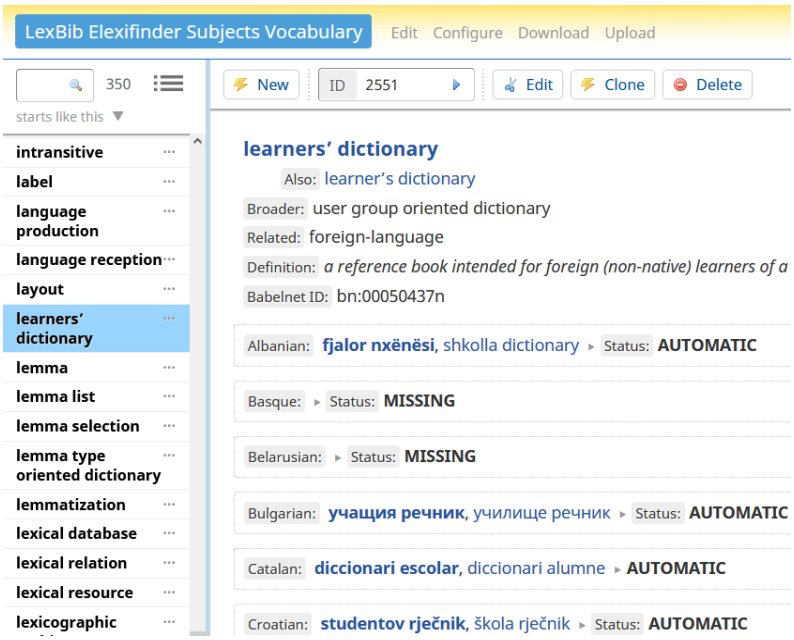

Figure 3: Lexonomy screenshot

The translation process of LexVoc will be carried out using the Lexonomy application,[21] where we have set up a set of XML-based multilingual dictionaries, one for each language, the lemmata of which are the English SKOS vocabulary terms. Contributors will access an editing form, where translation equivalents can be filled in or modified, and annotated with a status, as shown in Figure 3. During the editing process, the contents of those bilingual dictionaries will be regularly mirrored to a single multilingual resource that can be accessed by the interested public.
To facilitate the translation task, we have extracted translation candidates using the BabelNet[22] API (term labels with status "automatic" in Figure 3), so that the task for contributors consists of validating candidate translation equivalents, or providing translations from scratch, where no translation equivalent candidate could be provided. The search for contributors has been started at the time of writing this paper.

[15] Accessible at https://lexbib.elex.is/wiki/LexVoc.
[16] Multilingual labels of Wikidata items instances of class Q33742, "natural language".
[17] See https://grobid.readthedocs.io/en/latest/.
[18] The GROBID tool is trained on standardized research paper formats, as found in journals and proceedings. Book chapters are often not correctly parsed, as they usually present a different structure. Thus, for book chapters, we chose the manual approach. Different text encodings found in PDF are also problematic: Diacritics and other special characters may not be correctly recognized by standard tools.
[19] See https://lexbib.elex.is/wiki/Project:SPARQL/examples.
[20] Accessible at https://meet.elex.is/groups/lexicographic-concepts-vocabulary.
[21] Accessible at https://lexonomy.elex.is.
[22] See BabelNet homepage at https://babelnet.org/.

4 Elexifinder version 2

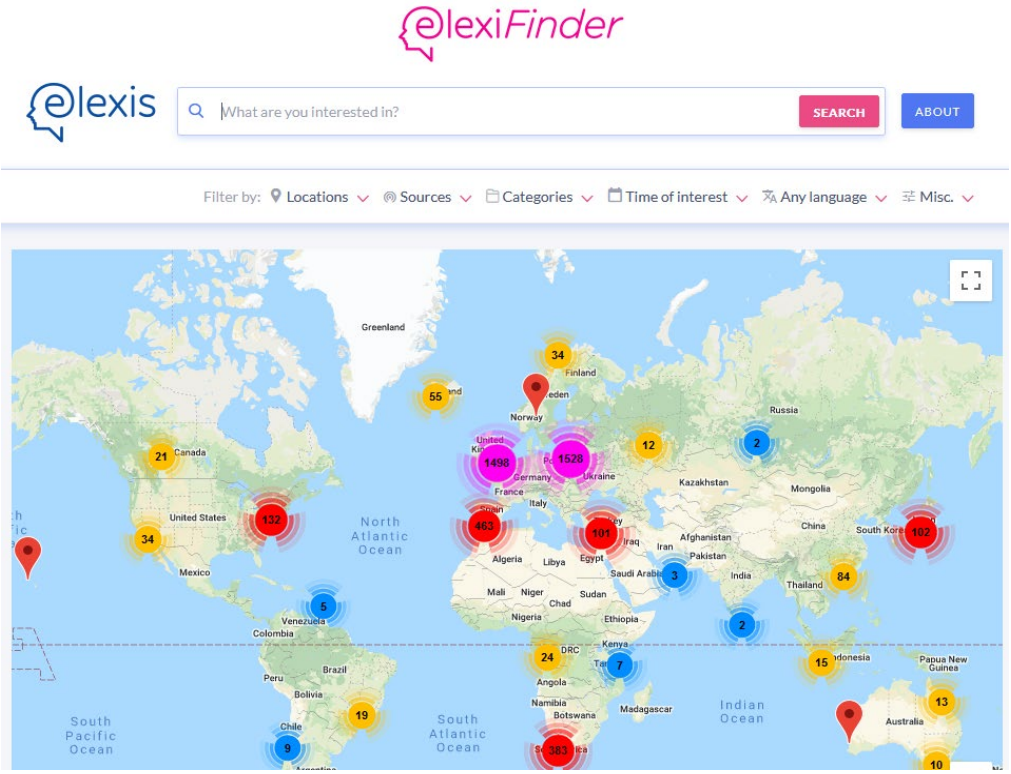

Figure 4: Elexifinder search portal at https://finder.elex.is

4.1 Contents

The most recent version of Elexifinder, updated in early 2021, contains 6,482 publications and 86 videos; in other words, 4,727 publications and 8 videos were added since version 1. The main share of new publications has come in the form of journal papers, and chapters in collective volumes and handbooks. The full list of the contents is as follows:[23]

- Conference proceedings (2,018 articles):
  - eLex (2009-2019)
  - Euralex (1983-2020)
  - Globalex (2016-2020)
- Journals (4,145 articles):
  - International Journal of Lexicography (1988-2019)
  - Lexikos (1991-2019)
  - LexicoNordica (1994-2017)
  - Lexicographica (1985-2019)
  - Nordiske Studier i Leksikografi (1992-2018)
  - Lexicon (1995-2018)
  - Asialex (2014-2020)
  - Slovenščina 2.0 (2013-2020)
  - Revista de Lexicografía (1994-2019)
- Collective Volumes and Handbooks (319 articles):
  - Gouws et al. (eds.). 2013. Dictionaries. An International Encyclopedia of Lexicography (HSK 5/4).
  - Teubert (ed.). 2007. Text Corpora and Multilingual Lexicography.
  - Fuertes Olivera (ed.). 2010. Specialised Dictionaries for Learners.
  - Müller-Spitzer. 2014. Using Online Dictionaries.
  - Jackson (ed.). 2013. Bloomsbury Companion to Lexicography
  - Braun et al. (eds.). 2003. Internationalismen II: Studien zur interlingualen Lexikologie und Lexikographie.
  - Domínguez Vázquez et al. (eds.). 2014. Zweisprachige Lexikographie zwischen Translation und Didaktik.
  - Domínguez Vázquez et al. (eds.). 2014. Lexicografía de las Lenguas Románicas.
  - Gottlieb & Mogensen (eds.). 2007. Dictionary Visions, Research and Practice.
  - Wiegand (ed.). 2000. Wörterbücher in der Diskussion IV: Vorträge aus dem Heidelberger Lexikographischen Kolloquium.
  - Wiegand (ed.). 2002. Perspektiven der pädagogischen Lexikographie des Deutschen II, Untersuchungen anhand

[23] Full reference of edited volumes is found on Zotero (http://lexbib.org/zotero).

des »de Gruyter Wörterbuchs Deutsch als Fremdsprache«.
    - Sterkenburg (ed.). 2003. A Practical Guide to Lexicography.
- Videos (86 items):
    - eLex 2011
    - Euralex 2018
    - WNLex workshop (2018)
    - 15 lexicographic presentations in Slovenia (in Slovene and English)

This dramatic increase in the number of publications has been accompanied by a similar increase in the number of languages represented in Elexifinder. It now contains publications written in 20 different languages:

- English: 3729
- German: 829
- Danish: 436
- Spanish: 390
- Swedish: 362
- French: 195
- Norwegian Bokmål: 178
- Afrikaans: 136
- Slovene: 81
- Italian: 40
- Nynorsk: 36
- Portuguese: 12
- Russian: 10
- Dutch: 7
- Modern Greek: 5
- Catalan: 4
- Belarusian: 1
- Finnish: 1
- Croatian: 1

English still dominates as the language of articles and videos. However, the distribution of publications per language has become more balanced. Most notably, in version 1, articles and videos in English represented 84% of total contents, while in version 2 the share has dropped to 57%.

4.2 Interface

Improvements have also been made to the Elexifinder interface, both in terms of functionality and user-friendliness. The two major upgrades are related to the improved author disambiguation workflow and subject indexation, mentioned in Sections 2 and 3 respectively. The author disambiguation procedure, which for Elexifinder allows the selection of one name representation for all author name variants, means that it is now much more straightforward to find all publications of a certain author. In version 1, this was done by using the Sources/Authors filter option and selecting all relevant name variants; in version 2, not only does the user select only one name, it is now also possible to obtain the list of publications by entering/finding the author's name in the main search window.
Elexifinder offers a category-based system search, which allows to browse articles according to certain content-describing terms, and/or to perform cascaded searches, i.e. to filter sets of displayed search results. In version 2, DMOZ all-domain categories, which were initially used as a temporary solution, have been replaced by the lexicography-oriented controlled vocabulary described in Section 3 above (see Figure 5 for a display of categories relevant in articles authored or co-authored by Patrick Hanks).

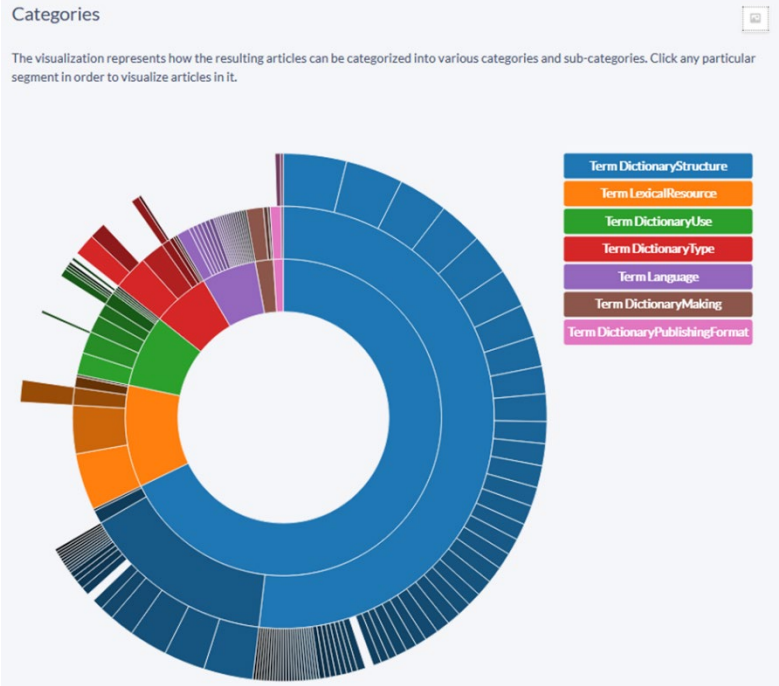

Figure 5: Categories relevant for author Patrick Hanks

With publications in many languages now represented in Elexifinder - and more languages to be added - it becomes necessary for the users to be able to find research publications on a topic of interest in all the languages. This need is addressed by a newly added cross-linguistic searching via concepts, which are identified by wikification, "a process of entity linking that uses Wikipedia as the knowledge base" (Leban, Fortuna & Grobelnik 2016). In other words, publications in different languages dealing with similar topics are indexed, and found, with the same concepts (in English), provided that the concepts and their translations are found on Wikipedia. While the LexVoc subject vocabulary (see Section 3) is being translated, this approach already offers an immediate solution for multilingual information retrieval.
Another search-related improvement is the option of searching the publications and videos published in a certain source. Using the search by group in the Sources/Sources filter, the users can now for example limit their search to only eLex conference proceedings or Lexikos journal contents, etc.

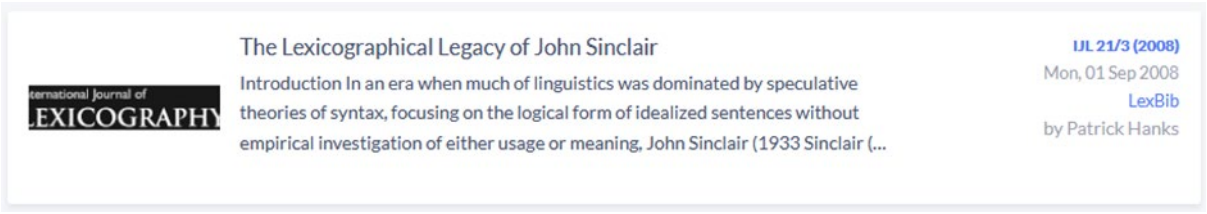

Figure 6: Elexifinder bibliographic item display

An important addition to the result display is the link for each publication to its corresponding entry in the LexBib collection on Zotero (see Figure 6). The decision to add this link was made in order to meet the needs of users who require the complete metadata of a publication, for example for citing purposes.
Lastly, each publication now also comes with a source logo image (see Figure 6). We have decided to add this feature in order to make the sources of publications more easily and quickly identifiable. This conveniently complements the information of the source in the top right corner of the item display, where the name of the source, including the year of publication and issue if relevant, has to be provided in a relatively short form due to space constraints. This abbreviated source information is, at the same time, a hyperlink that leads to the website associated with the source item.

5 Conclusions and Outlook

Elexifinder has come a long way in terms of contents and functionality; however, there is still much to be done. In addition to adding more publications and making the tool even more user-friendly, immediate challenges to be addressed are dissemination among the lexicographic community, and sustainability.
As the majority of major lexicographic conferences and journals, which mainly contain papers in English, are now found

in Elexifinder,[24] we have shifted our focus to acquiring (open access) lexicographic research in other languages, and to lexicographically-relevant research published in volumes or presented at conferences with a scope that is not strictly lexicographic. As identifying such publications is a difficult task, we have asked members of the lexicographic community for help, with initially turning to the lexicographers and researchers coming from ELEXIS partner and observer institutions, in order to test the approach. For the submission of suggestions for new content, and related discussions, we have decided to use LexMeet,[25] a newly developed platform for the lexicographic community to discuss issues, ideas, project results, collaborations etc.

As mentioned in Section 3, we have also launched a project of translating the controlled vocabulary of English terms. Contributors will be using the Lexonomy dictionary writing software on the ELEXIS website for editing translation equivalents. The LexMeet platform will be used for discussions, either general ones by the entire contributor group, or language-specific by contributors working on a specific language. The multilingual dataset will also be publicly accessible throughout the translation process, and when finalized, published in a CLARIN repository.

As far as the dissemination is concerned, a longer and global dissemination campaign is on the way, with the aim of making the community aware of the tool, and helping them with using it. At the same time, we will aim to obtain as much feedback as possible, for example by using an online survey, in order to identify further ways in which to improve Elexifinder.

Finally, it is important to dedicate some time to thinking about sustainability of the tool and its contents, especially as the end of ELEXIS project is drawing near. The greatest sustainability challenge is connected to contents; we need to find a way to keep them regularly updated. One of the solutions currently explored is to make the Elexifinder workflow a community-driven project, where lexicographers not only submit requirements or new bibliographical datasets, but also divide the task of data curation, and upload to Elexifinder.

6 References

Córdoba Rodríguez, F. (2003). *Bibliografía Temática de La Lexicografía*. A Coruña: Universidade da Coruña.

Engelberg, S., Storrer, A. (2016). Typologie von Internetwörterbüchern und -portalen. In A. Klosa & C. Müller-Spitzer (eds.) *Internetlexikografie. Ein Kompendium*. Berlin/New York: de Gruyter, pp. 31–63.

Kipfer, B.A. (2013). Glossary of Lexicographic Terms. In H. Jackson (ed.) The Bloomsbury Companion to Lexicography. London: Bloomsbury, pp. 391–406.

Kosem, I., Krek, S. (2019). ELEXIFINDER: A Tool for Searching Lexicographic Scientific Output. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (eds.) Electronic Lexicography in the 21st Century: Proceedings of the ELex 2019 Conference. Brno: Lexical Computing CZ s.r.o., pp. 506–18.

Leban, G., Fortuna, B., Brank, J. & Grobelnik, M. (2014). Event Registry: Learning about World Events from News. In *Proceedings of the 23rd International World Wide Web Conference, WWW14, Seoul, Korea, April 7-11, 2014*, pp. 107–10.

Leban, G., Fortuna, B. & Grobelnik, M. (2016). Using News Articles for Real-Time Cross-Lingual Event Detection and Filtering. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval (NewsIR16), Padova, Italy, March 20, 2016*, pp. 33–38.

Leban, G., Fortuna, B. & Grobelnik, M. (2017). Event Extraction from Media Texts. In C. Sammut & G.I. Webb (eds.) Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US, pp. 416–22.

Lindemann, D., Klaes, C. & Zumstein, P. (2019). Metalexicography as Knowledge Graph. Edited by Maria Eskevich, Gerard De Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek & Milan Dojchinovski. *Open Access Series in Informatics* 70.

Lindemann, D., Kliche, F. & Heid, U. (2018). Lexbib: A Corpus and Bibliography of Metalexicographcal Publications. In *Proceedings of EURALEX 2018*. Ljubljana, pp. 699–712.

Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Tbilisi State University, pp. 906–9.

Müller-Spitzer, C., ed. (2014). *Using Online Dictionaries*. Lexicographica Series Maior 145. Berlin: De Gruyter.

Romary, L. & Lopez, P. (2015). GROBID - Information Extraction from Scientific Publications. *ERCIM News, Scientific Data Sharing and Re-Use* 100 (January).

Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The authors wish to thank Christiane Klaes from University of Hildesheim, who in the framework of her MA thesis project has defined the algorithm used for author name variant clustering (see section 2).

[24] One notable omission in the existing dataset is Dictionaries, a journal of the Dictionary Society of North America, but we already have its whole content in our database and will be adding it with the next Elexifinder update.

[25] Accessible at https://meet.elex.is/.