New Developments in the LRE Map

Vladimir Popescu[1], Lin Liu[1], Riccardo Del Gratta[2], Khalid Choukri[1], Nicoletta Calzolari[2]
[1]ELDA, [2]ILC–CNR
[1]9 rue des Cordelières, Paris, France,
[2]Via Giuseppe Moruzzi nr. 1, Pisa, Italy
{vladimir,lin,choukri}@elda.org, {riccardo.delgratta,glottolo}@ilc.cnr.it

Abstract

In this paper we describe the new developments brought to LRE Map, especially in terms of the user interface of the Web application, of the searching of the information therein, and of the data model updates. Thus, users now have several new search facilities, such as faceted search and fuzzy textual search, they can now register, log in and store search bookmarks for further perusal. Moreover, the data model now includes the notion of paper and author, which allows for linking the resources to the scientific works. Also, users can now visualise author-provided field values and normalised values. The normalisation has been manual and enables a better grouping of the entries. Last but not least, provisions have been made towards linked open data (LOD) aspects, by exposing an RDF access point allowing to query on the authors, papers and resources. Finally, a complete technological overhaul of the whole application has been undertaken, especially in terms of the Web infrastructure and of the text search backend.

Keywords: Language resource, LRE Map, Information search and retrieval, Data modelling

1. Introduction

Initiated by ELRA and FlaReNet and introduced at LREC 2010, the LRE Map is a new mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources during the process of submitting articles to conferences.

It is a collective enterprise of the LREC community, as a first step towards the creation of a very broad, community-built, Open Resource Infrastructure. It is meant to become an essential instrument to monitor the field and to identify shifts in the production, use, and evaluation of LRs and LTs over the years.

At LREC 2010, nearly 2000 language resource forms have been filled in. Apart from providing a portrait of the resources behind the community, of their uses and usability, the LRE Map intends to be a measuring instrument for monitoring the field of language resources.

The feature has been so successful that it has been implemented also at COLING 2010, EMNLP 2010 and many other conferences, while other major events are in the pipeline, in addition to the LRE Journal.

Since the LRE Map service has been set up, several new developments have been made.

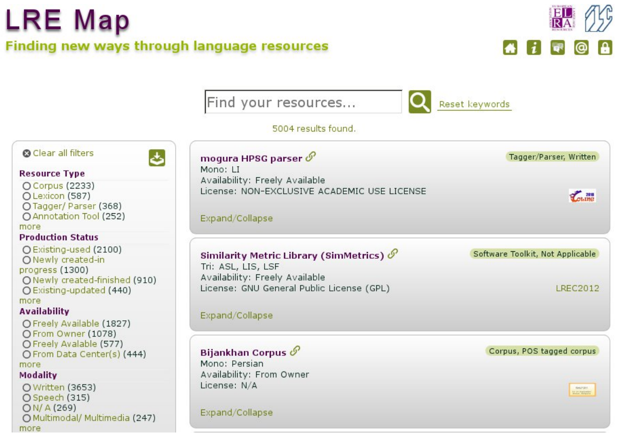
These developments are mostly concentrated in three directions:

1. Upgrades of the Web platform
2. Extensions to the database model
3. Cleanup and normalization of the data

The present paper documents the evolution of the LRE Map with respect to these directions.

2. LRE Map Platform Update

The LRE Map Platform Upgrading has been undertaken by ELDA and consisted of the following actions:



5004 results found.

Resource Type
○ Corpus (2233)
○ Lexicon (587)
○ Tagger/ Parser (368)
○ Annotation Tool (252)
more

Production Status
○ Existing-used (2100)
○ Newly created-in progress (1300)
○ Newly created-finished (910)
○ Existing-updated (440)
more

Availability
○ Freely Available (1827)
○ From Owner (1078)
○ Freely Avalable (577)
○ From Data Center(s) (444)
more

Modality
○ Written (3653)
○ Speech (315)
○ N/ A (269)
○ Multimodal/ Multimedia (247)
more

mogura HPSG parser — Tagger/Parser, Written
Mono: LI
Availability: Freely Available
License: NON-EXCLUSIVE ACADEMIC USE LICENSE

Similarity Metric Library (SimMetrics) — Software Toolkit, Not Applicable
Tri: ASL, LIS, LSF
Availability: Freely Available
License: GNU General Public License (GPL)
LREC2012

Bijankhan Corpus — Corpus, POS tagged corpus
Mono: Persian
Availability: From Owner
License: N/A

Figure 1: The new LRE Map Application.

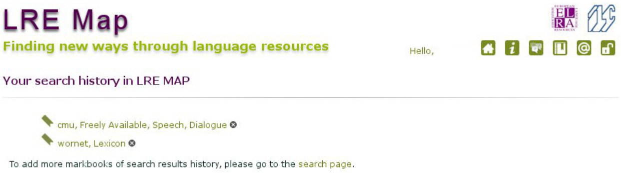
- The web platform technology has been migrated from a PHP/MySQL solution to a more modern Python/Django solution, backed by a PostgreSQL database [1].

  The home page of the new LRE Map is displayed in Figure 1.

  Migrating from a MySQL database to a PostgreSQL one integrated into the Django application took a significant amount of data migration work, which was undertaken by ELDA.

- Several new functionalities have been added
  - user registration and login, with per-user search bookmarking and contact form,
  - faceted search with dynamically-built facets and faceted search tags,
  - textual fuzzy search.

[1] A demo version of the new LRE Map application is available at: http://lremap.elra.info/.



LRE Map

Finding new ways through language resources

Hello,

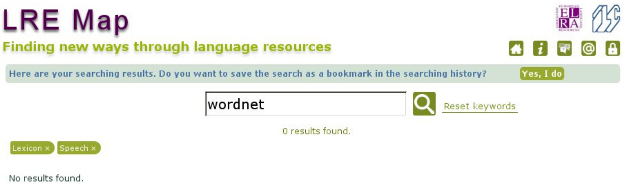
Your search history in LRE MAP

cmu, Freely Available, Speech, Dialogue

wornet, Lexicon

To add more markbooks of search results history, please go to the search page.

Figure 2: Bookmarks in LRE Map.



LRE Map

Finding new ways through language resources

Here are your searching results. Do you want to save the search as a bookmark in the searching history?

wordnet

0 results found.

No results found.

Figure 3: Combining faceted search with full-text search while keeping track of the selected filters.

The first functionality allows users to register and subsequently login, in order to be able to:

- send personalized feedback or information request,
- bookmark their search results; as the search process can be quite complex, combining faceted search with full-text search (see below), this functionality is, in our view, essential to improving the overall usability of LRE Map. An example of a set of bookmarks is shown in Figure 2.

Thus, after having done complex searches, users can come back, log in and reissue the same search queries, as they have been stored as bookmarks.

The second functionality resembles what has been provided before, safe for the fact that now faceted search and text-based search can be chained in any order, and the information of what is currently being searched for is available at any time, including when no results are available.

For instance, when choosing “Resource Type”: “Lexicon”, “Modality”: “Speech” and typing “Wordnet” in the full-text search bar, no results are retrieved. Nevertheless, we are able to either:

- discard the full-text search and back-off to faceted search only;
- or to discard one faceted search filter or both and back-off to full-text search only, or to a mixture of full-text search and fewer faceted search filters.

The example is depicted in Figure 3. Here, the user can, for instance, back-off the [“Modality”:] “Speech” filter, thus having a non-trivial result set, that she or he can further filter by using other criteria.

The third functionality allows the users to get more results, making the full-text searching process more robust to typos in the search queries or in the database itself. In order to understand the full potential of this functionality, one should bear in mind that LRE Map stores, for most of the data items two kinds of information:

1. Data as provided by the original LREC paper submitters to LRE Map (called “original” / “unnormalized” data)
2. Data that has been manually corrected [2] (called “normalized” data).

Now, fuzzy full-text search looks through the normalized data first, backing off to original data when no results are retrieved based on the normalized data, and retrieves the results sorted by their relevance, computed according to the TF-IDF of the terms of the query with respect to the documents being searched.

Thus, full-text search works at the same time for people typing e.g. “canonical” resource names (e.g. “MultiWordNet” for resource (Fondazione Bruno Kessler, 2014)) and for people typing misspelled or incomplete resource names (e.g. “wornet”).

Elasticsearch [3] has been used to power the fuzzy searching process. Backed by the Lucene [4] text search engine library, Elasticsearch is one of the most powerful and customizable (via a JSON RESTful API) search engines available.

Thus, in order to harness its power, a special NLP-oriented configuration has been chosen for Elasticsearch. In the following lines we will briefly describe it. First of all, Elasticsearch allows users to specify what they call the linguistic “analyzer”, which has several components: the language of interest (English in our case), the term tokenizer (in our case, after several trial-and-error experiments, we came to the conclusion that the “standard” [5] tokenizer suffices). The interesting part of the Elasticsearch analyzer configuration resides in the ability to define a pipeline of processing elements (called *filters* in the Elasticsearch parlance). Thus, several filters have been piped together in the Elasticsearch LRE Map fuzzy-search configuration:

1. At first, an English possessive stemmer is used, which removes possessive articles from the indexing and query tokens.
2. Secondly, the “lowercase” filter is used, which converts all indexing and query tokens to lowercase.
3. Then, Krovets stemming (Kstem) is used, which provides fast query token stemming, in a least aggressive manner, by means of a rule-based approach combining inflectional and derivational stemming (Krovetz, 1993). Unlike other stemmers, the Kstem avoids conflating variants with different meanings, thus it avoids changing the query semantics.
4. In order to keep the search precision as high as possible, the “keyword repeat” filter is used, in order to

[2] The corrections have been undertaken by ELDA, by the LIMSI and by the ILC.

[3] https://www.elastic.co/products/elasticsearch

[4] https://lucene.apache.org/

[5] The “standard” tokenizer in Elasticsearch implements the Unicode text segmentation algorithm, as documented in the Annex 29 of the Unicode standard (http://unicode.org/reports/tr29/). The algorithm is quite general in that it handles at the same time word boundaries and sentence boundaries.

allow for indexing a stemmed and an unstemmed version of each token side by side.

5. Then, to prune the search space even further, stopwords are removed by means of the “stop” filter, catered to English.
6. To improve search result recall, the word delimiter token filter is also used, which tokenizes compound words; original tokens are also preserved, in an attempt to limit precision degradation.
7. Then, Porter stemming [6] is also performed on the atomic tokens obtained at the preceding step, in an attempt to improve recall even further.
8. At last, to streamline the indexing process, the “unique” token filter is used, so that duplicate adjacent tokens are removed and thus only unique contiguous tokens are indexed. This last step is also useful for avoiding indexing duplicates produced via the “keword repeat” filter when a token has the same form as its stem [7].

Last but not least, Elasticsearch allows us to control the fuzzy search robustness in a fine-grained manner. Thus, also by trial and error, we have been able to set the tolerance threshold on the Levenshtein-Damerau similarity distance to 2, so that the distance between two “identical” strings is 2 at most, i.e. “worknet” is considered identical to “wornet” and to “wordnet”, and to “woknet” as well, but not to “wokmet” [8].

Had we set this threshold to a value greater than 2, we would have lowered the search precision too much. Conversely, had we used a lower threshold, we would have degraded recall, especially with respect to typos in the search queries, but also with respect to certain database fields which have not been normalized yet.

Moreover, different relevance weights have been assigned to some of the language resource fields. Thus, resource names are ranked higher than paper names, which, in turn, are ranked higher than author names. This entails, e.g., that a resource which contains a term appearing only once in the resource name should be ranked higher than the resource which contains the same term appearing only once, but in the paper title.

Besides these main development directions, several minor improvements have been made, such as making the resource URLs clickable, adding logos to conferences, providing CSV (Comma-Separated Values) export for a limited random sample of resource metadata[9], resource listing pagination, sliding menus for faceted search, modal conference metadata display, etc.

3. New Database Architecture

3.1. General Context

The LRE Map is a user interface and a database. The user interface is embedded in the START [10] article submission system, used for LREC.

The START user interface gathers the information on the used/cited resources and connects them to the paper and its authors. A communication protocol settled between ILC and START is responsible for feeding and updating the database with the information gathered with the user interface. Since its first appearance in the LREC 2010 conference (Calzolari et al., 2010), the LRE Map underwent many changes in its internal structure to address specific requirements of the referenced community. For example, the number of slots dedicated to the languages changed from 3 to 6 after the pioneering work on language matrices (Mariani and Francopoulo, 2012), and some additional details such as size and unit of the described language resource have been added to keep track of the amount of information a LR can provide to the community. All these changes caused variations in the record tracks and, consequently, in the database structure. Moreover, the LREC 2016 database contains the ISLRN table which connects the resources described with the official number provided by ELRA.

3.2. Data Normalization Process

The database contains the data collected through the interface and provided by paper authors: the use of acronyms for well-known resources as well as typos are frequent. The data collected needed to be normalized according to (at least) the following items:

- Acronym Resolution. Where it comes that BNC becomes British National Corpus (BNC), for example;
- Typo correction and clustering of similar terms. Where `Emglish` is corrected into `English` and `freely avail.` into `Freely-Available` and so on.
- Standardization. Values of important metadata, such as types and uses, provided by authors and that occur many times and/or are considered relevant for the community have been inserted in the list of official metadata provided by the LREC committee through the user interface.

The normalization process defines two distinct databases: one which contains the original uncorrected values and one with all revised records. Clearly these two distinct databases are connected through a unique key. Having two distinct data sources is important also for the new interface and services described in Section 1. In the same GUI or with the same service it is possible to search across both

[6] The Porter stemming algorithm is generally much greedier than Kstem, in that semantically unrelated words are sometimes stemmed in the same way, e.g. “universal”, “university” and “universe” are all stemmed to “univers”. See also (Porter, 1980).

[7] https://www.elastic.co/guide/en/elasticsearch/reference/1.4 /analysis-keyword-repeat-tokenfilter.html

[8] The latter is however considered identical to “wornet”, as there are two character differences, hence when querying for “wopnet”, “WorNet”-related results are retrieved, but not “Wordnet”-related ones, which, however, are retrieved when looking for “wornet”.

[9] More resources are exported for registered users.

[10] START is an integrated web-based solution for managing peer-reviewed conferences and workshops (https://www.softconf.com).

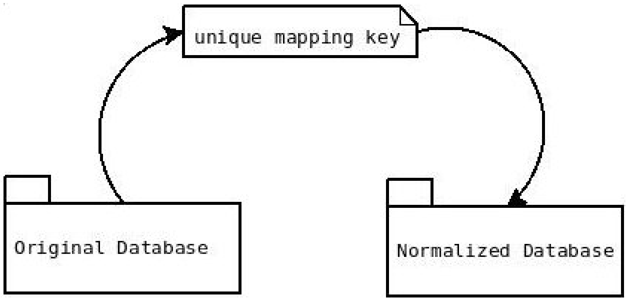


Figure 4: The two data sources are connected through a unique key.

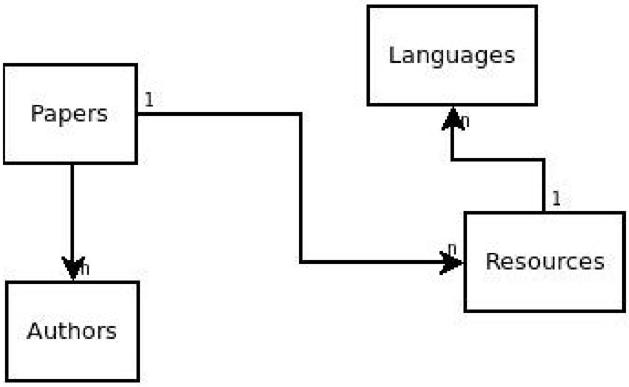


Figure 5: Entity types in the database.

data sources and it will soon be possible to surf from one source to the other and vice versa showing original and corrected values at the same time, as shown in Figure 4.

3.3. The LRE Map Database: entities and relations

In Figure 5 we show the complete picture of the parallel data sources.
The following entity types are defined:

1. Resources: they collect all described LRs during the submission phase
2. Authors: they collect the authors of all accepted conference papers
3. Papers: they collect all accepted papers
4. Languages: they collect all languages of a given LR instance.

While Resources and Languages provide pictures of the Language Resources landscape: their uses, status, availability, modality and languages, Authors and Papers add different dimensions to the LRE Map. Through Papers we can access topics and keywords as they have been selected by the authors and map them to the automatic results obtained by Saffron [11] for example. Authors, instead, provide information on affiliations and countries. The LRE Map is mapped onto the Earth, thanks to countries, and the language resources described acquire a geographical feature. In other words, the LRE Map can be analyzed according to geographical dimensions. For example, how many LRs created in the USA are used in the USA only and how many Under-Resourced Languages are used, where and by whom are types of analysis that can be easily carried out via the LRE Map.

3.4. Model Upgrades in the LRE Map Application

The new, much richer, database model has been integrated into the current LRE Map application and is currently pending internal validation. To this end, two steps have been undertaken:

1. the old LRE Map database model has been migrated to the model described in this paper
2. the database has been filled in with the supplementary information pieces, mostly concerning author, paper and submission information; currently, this information is available in the new database for LREC 2014.

Last but not least, users are now able to visualize normalized and non-normalized (original) data, in a standalone manner or in a side-by-side manner.

4. Conclusions and Perspectives

After several years of usage, LRE Map is now a mature Web platform for tracing the *usage history* of language resources, as reported in scientific conferences. However, there are still some missing spots: for example, in order to reach its full potential, we believe that the LRE Map would benefit from exposing an RDF / SPARQL end-point, thus integrating into the Linked Open Data (LOD) landscape.
Linked Open data (LOD) (Chiarcos and Nordhoff, 2012) represents a new modality of data presentation and storage. It is deeply inside the Semantic Web Paradigm and allows at connecting data sets serialized in RDF. Also the LREMap follows this trend, as described in (Del Gratta et al., 2014) and the data of LREC2014 have been partially released in RDF [12]. The LOD paradigm helps also in mapping the aforementioned entities into available ontologies: Authors onto Friend Of A Friend (FOAF) [13], Papers onto Bibliographic Ontology (BIBO) [14] and Languages onto LEXVO [15]. In this way, the LREMap database is automatically connected to the LOD cloud [16].
In order to reap the full benefits of this, the newly-developed LRE Map application developed by ELDA needs to be integrated with the LOD provisions developed by the ILC.

5. Bibliographical References

Calzolari, N., Soria, C., Gratta, R. D., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The lrec map of language resources and

[11] http://saffron.insight-centre.org/

[12] See http://www.resourcebook.eu/lremap/owl/instances/.

[13] http://xmlns.com/foaf/spec/

[14] http://bibliontology.com/

[15] http://www.lexvo.org/

[16] http://lod-cloud.net/

technologies. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh Language Resources and Evaluation Conference*, pages 949–956, Valletta, Malta. European Language Resources Association (ELRA).

Chiarcos, C. and Nordhoff, S. (2012). *Linked Data in Linguistics*. Springer, Berlin.

Del Gratta, R., Pardelli, G., and Goggi, S. (2014). Lremap disclosed. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3534–3541. European Language Resources Association (ELRA).

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202. Association for Computer Machinery (ACM).

Mariani, J. and Francopoulo, G., (2012). *The Language Matrices and the Language Resource Impact Factor*, pages 441–471. Springer, Berlin.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

6. Language Resource References

Fondazione Bruno Kessler. (2014). *MultiWordNet*. Fondazione Bruno Kessler, WordNet, 1.0, ISLRN 103-401-284-171-1.