OR2019 Presentation Submission
14th International Open Repositories Conference, June 10th-13th, Hamburg, Germany

Improving LA Referencia metadata by linking research profiles to repositories: the case of the Brazilian Digital Library of Thesis and Dissertations (BDTD) and the Lattes CV Platform

Lautaro J. Matas, LA Referencia / Observatorio CTS - OEI, lmatas@gmail.com;
Washington L. R. de Carvalho-Segundo, IBICT, washingtonsegundo@ibict.br;
Thiago M. R. Dias, CEFET-MG, thiagomagela@gmail.com;

Session Type

- Presentation

Abstract

LA Referencia community aims to provide more and better services to users. One of the key problems is identifying *relationships* and linking metadata from different sources. This presentation briefly describes the last results of a collaborative effort of different Latin American institutions for building a common software platform capable of process metadata from different heterogeneous sources. The case study for this initial phase is the *linkage* of the Brazilian Digital Library of Thesis and Dissertations (BDTD) and the Lattes CV platform, applying an improved trigram based strategy. The preliminary results show a promising path to follow towards reaching a production-grade implementation for all LA Referencia community.

Conference Themes

- Discovery, use, and impact
- Repositories - evolution or revolution?
- Repositories and global knowledge

Keywords

Linking data, String similarity, Harvesting, Metadata enrichment, CRIS systems, BDTD, Lattes Platform, LA Referencia software

Audience

Repository managers, developers, data producers, librarians, research data managers.

Background

The LA Referencia software (LRHarvester) is a platform for metadata harvesting, validation, and transformation (enrichment/curation). The platform is currently installed in ten national nodes in Latin America, and periodically harvests and processes more than 1.7 million metadata records.

The Brazilian Digital Library of Thesis and Dissertations (BDTD) <http://bdtd.ibict.br> is a network of more than a hundred institutions that aggregate more than a half million full-text open access electronic thesis and dissertations. This national repository uses the LRHarvester. Also, the content of

BDTD is harvested by the LA Referencia network via OASISBR <http://oasisbr.ibict.br> (Carvalho-Segundo et al, 2017). According to the NDLTD search portal <http://search.ndltd.org/>, the BDTD is the second most large national repository of the world.

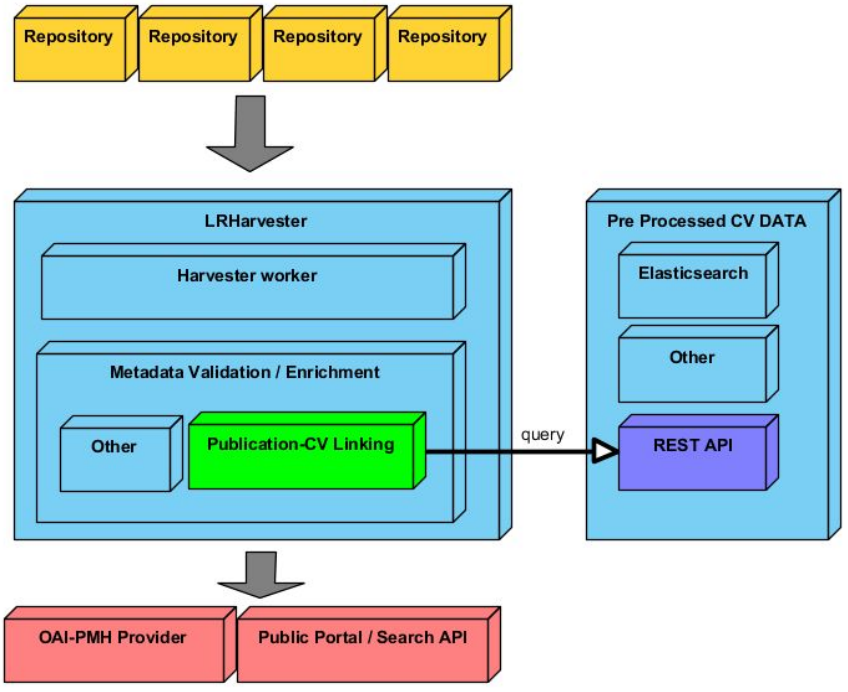
On the other hand, the Lattes CV Platform <http://lattes.cnpq.br/> is a database that has more than 6 million research profiles. The researcher declares in this platform his formation, academic production, participation in conferences and projects, academic awards, etc. In Brasil, having a Lattes profile it is a requirement for submitting a proposal for funding. Moreover, the government agencies have been putting effort in the creation of services of interoperability between ORCID, Lattes CV Platform, open access scientific repositories, and funding platforms.

In Brazil, the records of BDTD have a richer metadata schema than the standard repositories of scientific publications. For instance, authors, advisors, supervisors, and referees are able to attach their profiles IDs via specific fields of the metadata schema. Unfortunately, this task of filling the IDs is done manually, and a small amount of the records have it properly filled. However, researcher IDs are an important step towards the construction of metrics and data analysis over the repositories. Another important aspect is that these linking strategies are a step towards the construction of Current Research Information Systems (CRISs).

The scope of the present work is the implementation of an automatic linking strategy between the records of BDTD and Lattes CV Platform. This strategy is being implemented in the LRHarvester.

Content

LA Referencia harvester platform: the metadata processing model. LRHarvester architecture allows to easily perform metadata transformation rules, which can be integrated into the harvesting, validation, transformation and publication pipeline. For the *linking phase*, an external component was implemented in order to provide a set of preprocessed research profile data. This component is going to be integrated into the platform for the production stage (see figure below).



Automatic publication metadata and CV linking. Since this is a collaborative initiative, one of the main objectives was to provide a common platform so that developers can contribute to different strategies and compare results using the same metadata sets. In this phase, two strategies from different research groups are being integrated. In both cases, the main idea is to use title, authors and other metadata fields to infer the relationships between a repository record and curriculum vitae record. In the following, a trigram-based strategy is presented. Another implementation based on string transformation has been submitted separately in (Dias et al, 2018).

Trigram-based strategy. Trigram string similarity is a method of identifying phrases that have a high probability of being variants of the same original phrase. It is based on representing each phrase by a set of character trigrams. These are used as key terms of the phrase. The similarity of phrases is then computed using the vector cosine similarity measure. ("Trigram Algorithm". NLM). The similarity between the two phrases is computed as the cosine of the angle between them. This is always a number between 0 and 1. When the cosine is roughly 0.7 or greater, the probability that the two phrases are originally the same is very high. The key to this strategy is to calculate the distance for the title and author strings of the records separately. The hypothesis is that the joint probability of two records having high coefficients in the two fields and not being the same record is extremely low. Effectively, the strategy proved to be very accurate, but with impractical computational cost. To compare a base with M records against another with N, *O(MxN)* comparisons are necessary. Comparing millions of records can take weeks of processing.

Elasticsearch trigram indexing and "More Like This Query" heuristic. To solve the scaling problems of the trigrams algorithm in large collections, a previous step was implemented using the Elasticsearch platform. The idea is to use the following heuristic: given the title of a record, to obtain (efficiently) a small list of possible candidates to be compared with the previous trigram-based strategy. This should reduce the computational effort since the number of comparisons between phrases is drastically reduced.

To elect the possible candidates, it is created an index in Elasticsearch for the records of the Lattes CV Platform. This index uses a tokenizer that decomposes each title in trigrams. In this way, the resulting index allows searching for records using the trigrams of a title of the BDTD records. Elasticsearch provides an efficient implementation of a "More Like This" query. This feature is used to obtain a list of similar records using the titles that were indexed by trigram decomposition. In this way, for a given record of BDTD, a list of similar candidates of the Lattes CV Platform is obtained. Thus, a more accurate title and author trigram comparison are made over the candidates. As a result, the method can be used to compare two arbitrary collections of millions of records in a few hours.

BDTD collection and CV Lattes linking results. The Elasticsearch+Trigram-based strategy was applied do the BDTD collection (543,161 metadata records) with Lattes CV Platform thesis collection (1.364.279 records). Additionally, a subset of BDTD collection (87.341 records) that have the ID Lattes assigned was used as a control set and for error calculation. The was executed using .60 as threshold trigram cosine distance for author and title comparisons, and the candidates were selected considering titles with 55% of coincident trigrams for "More Like This" queries. As result 401.723 BDTD records (73,96%) were identified in the Lattes CV Platform. Regarding the control subset, 65.981 (75,54%) were matched with 6 (0,01%) wrong matches (Error Type I: positive match for different IDs Lattes). Regarding Error Type II (not matched with same Lattes ID) 17.085 records (19,56%) were missed. Nevertheless, a careful analysis of the data showed that most cases correspond to titles declared in different languages (English versus Portuguese). The first test of the strategy shows a promising path to follow, considering the very low Error Type I and the relatively high percentage of matching for consistently declared titles.

Conclusion

LA Referencia software can be used as a common collaboration platform to implement different strategies for metadata enrichment and linking, considering the extensible architecture and the collection coverage of most relevant Latin America repositories. The trigram algorithm applied to the BDTD / Lattes CV Platform case study showed a promising path to follow, with use of Elasticsearch MLT to filter candidates, and made the method computationally applicable to medium size collections. Other optimization strategies are being considered to improve computing times.

References

Carvalho-Segundo, W.; Matas, L.; Cabezas, A.; Amaro, B.; Gomes, G.. (2017). *The LA Referencia Software and the Brazilian Portal of Scientific Open Access Publications (oasisbr),* Open Resporitories 2017.

Dias, TMR; Carvalho-Segundo, W; Matas, L. *Using the LattesDataXplorer framework to Automatically linking the Lattes CV Platform to the Brazilian Digital Library of Thesis and Dissertations (BDTD)*.

Available at <http://bit.ly/2Fut7J3>. Accessed in jan/2018.

"Trigram Algorithm", Indexing Initiative (II), U.S. National Library of Medicine (NLM)

https://ii.nlm.nih.gov/MTI/Details/trigram.shtml

TriLite, An Inverted Trigram Index for Accelerated String Matching in Sqlite.
https://github.com/jonasfj/trilite

"More Like This Query", Elasticsearch API Reference,

https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html

DIAS TMR. 2016. Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes. 2016. 181p. (Thesis doctoral). Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte.

Yu, M., Li, G., Deng, D. et al. String Similarity Search and Join: A Survey. Front. Comput. Sci. (2016) 10: 399. https://doi.org/10.1007/s11704-015-5900-5