Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 1–5
Marseille, 20 June 2022

Towards the Profiling of Linked Lexicographic Resources

Lenka Bajčetić[1], Seung-Bin Yim[1], Thierry Declerck[2]
[1] Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Vienna, Austria
[2] DFKI GmbH, Multilinguality and Language Technology Lab, Saarland University Campus D3 2, Germany
[1]Lenka.Bajcetic@oeaw.ac.at, Seung-Bin.Yim@oeaw.ac.at, declerck@dfki.de

Abstract

This paper presents Edie: ELEXIS Dictionary Evaluator. Edie is designed to create profiles for lexicographic resources accessible through the ELEXIS platform. These profiles can be used to evaluate and compare lexicographic resources, and in particular they can be used to identify potential data that could be linked.

Keywords: ELEXIS, Lexicographic Profiling, Dictionary evaluation

1. Introduction

The work described in this paper is done in the context of the ELEXIS project,[1] which is dealing with the building of a large European lexicographic infrastructure. It pursues this goal by providing the lexicographic infrastructure with interactions with Natural Language Processing (NLP) tools and resources, for both access to and creation of linked lexical data. The resulting multilingual infrastructure is intended to be used by academics, students, researchers, programmers, dictionary creators, etc.

At the core of ELEXIS is the so-called dictionary matrix, a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, multilingual, monolingual, modern, historical etc. Data from the dictionary matrix is available through a RESTful Web service, which also make the data available for consumption through tools to Sketch Engine and Lexonomy.[2] Edie is situated at this access interface. Figure 1 shows the overall architecture of ELEXIS and the place the dictionary matrix has in this infrastructure.

ELEXIS offers a well-defined interface (McCrae et al., 2019) that supports the access to the data sets hosted by the ELEXIS infrastructure, but it also guides users by the creation, modification, and publication of dictionaries with the ELEXIS infrastructure. Figure 2 sketches the access procedure to (linked) lexical data included in the dictionary matrix, where we can see that the data is serialized in three different formats: TEI,[3] OntoLex-Lemon,[4] or JSON.[5] This is the lexical data which EDIE is accessing and profiling. Edie can retrieve this data via the Lexonomy interface as a dictionary, a lexical entry or a lemma, and generate profiles based on this information, both at the level of metadata and data.
Table 1shows the kind of information Edie is accessing, when querying for a dictionary within the ELEXIS infrastructure.
and the Table 2 shows the type of information that is accessed by Edie when querying for an individual entry of a dictionary.
Edie can also access lemma information.
Since there are numerous possible use-cases, as well as different types of end users, we needed to create a generic dictionary assessment tool which would work best under these ambiguous circumstances. Since we cannot make any definitive assumptions regarding the goal of the end users and their priorities regarding dictionary quality, we have decided to create a tool which would leave the final evaluation to the end users, while providing them with a profile with enough information to make their own estimate. The tool is described in the next section.

2. Edie

EDIE is an acronym for the ELEXIS DIctionary Assessment tool[6]. This tool is aimed to assist users with context-dependent qualitative assessment of linguistic resources by creating lexicographic profiles which can be easily compared and evaluated by the end user.

2.1. Implementation

The EDIE infrastructure consists of three main components:

- the main evaluator which consists of three evaluator modules

[1] See `https://elex.is/` and (Woldrich et al., 2021) for more details.

[2] See `https://www.sketchengine.eu/` and `https://www.lexonomy.eu/` repsectively

[3] See `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`

[4] See `https://www.w3.org/2016/05/ontolex/`

[5] See `https://www.json.org/json-en.html`

[6] The code is available here: https://github.com/ELEXIS-eu/edie and the service will be deployed shortly on the ELEXIS platform

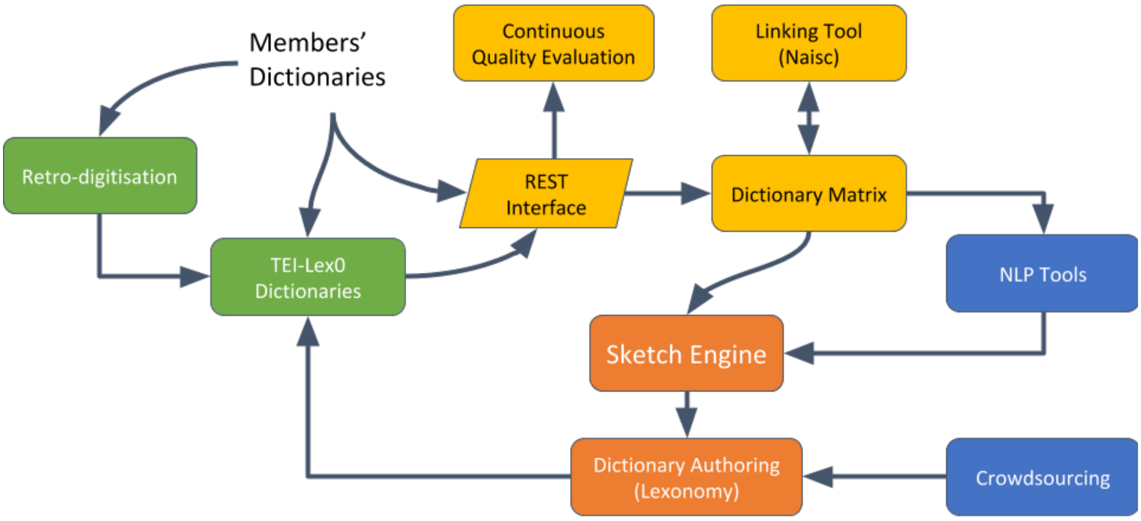


Figure 1: The interface for accessing lexical data in the dictionary matrix, taken from (McCrae et al., 2019)

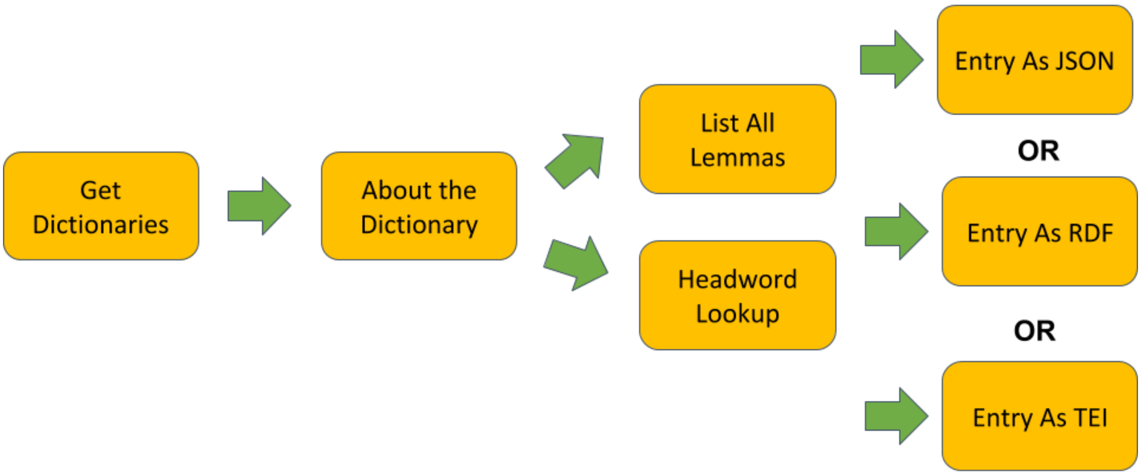


Figure 2: Overall Architecture of ELEXIS, taken from (McCrae et al., 2019)

- API client which retrieves necessary data
- helper functions

The three evaluator modules are designed to assess different aspects of the resources, and in combination they create the resource's profile. The content of a lexical resource is represented on entry level by a model which has all the fields an entry could have, e.g. lemma, senses, examples, part of speech, etc. Iterating through the entries of a lexical resource, EDIE creates a statistical overview of a 'typical' entry, defining the average structure and type of information which can be found in such a dictionary and providing the user a quick insight into the dictionary structure, sense granularity, and the type of information they can expect to encounter.
Besides the content of a dictionary, EDIE also takes into account the resource's metadata. The metadata information which can be found in the Elexis infrastructure is represented by the metadata model which has all fields defined by Dublin Core, and those used by the whole Elexis infrastructure. Since an automatic verification of the accuracy or quality of the metadata is too advanced, the metadata evaluation only takes into account the completeness of the data. This means the final profile of the resource will consist of a summary of the existing metadata, accompanied with a list of any missing information.

Finally, the provided metadata is also used to perform context-specific profiling and resource comparison. We call this "aggregated" profiling because it aims to contextualize a particular resource by comparing it to others, thus providing a more comprehensive resource profile. The language and type of a resource are used so that the output of our assessment would provide the user information within a sensible context. If a dictionary is categorized as a terminological dictionary of French, we can compare its properties to other terminological dictionaries of French. This way, we make sure that the comparisons we make are useful and reasonable. For instance, if a user wants to make sure that

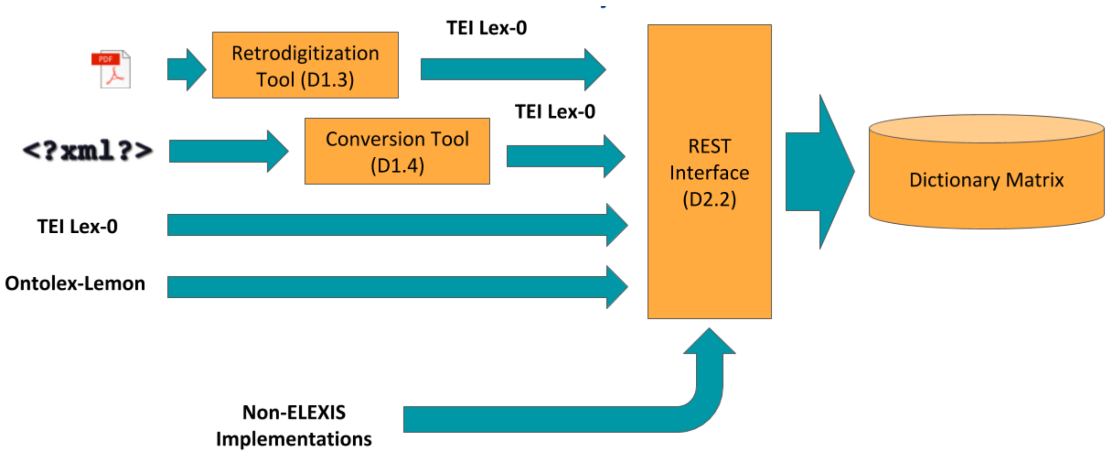


Figure 3: How to upload lexical resources to the dictionary matrix, taken from (McCrae et al., 2019)

Method Name:	`/about`
Parameters:	The dictionary ID
Returns:	An object describing the dictionary
Example Request:	`http://www.example.com/about/example-dictionary`

Example Response:	`{`
	`"release": "PUBLIC",`
	`"sourceLanguage": "en",`
	`"targetLanguage": [ "en", "de" ],`
	`"genre": [ "gen" ],`
	`"license": "creativecommons.org/licenses/by/4.0/",`
	`"title": "The Human-Readable Name of this resource",`
	`"creator": [{`
	`    "name": "Institute of This Resource",`
	`    "email": "contact@institute.com"`
	`    }],`
	`"publisher": [{`
	`    "name": "Publishing Company"`
	`    }`
	`]}`

Table 1: Type of information returned by querying for a dictionary within the ELEXIS infrastructure

they are using the largest available resource in a particular category, they can easily see how the resource compares in size with the other resources in that category.

2.2. Usage

As previously mentioned, EDIE is situated at the access interface for the dictionary matrix, and it can be accessed through a RESTful Web service. Since there are many dictionaries with several thousands of entries, creating their profiles can take time. Additionally, we can assume that the data will not be changed frequently. In order to save time, a resource is profiled as soon as it is added to the dictionary matrix, and this profile is later accessed on user demand. If the resource content or metadata is altered in any way, the profile is created anew. Since aggregated evaluation takes into account several dictionaries depending on the categorization created by the user, this cannot be done in advance. However, aggregating does not take too long because the system works with the existing profiles.
Once a user selects the resource they are interested in, or the category they wish to compare using aggregated profiling, they can send a parameterized request to EDIE using the REST API, and quickly get a response in JSON format. The response is EDIE's end report which consists of the resource's content statistics, metadata with the missing data pointed out, formatting errors, and the aggregation profile if requested. A sample of the end report can be seen in Figure 4.

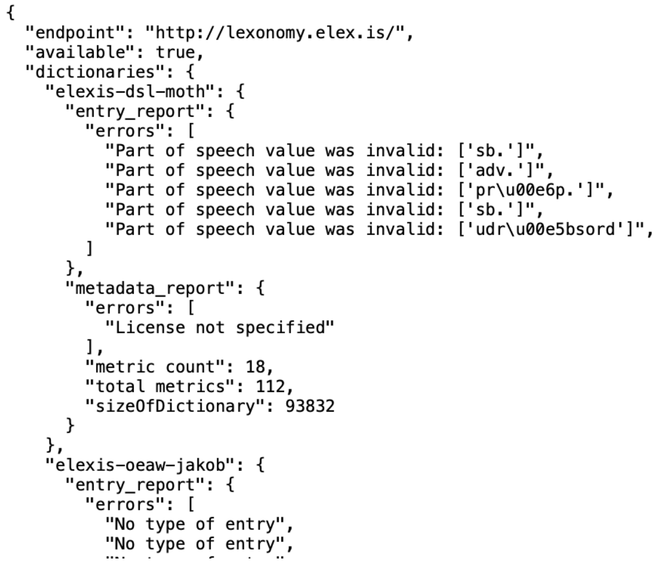
3. Related work

Evaluation of dictionaries and linguistic resources relies on the accuracy and thoroughness of the metadata which accompanies them. Without relevant information regarding the resource, the user cannot create a

Method Name:	`/list/`*dictionary*
Parameters:	A limit and an offset
Returns:	A list of lexical entry descriptions
Example Request:	`http://www.example.com/list/example-dictionary?limit=2`

Example Response:	`[`
	`{`
	`"release": "PUBLIC",`
	`"lemma": "work",`
	`"language": "en",`
	`"id": "work-n",`
	`"partOfSpeech": [ "NOUN" ],`
	`"formats": [ "tei" ]`
	`}, {`
	`"release": "PUBLIC",`
	`"lemma": "work",`
	`"language": "en",`
	`"id": "work-v",`
	`"partOfSpeech": [ "VERB" ],`
	`"formats": [ "tei" ]`
	`}`
	`]`

Table 2: Type of information returned by querying for an individual entry of a dictionary within the ELEXIS infrastructure



```
{
  "endpoint": "http://lexonomy.elex.is/",
  "available": true,
  "dictionaries": {
    "elexis-dsl-moth": {
      "entry_report": {
        "errors": [
          "Part of speech value was invalid: ['sb.']",
          "Part of speech value was invalid: ['adv.']",
          "Part of speech value was invalid: ['pr\u00e6p.']",
          "Part of speech value was invalid: ['sb.']",
          "Part of speech value was invalid: ['udr\u00e5bsord']",
        ]
      },
      "metadata_report": {
        "errors": [
          "License not specified"
        ],
        "metric count": 18,
        "total metrics": 112,
        "sizeOfDictionary": 93832
      }
    },
    "elexis-oeaw-jakob": {
      "entry_report": {
        "errors": [
          "No type of entry",
          "No type of entry",
```

Figure 4: A sample of EDIE's end report

verdict about the quality or the usability of a particular resource for their purpose. The assessment of metadata provided with a lexicographic resource is also called *metalexicography* (Swanepoel, 2008).

One example of metadata schema used to evaluate and connect language resources is given by the META-SHARE ontology, which is described in (Gavrilidou et al., 2012).[7] While the META-SHARE ontology is a very important resource for our work, we are not aware of any initiative using it for (automatic) usability assessment of lexical resources.

Another initiative related to this topic of accessing metadata of linguistic resources is ”LingHub” ((McCrae and Cimiano, 2015))[8], which is combining metadata from different schemes, like LRE-MAP, META-SHARE, CLARIN and more. This integration is resulting in an RDF-based set of metadata that are greatly improving the discovery of language resources. But

[7] The latest version of the META-SHARE ontology is available at `http://www.meta-share.org/ontologies/meta-share/meta-share-ontology.owl/documentation/index-en.html`.

[8] See also `https://linghub.org/`.

LingHub is not dealing directly with the data itself, and the quality issues dealt with by the developers of LingHub are primarily concerning the encoding of the metadata.

In the field of profiling Knowledge Graphs (KG) We are aware of work pursued within the COST Action ”NexusLinguarum”[9] and dealing with data profiling in the Linguistic Linked Open Data (LLOD)[10], using for this the ABSTAT tool ((Spahiu et al., 2018) ; (Principe et al., 2018))[11] This work is dealing primarily with the establishment of specific metrics to describe the structural features, or schema-level patterns, of knowledge graphs encoding linguistic data – basically the data sets included in the LLOD cloud. But it doesn't address directly the linguistic features included in those data, and their compliance to a standardized vocabulary.

As it has been noticed by (Rabby et al.,), sets of schema-level patterns delivered by profiling tools such as ABSTAT ((Principe et al., 2018)), may be huge, and might deal with very generic features. Therefore our approach in Edie is focusing directly on the content of the RDF-based lexical data sets included in the dictionary matrix.

4. Conclusions and Future work

We have presented EDIE, the tool designed for profiling lexicographic resources within the ELEXIS infrastructure. EDIE is designed to allow users to assess different aspects of dictionaries based on their metadata and entries. Furthermore, users can utilize aggregated profiling to compare relevant dictionaries for their specific use cases. The current implementation of EDIE does not have any graphical user interface for interactive exploration of the lexicographic resources. Such an user interface in combination with different statistics and comparative visualizations based on different criteria selected by users (dictionary types, genres, languages, etc.) would help the users to assess different dictionaries in a more user-friendly manner.

5. Acknowledgements

This paper is based upon work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015).

6. Bibliographical References

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The METASHARE metadata schema for the description of language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1090–1097, Istanbul, Turkey, May. European Language Resources Association (ELRA).

McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In Agata Filipowska, et al., editors, *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS 2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15) 11th International Conference on Semantic Systems - SEMANTiCS 2015, Vienna, Austria, September 15-17, 2015*, volume 1481 of *CEUR Workshop Proceedings*, pages 88–91. CEUR-WS.org.

McCrae, J. P., Tiberius, C., Khan, A. F., Kernerman, I., Declerck, T., Krek, S., Monachini, M., and Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*, pages 642–659, Sintra, Portugal, 10.

Principe, R. A. A., Spahiu, B., Palmonari, M., Rula, A., Paoli, F. D., and Maurino, A. (2018). ABSTAT 1.0: Compute, manage and share semantic profiles of RDF knowledge graphs. In Aldo Gangemi, et al., editors, *The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers*, volume 11155 of *Lecture Notes in Computer Science*, pages 170–175. Springer.

Rabby, G., Keya, F., Svátek, V., and Principe2, R. A. P. (). Effect of heuristic post-processing on knowledge graph profile patterns: cross-domain study. unpublished.

Spahiu, B., Maurino, A., and Palmonari, M. (2018). Towards improving the quality of knowledge graphs with data-driven ontology patterns and SHACL. In Martin G. Skjæveland, et al., editors, *Proceedings of the 9th Workshop on Ontology Design and Patterns (WOP 2018) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 9th, 2018*, volume 2195 of *CEUR Workshop Proceedings*, pages 52–66. CEUR-WS.org.

Swanepoel, P. H. (2008). Towards a framework for the description and evaluation of dictionary evaluation criteria. *Lexikos*, 18:207–231.

Woldrich, A., Goli, T., Kosem, I., Matuška, O., and Wissik, T. (2021). ELEXIS: Technical and social infrastructure for lexicography, July. Published in K Lexical News (28), pp. 45-52.

[9] nexuslinguarum.eu/.

[10] See https://linguistic\protect\discretionary{\char\hyphenchar\font}{}{}lod.org/ for more details on the LLOD cloud.

[11] See the ”Intermediate Activity Report Working Group 1 'Linked data-based language resources'” of the NexusLinguarum COST Action at https://nexuslinguarum.eu/wp-content/uploads/2021/11/D1.3_IntermediateActivityReport.pdf.