An ETL Strategy for Integrating the LA Referencia Platform and VIVO for the Brazilian CRIS

Vivian S. Silva[1], Lautaro Matas[2], Tales Moreira[3], Washington C. Segundo[4]

[1] viviansilva@ibict.br, IBICT

[2] lautaro.matas@lareferencia.redclara.net, LA Referencia

[3] talesmoreira@ibict.br, IBICT

[4] washingtonsegundo@ibict.br, IBICT

Introduction

The BrCris project [1] aims at collecting, integrating, and making available a wide range of information regarding the Brazilian research ecosystem. It is conducted by the Brazilian Institute for Information in Science and Technology (IBICT), and has as its main goal to provide the academic community with easily accessible, consolidated data about the national scientific production, through the use of free software and international standards.

To populate BrCris, we gather data from many distinct sources, such as the Lattes Platform, which hosts researchers' academic CVs, Oasisbr, a repository for open access Brazilian publications, and BDTD, the Brazilian Theses and Dissertations Digital Database. Institutions and Journals directories and complementary publication repositories are harvested as well.

For dealing with such an amount of heterogeneous data, we rely on the LA Referencia Platform [2], which provides a suite of tools for managing the data load, deduplication, linking, conversion, and export, so the resulting homogeneous data sets can be explored in different visualization and analysis tools. LA Referencia employs the concept of an Entity-Relation metamodel, where the entities of concern, such as Person, Organization, Publication, Project, etc., their attributes, and the relationships among them are defined as an XML model, which can be updated regularly without affecting the actual relational database physical schema. For BrCris, the Entity-Relationship metamodel is loosely based on CERIF [3].

One of the visualization tools chosen for the exploration of BrCris data is the VIVO Platform, a widely adopted CRIS data navigation environment. In fact, the BrCris semantic model [4], a translation of the Entity-Relationship metamodel aimed at adhering to international standards, was built upon the VIVO Ontology [5], reusing many of its classes and properties. A local extension, identified by the "*brcris*" namespace, was also added to the model, so the particularities of the data collected within the project, and not originally covered by the VIVO Ontology, could be accommodated, avoiding loss of information.

To convert the pre-processed relational data from LA Referencia to an RDF graph ready to be visualized in VIVO, we developed an Extraction, Transformation, and Load (ETL) tool, which populates the BrCris semantic model, integrating seamlessly both platforms.

The LA Referencia ETL Tool for Populating VIVO

The LA Referencia Platform provides a set of tools for processing data coming from a wide variety of repositories, allowing the treatment of heterogeneous data sets for creating a unified, deduplicated database. Since it serves different countries, each one having its own target repositories and entities of interest, the platform allows the definition of the data model as an XML file. This way, although the data is internally stored in a relational database, its schema is transparent to the users, and each country can add or remove entities, attributes, and relationships to/from their own models without any impact on the platform's internal structure.

Once the data is loaded, deduplicated, and consolidated, it can then be exported to other tools for visualization and analysis. For example, it can be indexed in Solr for retrieval through a REST API, or in

ElasticSearch for the construction of dashboards. For BrCris, we chose the VIVO Platform for visualizing the data as a knowledge graph. To make the data available in such a format, we extended LA Referencia export functionalities, assembling an ETL tool to integrate the two environments, comprised of the following components:

1. **Extraction**: the extraction component is shared by all the indexers/exporters, and is responsible for retrieving all the entities by type (e.g., Person, Publication, Patent, etc.) from the relational database, along with their attributes and relationships to other entities. A single entity type, if specified, can be retrieved, or all the available entity types can be extracted at once.
2. **Transformation**: the data retrieved by the extraction component feeds a transformation procedure. This transformation implements a pre-defined mapping between the Entity-Relation metamodel and the ontology-based semantic model. The mapping is also represented as an XML file, making this step fully configurable. For each entity type, all its attributes and relationships are listed, and each one of them is mapped to one or more target triples in the final RDF model. When describing these triples, all the elements necessary to create their object, predicate, and subject, such as their class, namespace, id, etc., are specified. Using the Jena API[1], small batches of entities are processed and their attributes and relationships are converted to RDF triples, which are added to an in-memory RDF submodel, which is, in turn, sent to VIVO and emptied before the next batch is processed.
3. **Load**: once a batch is finished, the resulting submodel is sent to VIVO through its SPARQL Update API[2]. Since this is a standard data ingestion channel, VIVO automatically computes the inferences and rebuilds the index to reflect the changes, making the new data immediately available for navigation.

Upon the definition of the mapping configuration file, the ETL pipeline is triggered by a single line command, finishing with the creation of a complete knowledge graph ready for exploration in VIVO.

Conclusion

The BrCris is a wide-range project aiming at providing a unified view of the Brazilian research ecosystem. It relies on the LA Referencia Platform for processing heterogeneous data coming from diverse sources, and the VIVO Platform for data visualization. We developed an ETL strategy for integrating both platforms, making the data flow from a relation representation to an RDF knowledge graph which follows a VIVO Ontology-based semantic model, leveraging the advantages of both tools while maintaining the data consistency between them, through a single command-line synchronization.

References

[1] Pinto, A.L., Segundo, W.C., Quoniam, L. and Dias, T.M., The Brazilian Current Research Information System: BrCris. Colecção CA–Ciência Aberta. 2021.

[2] Segundo, W.C., Cabezas, A., Matas, L., Amaro, B., and Gomes, G. The LA Referencia Software and the Brazilian Portal of Scientific Open Access Publications (Oasisbr). International Conference on Open Repositories. 2017.

[3] Jörg, B. CERIF: The common European research information format model. Data Science Journal. 2010

[4] Silva, V.S., Moreira, T., Dias, T.M., Gomes, J. and Segundo, W.C., Um Modelo Semântico Baseado em Ontologia para o CRIS Brasileiro. Colecção CA–Ciência Aberta. 2021.

[5] Corson-Rikert, J., Mitchell, S., Lowe, B., Rejack, N., Ding, Y., and Guo, C. The VIVO ontology. Synthesis lectures on semantic web: theory and technology. 2010.

[1] https://jena.apache.org/

[2] https://wiki.lyrasis.org/display/VIVODOC112x/SPARQL+Update+API