Addressing search in scientific open data repositories: A semantic metasearch platform

Gustavo Caetano Borges[1], **Julio Cesar dos Reis**[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing, University of Campinas (UNICAMP)

borges.gustavo.comp@gmail.com, {jreis,cmbm}@ic.unicamp.br

Abstract. *Scientific research in all fields has advanced in complexity and in the amount of data generated. The heterogeneity of data repositories, data meaning and their metadata standards makes this problem even more significant. In spite of several proposals to find and retrieve research data from public repositories, there is still need for more comprehensive retrieval solutions. In this article, we specify and develop a mechanism to search for scientific data that takes advantage of metadata records and semantic methods. We present the conception of our architecture and how we have implemented it in a use case in agriculture.*

1. Introduction

Open Science is a growing movement that preconizes that science should advance through collaboration regardless of geographic, political or temporal constraints. This collaboration is enabled by publishing, in open institutional repositories, digital objects associated with a research project, such as publications, data, software, and all associated documentation. This investigation focuses on aspects concerning the *search for scientific data* in such repositories.

Openness of scientific data enables research reproducibility, auditing, and transparency. This entails savings in project costs, through reuse of openly published data. For these reasons, several funding agencies require that all data produced by projects they fund be made publicly available. For instance, in Brazil, FAPESP's open data policy, part of its open science policy, indicates that "outputs of the research financed by the Foundation are a public good and must be made public as soon as possible, while respecting the principles of scientific ethics, privacy and security, as well as protection of intellectual property."[1]

Whereas the principle of collaboration through data is part of good scientific practices, the implementation aspects of data sharing presents countless challenges. Such challenges range from human aspects (*e.g.*, researchers' resistance to opening "their" data), to e-infrastructure issues (such as appropriately managed repositories), and countless other issues related to, *e.g.*, domain-specific requirements, curation, pseudonymization of sensitive data, or compliance with standards.

In particular, FAIR principles for data sharing and reuse [Wilkinson et al. 2016] present a set of requirements for data to be Findable, Accessible, Interoperable and Reusable to comply with the open science movement. FAIR-ness demands, among others, that data be appropriately documented via metadata standards and stored in repositories

[1]http://www.fapesp.br/openscience/en

that follow good data management practices (*e.g.*, such as certification[2]). The number of internationally recognized repositories is representative of the difficulties for searching because different repositories publish data files using distinct processes. The global platform for registry of research data (RE3Data) indexes approximately 2500 repositories in the most diversified search fields[3]. Each repository may store information about a specific research field or be a multifield (generalist) repository. A given domain may adopt many consensual metadata standards (cf. the RDA directory of metadata standards[4]).

Many search mechanisms rely on metadata [Gottardi et al. 2020]. This requires finding the correct metadata elements and their contents, which depends on knowing the standard used. In addition, it is necessary to know how one standard maps to others (so that the appropriate field is searched for). Mappings among standards have been defined by research groups (*e.g.*, mappings between ABCD and Darwin core, two among many biodiversity standards[5]). However, most mappings require manual correspondence among standards, which is an arduous task. Even when researchers document their data using the same metadata standard, there is heterogeneity of values stored in each element. Each research field has its way to reference things, naming an item with several names. Due to this, homogenization, although desirable, can be impracticable.

In searching for data, researchers need to access all repositories related to the research field and check for related files. This may be simplified in some cases when sets of repositories offer a single metadata interface (*e.g.*, in the network of research data repositories of the State of Sao Paulo[6]).

This research aims to alleviate this burden, by designing and developing a search engine for scientific datasets that accommodates several metadata standards. In our approach, we explore the use of domain ontologies to support semantic search. Our solution is based on a multi-step process that involves: (1) harvest metadata records from files published by multiple scientific repositories; (2) map each record's metadata structure to a basic metadata template that we designed; (3) perform semantic search against these converted records, ranking the results. We name this process *semantic metasearch*.

The remaining of this article is organized as follows. Section 2 presents background literature. Section 3 reports on our proposal. Section 4 presents the development aspects of our platform and preliminary results in a case study. Section 5 presents conclusions and ongoing work.

2. Background

According to [Breeding 2005] "metasearch is the ability to search multiple resources simultaneously". Our work involves metasearch on sets of metadata records harvested from multiple repositories. It combines *mapping among metadata standards*, and *semantically processing harvested records* – our *semantic metasearch*.

Metadata simultaneously serve to document data and facilitate the search process [Simionato 2017]. Indeed, the work of Kaiser *et al.* [Kaiser et al. 2020], written in the

[2]https://www.coretrustseal.org

[3]https://www.re3data.org/metrics

[4]https://rd-alliance.github.io/metadata-directory/

[5]https://www.bgbm.org/TDWG/CODATA/Schema/Mappings/DwCAndExtensions.htm

[6]https://metabuscador.uspdigital.usp.br/

context of supporting COVID-19 research, calls metadata "research accelerant". Its emphasis is on how metadata can support discovery of scientific literature, datasets, and developing policies.

Pierre and Laplant [Pierre and LaPlant 1998] defined a metadata standard as a way to specify in items a set of elements, attributing meaning to each element. Research such as that of Costa and Braga [Costa and Braga 2016] and Sanchez, Silva and Vechiato [Sanchez et al. 2019] analyzed the usage of metadata standards in scientific data repositories. These investigations highlighted that the most frequently used generic standards are *Dublin Core*[7], *Data Documentation Initiative*[8] (DDI) and *ISO 19115*[9].

A *crosswalk* defines the process of mapping a metadata standard to another. [Pierre and LaPlant 1998] highlighted how challenging and error-prone this process is, requiring domain experts with in-depth knowledge. Crosswalks can be manual or semi-automated. For instance, Yan *et al.* [Yan et al. 2013] presented a tool that uses a web service to transform geographic data in a given standard to another standard. An example of manual crosswalk appears in [do Espírito Santo et al. 2019] to support multi-database queries.

The COVID-19 pandemic brought about the urgent need for data sharing[10]. This prompted research geared to the coronavirus (rather than generic search mechanisms). For instance, Izquierdo *et al.* [Izquierdo et al. 2020] proposed a platform to search COVID-19 data. Their platform receives natural language queries, extracts keywords, and applies the search over the contents of two COVID-19 repositories, the Brazilian NSG *(Notificações de Síndrome Gripal)* and the one provided by Johns Hopkins University.

Semantic annotation covers two concepts: The act of annotating; and the annotation itself: a tuple $< o, a >$, where o is the object being annotated and a is the annotation. Semantic annotations can help refining data retrieval because they support extending queries to identify data that is relevant to the query, but which is described with different terms. Ávila *et al.* [Ávila et al. 2017] explored the semantic linkage using the SKOS predicate. Their research proposed semantic enrichment using SPARQL queries and SKOS vocabulary. Using the Schlumberger Oilfield Glossary, they annotated terms and presented links between terms. Gavankar [Gavankar et al. 2020] compared different semantic search systems, namely Swoogle, BioPortal, Watson, Falcons, Hakia, Lexxe, SenseBot, and DuckDuckGo. Their comparison analyzed the search methodology used in each system, their resources, working logic, pros, and cons. Search methodologies included metasearch or even RDF indexing, resources as REST interface, and others.

Rather than annotating data, we annotate metadata. Few investigations consider annotating scientific metadata. An example of a semantic annotator of (biomedical) data was proposed by Jonquet *et al.* [Jonquet et al. 2009]. Their workflow consists of basically two steps: the user provides a text entry, and the tool processes it together with a dictionary (UMLS and NCBO ontology).

[7]https://dublincore.org/specifications/dublin-core/

[8]https://ddialliance.org/Specification/DDI-Codebook/2.5/

[9]https://www.iso.org/standard/53798.html

[10]http://www.oecd.org/coronavirus/en/data-insights/international-scientific-collaboration-on-covid-19-medical-research

3. A platform for semantic metasearch for scientific data

This section presents our proposal of a platform to help search for data from multiple public research repositories (see Figure 1). Our solution is based on semantic metasearch, in which we first map metadata records into a template we developed, and then perform the metasearch against these records. The template uses the classification of metadata fields of [Riley 2017] and contains the following elements: 1) Author; 2) Date; 3) Description; 4) URI; 5) Language; 6) Rights; 7) Source; 8) Subject; 9) Title; 10) Type; 11) Ad_descriptive; 12) Ad_administrative; 13) Ad_structural; and 14) Ad_markup – where elements 11 through 14 represent additional descriptive, structural, administrative and markup elements. The template was created by this paper's first author, based on a survey of major scientific metadata standards, which included the work of [Simionato 2017]. This Metadata template was designed to address the problem of heterogeneity of metadata standards across repositories.

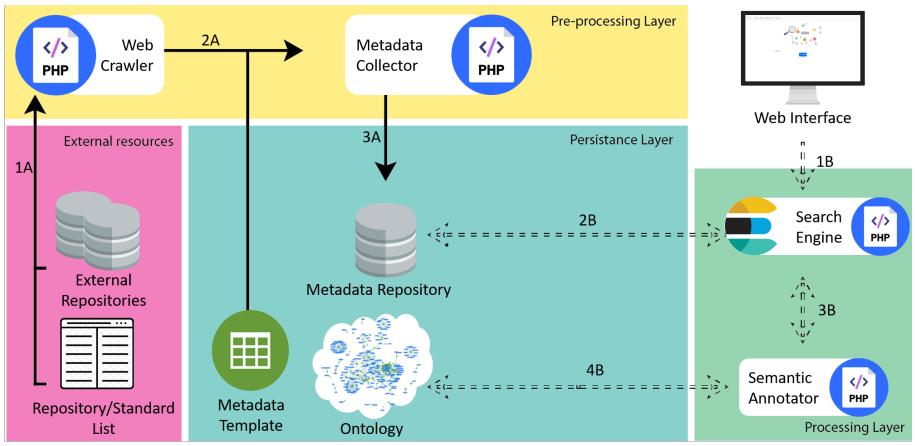

Figure 1. Architecture of the Scientific Data Semantic Metasearch Platform.

Figure 1 presents our semantic metasearch architecture. Metadata records are retrieved from external data repositories (labelled 1A in Figure 1) by a web crawler. The "Metadata Collector" rewrites each such metadata record into our Metadata template (item 2A in Figure 1) and stores the rewritten record into the "Metadata Repository"(3A in Figure 1). Metasearch is performed against data using queries that are semantically enriched via ontologies.

Queries via the "Search Engine" occur in two ways: (1) based on exact match between query terms and metadata elements in the "Metadata Repository" (cf. 2B in Figure 1); and (2) semantic metasearch in which an input query from the user is extended by the "Semantic Annotator" with domain ontologies (cf. 3B and 4B in Figure 1).

The construction of the "Metadata Repository" involves the following elements (cf. Figure 1):

- **External Repository:** Scientific data repositories external to our system. We assume that each repository uses its own metadata standards.
- **Repository/Standard List:** A document containing URLs of scientific repositories and their respective metadata standards. This can be manually generated or harvested from a webpage such as re3data[11]. Our system depends on this set of

[11]https://www.re3data.org

URLs because it harvests data from external repositories.

- **Metadata template:** Our basic Metadata Template, created to standardize collected metadata records. External metadata records are mapped to this template via a (manual) crosswalk process.This manual crosswalk is performed only once per external repository.
- **Metadata Collector:** Responsible for receiving and extracting the metadata records from a web crawler, transforming them into the Basic Template and storing them in the Metadata Repository.
- **Metadata Repository:** Our internal metadata repository, that stores metadata records rewritten into our basic template.

The search process is applied to the "Metadata Repository" and involves the following elements:

- **Web Interface:** Natural language interface for user queries for files of interest. Users can pose queries either without semantic processing, or expanded (semantic metasearch) queries. This component is represented in Figure 1 as a computer icon.
- **Search engine:** Module responsible for processing search strings and generating a ranked list of metadata records that are returned to the users.
- **Semantic Annotator:** Module responsible for semantically annotating metadata. It is invoked by the Search Engine to process semantic metasearch. To this end, it uses a set of online domain ontologies.
- **Ontologies:** Module that provides several formal knowledge models representing sets of domain concepts and the relationships among them.

4. Development and Evaluation

To check our architecture and implementation, we conducted a case study in the agriculture domain - in which users want to find open data related to specific agriculture research topics. We selected this domain because our group has a long history of projects in this field. Thus, we could count on expert collaborators, and on our previous experience with agriculture data. This corresponds to the first version of the implementation of our architecture (cf. Figure 1).

Given our application domain, our software was tailored to connect to two repositories: the International Institute of Tropical Agriculture (IITA) data repository[12] and a repository of images of plant diseases made available by Embrapa (CNPTIA), whose metadata are exposed via the central node of the network of public research data repositories of the state of Sao Paulo[13]. The IITA repository was selected because it contains a variety of agriculture-related curated data, as well as curated metadata. It contains more than 2500 scientific dataset entries. The Embrapa repository was chosen because of its quality and data types, and by the quality of the metadata records, which are published independently from the original repository using a basic standard format, thereby helping metadata harvesting.

[12]https://data.iita.org/about

[13]https://metabuscador.uspdigital.usp.br

We used two different kinds of metadata harvesting. Metadata from Embrapa was harvested directly from the central metadata repository of the network of public repositories of the state of Sao Paulo[14]. For IITA, we developed an algorithm to harvest metadata which sends a REST request to its data server. IITA limits harvesting to ten datasets at a time, so this had to be performed in a loop of requests.

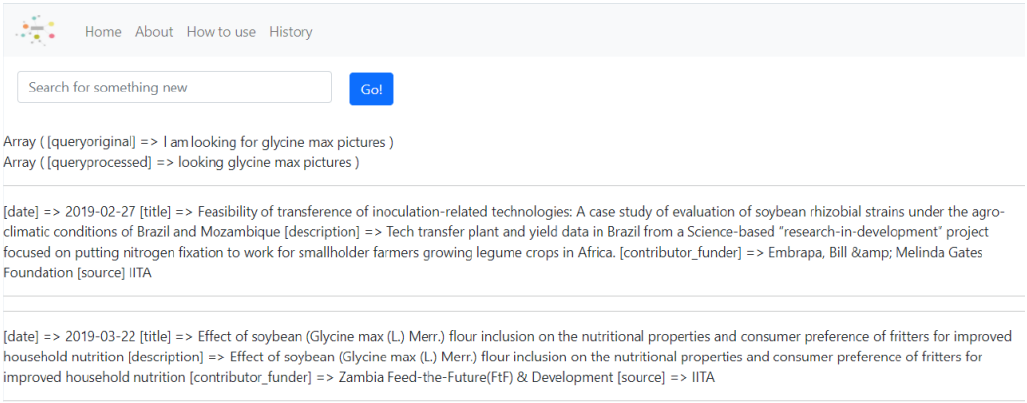

Figure 2. Search results - partial interface.

Preparing the Metadata Repository: We created the Metadata Repository (cf. Figure 1) for the case study as follows. First, a web crawler accessed Embrapa and IITA to collect metadata records, which were delivered to the Metadata Collector. Then, these collected metadata records were rewritten to our metadata template and stored in the Metadata Repository. Once this was done, the user could submit queries via the Web Interface.

User query: On the user's side, metasearch was implemented as follows. The user poses a query in natural language for example "pictures of diseases in soybeans", and receives as result a list of metadata records, each of which pointing to a different scientific data file (that can be in either the IITA or the Embrapa repositories). The user can either request an "exact match" query, which will not require the use of ontologies, or "extended query", which will require ontological processing.

Metasearch processing: Metasearch was implemented as follows. The Search Engine receives the user natural language query, removes stopwords and generates a set of keywords. For an "exact match" request, these keywords are input to the *ElasticSearch* engine[15]. For an "extended query", the keywords are first forwarded to Agroportal for semantic processing, and the result is then input to *ElasticSearch*.

AgroPortal[16] in our present implementation plays the role of the "Semantic Annotator" of our architecture. It is a platform to identify, host and use vocabularies and ontologies in agro-informatics applications and is widely used in implementations involving semantic processing in the agriculture domain.

[14]https://metabuscador.uspdigital.usp.br
[15]https://www.elastic.co/pt/what-is/elasticsearch
[16]http://agroportal.lirmm.fr/about

In more detail, in extended queries the keywords are forwarded to Agroportal; the latter, in turn, returns an expanded set of keywords, based on ontological relations. This expanded set of keywords is used to construct a set of queries that are sent to the *ElasticSearch* engine. Our Search Engine invokes *ElasticSearch* via the *ClientBuilder*. *ClientBuilder* requests are built using several types of queries, such as a typical client connect or REST API. Our solution connects to the *ElasticSearch* framework using a standard connection function from PHP.

Stopword removal used the NLTK package[17], introduced to improve query result. Elasticsearch does not differentiate between query terms and stopwords. For example, consider the query "*I am looking for glycine max pictures*". Before the stopword removal procedure, 2377 metadata records were retrieved. After stopword removal, only 10 metadata records were returned. Figure 2 presents a partial screen copy of results for this query. The figure shows that results provide metadata records (under our basic template) containing date, title, description, and source of the original metadata information.

5. Conclusions and ongoing work

Metasearch still requires further studies to exploit the full possibilities of ontological aspects for semantically enabled search engines. We proposed a semantic metasearch software architecture for retrieving scientific data from open repositories. To the best of our knowledge, our solution is the first proposal that combines standard metadata harmonized with ontological processing in a generic and extensible architecture.

We exemplified the applicability of our solution via a real-world case in the agriculture domain. In this case, data (and metadata) are published in one of the largest databases on agricultural data from the African continent, and on the Embrapa repository. This allowed us to identify key challenges faced in metasearch – such as the intrinsic heterogeneity of metadata, even within a single database. Our solution requires addressing dependencies on non-consensual vocabularies and ontologies. While some of these challenges are specific to our implementation environment, others are generic - such as having to cope with different repositories and metadata standards.

Our ongoing work involves both research and development activities. On the latter side, we aim to improve the way results are shown to end users, and continue our development efforts to include additional ontologies and vocabularies. Researchers in agricultural sciences are helping us to express requirements and validate results.

Further research needs to be conducted on, among others, metadata standards, and queries that take versioning of data into account. In this sense, search results to a query would be "time series" of data. This last extension is much harder to design (and implement) because it requires deciding on which kind of metadata timestamp to consider - such as "deposit time", "creation time", etc.

Acknowledgements Work partially supported by the São Paulo Research Foundation (FAPESP) (#2013/08293-7, #2017/02325-5), and CNPq (#428459/2018-8 and #305110/2016-0).

[17] https://nltk.org

References

Ávila, R., Santos, S., Araújo, D., Vidal, V. M. P., and de Macêdo, J. A. F. (2017). Ligações Semânticas Utilizando Predicados SKOS. In *SBBD*, pages 88–99.

Breeding, M. (2005). Plotting a new course for metasearch. *Computers in Libraries*, 25(2):27–29.

Costa, M. and Braga, T. (2016). Repositórios de dados de pesquisa no mundo. *Cadernos BAD*, 0(2):80–95.

do Espírito Santo, J., de Paula, E. V., and Medeiros, C. B. (2019). Exploring Semantics in Clinical Data Interoperability. In *Advances in Conceptual Modeling*, pages 201–210. Springer International Publishing.

Gavankar, C., Bhosale, T., Gunda, D., Chavan, A., and Hassan, S. (2020). A comparative study of semantic search systems. *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, pages 1–7.

Gottardi, T., Medeiros, C. B., and Reis, J. D. (2020). Understanding semantic search on scientific repositories: Steps towards meaningful findability. In *1st virtual workshop on Research data management for Linked Open Science-DaMaLOS*.

Izquierdo, Y. T., Garcia, G. M., Lemos, M., Novello, A., Novelli, B., Damasceno, C., Leme, L. A., and Casanova, M. A. (2020). Keyword Search over the COVID-19 Data. In *Anais XXXV SBBD*, pages 205–210, Porto Alegre, RS, Brasil. SBC.

Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56–60.

Kaiser, K. A., Chodacki, J., Habermann, T., Kemp, J., Paglione, L., Urberg, M., and Scott Plutchak, T. (2020). Metadata: The accelerant we need. *Information Services & Use*, (Preprint):1–11.

Pierre, M. S. and LaPlant, W. P. J. (1998). Issues in Crosswalking Content Metadata Standards. *National Information Standards Organization - White Papers*.

Riley, J. (2017). Understanding metadata. *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)*, 23.

Sanchez, F. A., Da Silva, N. B. P., and Vechiato, F. L. (2019). Padrões de metadados para representação e organização da informação em repositórios de dados de pesquisa. *Informação & Tecnologia*, 5(1):37–51.

Simionato, A. C. (2017). Mapeamento dos metadados para dados científicos. In *XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XVIII ENANCIB)*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., and et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Yan, Q., McMahon, M. J., Dascalu, S., Harris, F. C., and Ravi, L. (2013). Community metadata ISO 19115 adaptor. *28th International Conference on Computers and Their Applications 2013, CATA 2013*, pages 213–218.