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

15th International Conference on Current Research Information Systems (CRIS2022)

Dubrovnik, Croatia, 13th of May of 2022

V. Silva  
(viviansilva@ibict.br)

L. Matas  
(lmatas@gmail.com)

T. Moreira  
(talesmoreira@ibict.br)

W. Segundo  
(washingtonsegundo@ibict.br)
Summary

The Context
Open Science Background
The BrCris Project

The System Architecture

The Semantic Model

The ETL Tool

Conclusions
Main National Partnerships
Main International Partnerships
Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Mexico, Panama, Perú, Spain and Uruguay

Portugal
Brasilian Portal of Publications and Scientific Data in Open Access
**BrCris** - Information System on the Ecosystem of the Brazilian Scientific Research

https://brcris.ibict.br/vivo/
History of the BrCris Project
Laboratory of the Brasilian Scientific Research Ecosystem (LaEPeCBr)  http://dgp.cnpq.br/dgp/espelhogrupo/9750187028652303

https://pnipe.mctic.gov.br/laboratory/3911
BrCris
Ecossistema da pesquisa brasileira

LA Referencia
Red de repositorios de acceso abierto a la ciencia
LA Referencia Support for BrCris

Data collected from many distinct sources...

... loaded and processed into LA Referencia...

... and exported to different BrCris front ends
Mapped entities

1. Person
2. OrgUnit
3. 1. Journal; 2. Event
4. Service
5. Publication (*inclui TDs*)
6. Dataset
7. Course
8. GraduateProgram
9. Patent
10. Project
11. Funding
12. ResearchArea
13. BibliographicReference
14. ResearchGroupMembership
15. ProjectMembership
Integrating LA Referencia and VIVO

Once the data is loaded, deduplicated, and consolidated in LA Referencia:

it is extracted from the relational tables

transformed into a set of RDF triples

and loaded into VIVO
The ETL Tool - Extraction

We can extract a single entity type (e.g. Person, Publication, etc.) from the relational database, or all entity types at once.
The ETL Tool - Transformation

The transformation uses an XML mapping configuration file. Elements from the entity-relation model are mapped to a set of triples...
The ETL Tool - Transformation

... which reflect the VIVO Ontology-based BrCris semantic model
The ETL Tool - Transformation

All the elements necessary to build the triple components (subject, predicate, object), such as their type, namespace, id, etc., are specified in the mapping configuration.

```xml
<source-attribute name="birthDate">
  <target-triples>
    <triple>
      <subject type="foaf:Person" namespace="local" prefix="pers_" idType="UUID"/>
      <predicate type="objectProperty" value="obo:ARG_20000028"/>
      <object type="vcard:Individual" namespace="local" prefix="ind_" idType="UUID"/>
    </triple>
    <triple>
      <subject type="vcard:Individual" namespace="local" prefix="ind_" idType="UUID"/>
      <predicate type="dataProperty" value="vcard:birthdate"/>
      <object type="xsd:dateTime" value="$value"/>
    </triple>
  </target-triples>
</source-attribute>
```

The triples are stored in an in-memory RDF submodel with the aid of the Jena API.
The ETL Tool - Load

Once a batch of entities is processed, the resulting RDF submodel is sent to VIVO through its SPARQL Update API.

VIVO then automatically computes inferences and rebuilds the index to reflect the loaded data.
To sum up

1. Information retrieval systems
To sum up

1. Information retrieval systems

![Image of a magnifying glass over a document]

2. Dashboards

![Image of a dashboard with various charts and graphs]
To sum up

1. Information retrieval systems

2. Dashboards

3. Traditional and semantic APIs
Conclusions

- The process of harmonising heterogeneous sources can be very tedious, but on the whole, mechanisms are mapped it can be performed periodically.
Conclusions

- The process of harmonising heterogeneous sources can be very tedious, but on the whole, mechanisms are mapped it can be performed periodically.

- VIVO is very nice and fast to perform information retrieval and visualisation in research graphs.
Conclusions

▶ The process of harmonising heterogeneous sources can be very tedious, but on the whole, mechanisms are mapped it can be performed periodically.

▶ VIVO is very nice and fast to perform information retrieval and visualisation in research graphs.

▶ Nowadays, we are working to scale the realised tests to big collections.
Conclusions

- The process of harmonising heterogeneous sources can be very tedious, but on the whole, mechanisms are mapped it can be performed periodically.
- VIVO is very nice and fast to perform information retrieval and visualisation in research graphs.
- Nowadays, we are working to scale the realised tests to big collections.
- To give an idea of the dimensions, we have potentially more de 7 Million journal articles and more de 1.5 Million researchers in the Brasilian Research Ecosystem.
Conclusions

- The process of harmonising heterogeneous sources can be very tedious, but on the whole, mechanisms are mapped it can be performed periodically.
- VIVO is very nice and fast to perform information retrieval and visualisation in research graphs.
- Nowadays, we are working to scale the realised tests to big collections.
- To give an idea of the dimensions, we have potentially more de 7 Million journal articles and more de 1.5 Million researchers in the Brasilian Research Ecosystem.
Hvala vam!

viviansilva@ibict.br
lmatas@gmail.com
talesmoreira@ibict.br
washingtonsegundo@ibict.br