Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 29, Num. 01 (2022) 91-101

RESEARCH ARTICLE

SSM: A Semantic Metasearch Platform for Scientific Data retrieval

Scientific Semantic Metasearch (SSM): Uma Plataforma para Recuperação de Dados Científicos

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

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 the agriculture domain.
Keywords: Metadata — Query expansion — ontology — agriculture

Resumo: Pesquisas científicas em todos os campos têm avançado em complexidade e na quantidade de dados gerados. A heterogeneidade dos repositórios de dados, significado dos dados e seus padrões de metadados tornam esse problema ainda mais significante. Apesar das diversas propostas para se recuperar dados de pesquisas em repositórios públicos, ainda há a necessidade de soluções de recuperação de informação mais compreensivas. Neste artigo, especificamos e desenvolvemos um mecanismo para busca de dados científicos que aproveita os registros de metadados e métodos semânticos. Apresentamos a concepção da nossa arquitetura, sua implementação e avaliação em um caso de uso no domínio da agricultura.
Palavras-Chave: Metadados — expansão de consulta — ontologias — agricultura

[1] *Institute of Computing, University of Campinas (UNICAMP), Campinas - São Paulo, Brazil*
*__Corresponding author__: borges.gustavo.comp@gmail.com
DOI: http://dx.doi.org/10.22456/2175-2745.119164 • **Received:** 07/10/2021 • **Accepted:** 03/01/2022

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 [1] 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 that follow good data management practices (*e.g.*, such as certification[2]). The number of internationally recognized repositories is representative regarding the

[1] http://www.fapesp.br/openscience/en
[2] https://www.coretrustseal.org

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 [2]. 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, which diverge semantically. Due to this, homogenization, although desirable, can be impracticable.

An analogy can be drawn to Google Search, in which a search string returns a list of links that have to be checked individually. 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 as *semantic metasearch*.

This paper extends a short paper of ours [3] introducing our SSM platform. In the current contribution, we considerably extend that paper, including algorithms designed and implemented in our SSM platform, detailing its implementation, data flow, and providing evaluation scenarios in the agriculture domain.

The remaining of this article is organized as follows: Section 2 presents the background by including fundamental concepts and related work. Section 3 describes our proposed platform. Section 4 reports on a case study to present evaluation scenarios. Section 5 discusses our findings whereas Section 6 concludes this paper.

2. Background

This section provides key definitions and presents a summary of related work.

2.1 Fundamental concepts

Metasearch. According to Breeding [4], "metasearch is the ability to search multiple resources simultaneously". Our work involves metasearch on sets of metadata records harvested from multiple repositories. Our solution combines *mapping among metadata standards* and *semantically processes harvested records* – our *semantic metasearch*.

Metadata. Metadata simultaneously serve to document data and facilitate the search process [5]. Indeed, the work of Kaiser *et al.* [6], written in the context of supporting COVID-19 research, calls metadata as "research accelerant". Its emphasis is on how metadata can support discovery of scientific literature, datasets, and developing policies.

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

Crosswalks. A *crosswalk* defines the process of mapping a metadata standard to another. Pierre and Laplant [7] 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.* [10] presented a software tool that uses a web service to transform geographic data in a given standard to another standard. Santo *et al.* [11] described an example of manual crosswalk to support multi-database queries.

Semantic annotations. A *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.

2.2 Related work

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). An example of COVID-directed data research is the work of Izquierdo *et al.* [12], who proposed a platform to search COVID-19

[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/

[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

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 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.* [13] 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 [14] 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.* [15]. Their workflow consists of basically two steps: the user provides a text entry, and their software tool processes it together with a dictionary (UMLS and NCBO ontology).

3. A platform for semantic metasearch for scientific data

We present our SSM platform based on a top-down perspective. To this end,

Figure 1 presents the architecture of SSM; Figure 4 shows the data flow across the designed modules (that embed the corresponding algorithms). Figure 3 revisits Figure 1 showing the technologies chosen to implement each module.

3.1 Conceptual view

We present our proposal of SSM – a platform to help search for data from multiple public research repositories (cf. 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.

Our metadata template is based on the classification of metadata fields of [16] 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 [5]. This Metadata template was designed to address the problem of heterogeneity of metadata standards across repositories.

Figure 1 presents the SSM, our semantic metasearch architecture. Metadata records are retrieved from external data repositories (1A in Figure 1) by a web crawler.

The "Metadata Collector" rewrites each metadata record into our Metadata template (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" (2B in Figure 1) occur in two ways: (1) based on exact match between query terms and metadata elements in the "Metadata Repository"; and (2) semantic metasearch in which an input query from the user is extended by the "Semantic Annotator" (3B in Figure 1) based on domain ontologies (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, which applies to all data files within the repository. All these standards are subsequently mapped into our template.
- **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 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 template.

From a high level point of view, the search process is applied to the "Metadata Repository" and involves the following elements:

- **Web Interface:** Natural language interface for user queries for scientific data files of interest.
- **Search engine:** Module responsible for processing search strings and generating a ranked list of metadata records that are returned to the users. Queries are processed via a semantic processing and expanded (semantic metasearch).

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

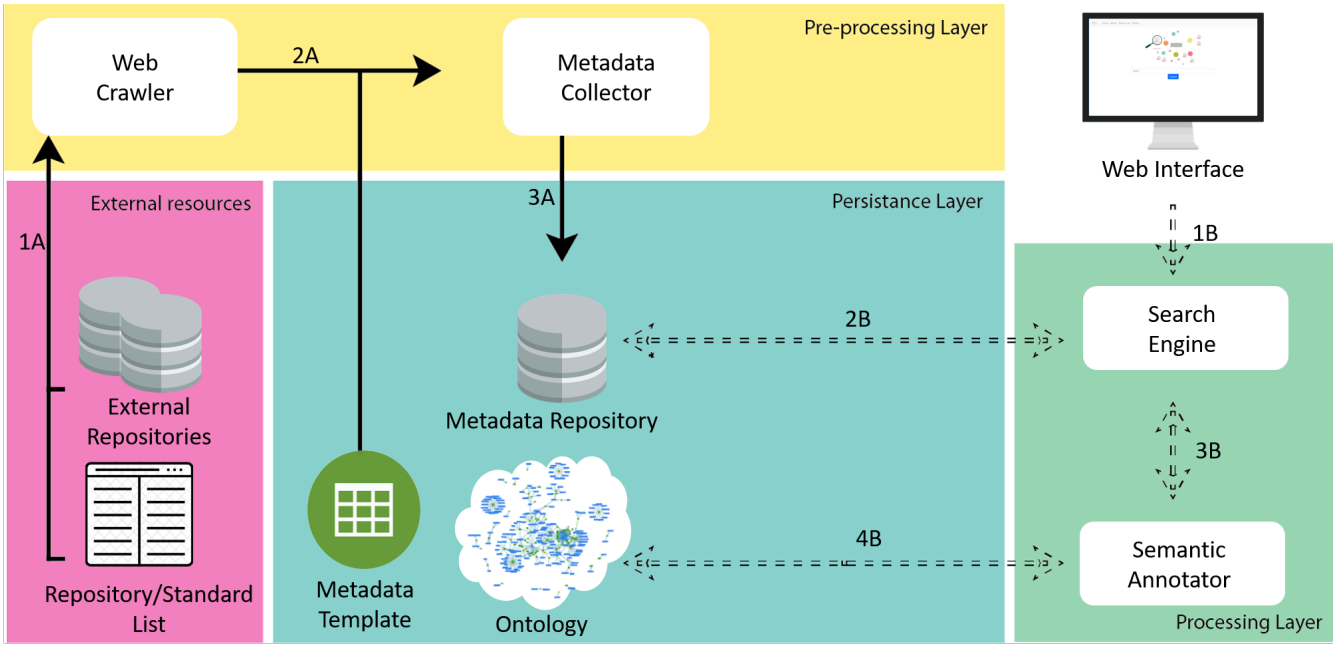

Figure 1. Architecture for SSM - modules and their interactions.

- **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 the access to formal knowledge models representing sets of domain concepts and the relationships among them.

3.2 Processing query strings

Figure 2 presents the execution flow for the search process – with and without ontologies. Blue boxes in Figure 2 represents the steps towards processing of the user-provided query strings from input to two kinds of outputs - without semantic processing (represented by the green flow) and with semantic processing (orange flow).

First, the user's input string is preprocessed (A) by eliminating stopwords. This stage produces a set of key terms that are forwarded either directly to the Search process (B), or go through semantic preprocessing (C, D in Figure 2). The Search process itself is detailed in the lower part of Figure 2.

Semantic preprocessing (C and D in Figure 2) results in two lists of terms ($L1$ and $L2$), which are input to the Search process. Each list contains terms that are either identical or ontologically related to the input terms: $L1$ contains those that are considered the closest; $L2$ contains those that are the most distant in a semantic way.

In more detail, process (C in Figure 2) retrieves all terms ontologically related to each input term. Process (D in Figure 2) builds the lists. As an example, if the terms provided by the user are t_1, t_2 and t_3, then $(L1 = T_1, T_2, T_3)$ where T_i is either t_i or the closest term to t_i in the set of ontologies processed. By the same token, $(L2 = F_1, F_2, F_3)$, where F_i is either t_i or the most distant term from t_i in the set of ontologies.

The lower part in Figure 2 shows how the terms are processed in the search process. The input is a set of terms and the output is a set of metadata records which contains data that either are exactly matches of the input terms, or are semantically similar to these terms.

First, key terms that are input to the Search are processed to check their similarity with the contents of the metadata records in the Metadata Repository (B1 in Figure 2). Next, the metadata records that meet the similarity check are ranked as to their similarity with the set of input terms (B2 in Figure 2). Duplicates in the set of records are discarded (B3 in Figure 2) and the final result of ranked metadata records is returned to the user.

3.3 Data flow in SSM

Figure 3 shows the data flow in our SSM platform.

- **1A in Figure 3** External metadata records are retrieved from repositories formatted according to any metadata standard;

- **2A in Figure 3** The Web Crawler forwards the external records to the Metadata Collector to be processed and mapped to our basic Metadata Template;

- **3A in Figure 3** External metadata records are mapped into our template and stored in our Metadata Repository;

Steps [1A in Figure 3] through [3A in Figure 3] are required to create (or update) the Metadata Repository. From

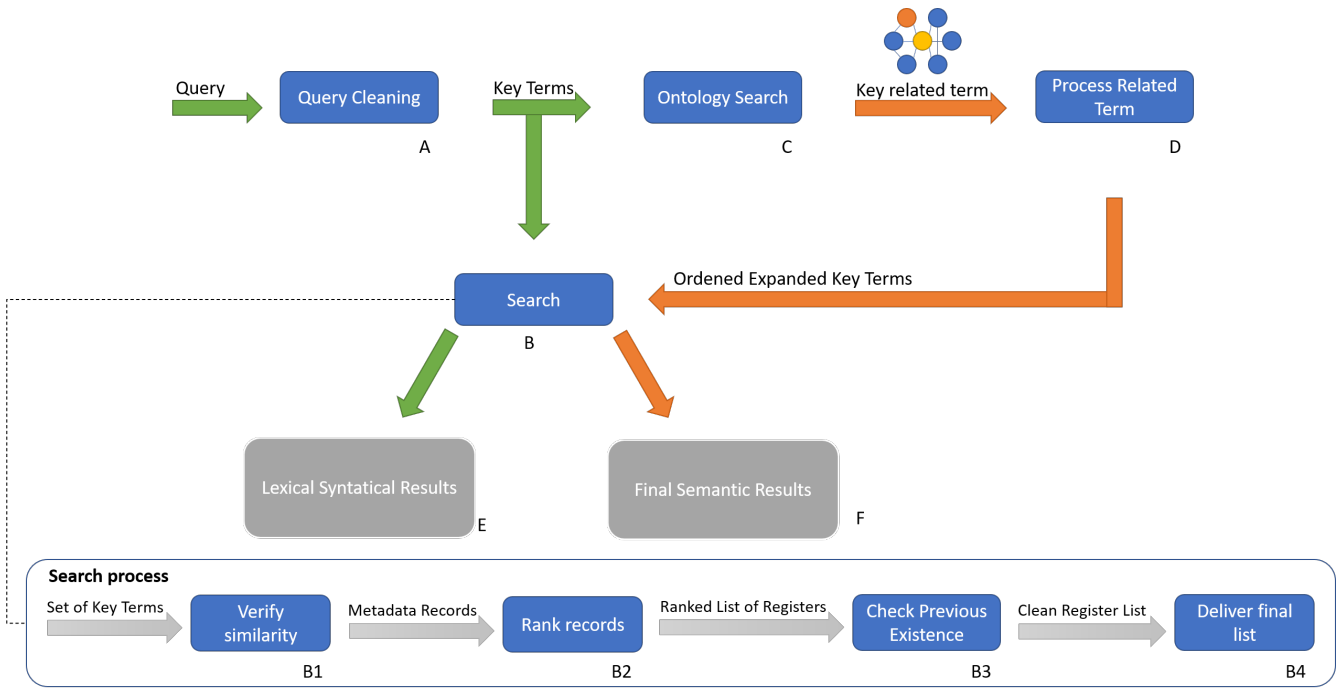

Figure 2. Search process (with and without semantics).

then on, the user can start querying the system. At this stage, data flow occurs as the following:

- **1B in Figure 3** The user provides a search string to the Search Engine;
- **2B in Figure 3** The Search Engine processes the input string according to two flows – the green and/or the orange ones in Figure 2. In the green flow, the terms are directed to the Metadata Repository for similarity search (as in the lower part of Figure 2), and the resulting records are returned to the user. In the orange flow, the records are processed by the Semantic Annotator, and then directed via the Search Engine to the Metadata Repository, where after processing, they return a set of semantically ranked metadata records.
- **3B in Figure 3** and **4B in Figure 3** represent the alternative path for semantic processing (orange flow). This occurs after the corresponding key terms are processed against the Metadata Repository and returned.

3.4 Implementation aspects

This section presents the algorithms designed in the SSM platform and the corresponding software modules. SSM was conceived with scalability concerns, both to facilitate addition of new functionality or ontologies and to improve performance issues.

Figure 4 illustrates the technologies used in the implementation of SSM modules. There are two kinds of modules – those that perform web crawling and create the Metadata Repository; and those to process user queries once the Repository is created. We concentrate on the latter. The two most important for query processing are:

- **Record retrieval**: Responsible for the communication between the Search Engine, the Metadata Repository, and the User Interface.
- **Annotator**: Implements the semantic processing described in Section 3.2.

Algorithm 1 performs an expansion of the original input terms via ontology processing. For instance, if the user inputs term “coffee”, the algorithm via an ontology query returns the concept “Coffea arabica” (one of the species of coffee plants). This is one of the terms that may expand the user query.

The main steps of Algorithm 1 are:

- *input*: user input terms and a set of ontology URLs;
- Check the input ontologies for terms related to the input ones;
- *output* Sets of lists where each original term is followed by additional related terms. Figure 5 presents an example of the output. The example explores ontologies available in AgroPortal[12]. Figure 5 illustrates the step-by-step processing of term “Soybeans” resulting in 5 concepts in the final query expansion.

[12]http://agroportal.lirmm.fr

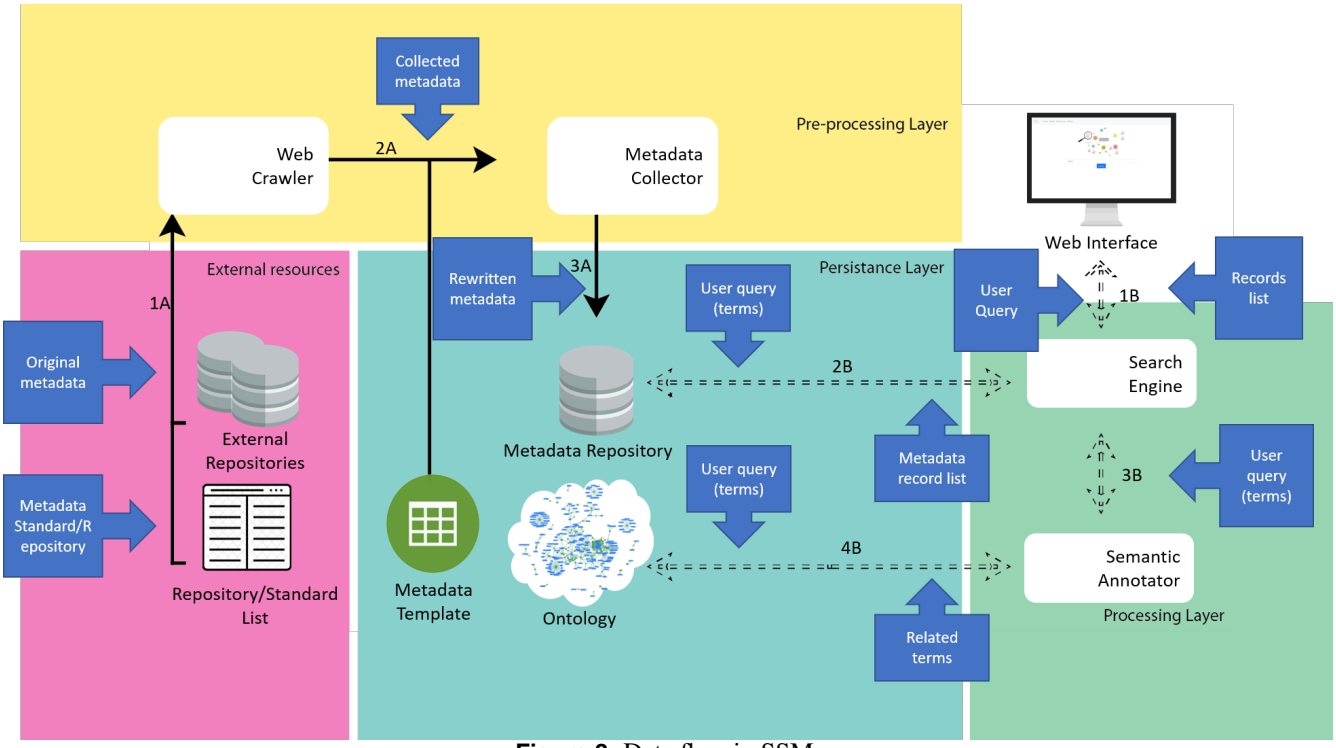

Figure 3. Data flow in SSM.

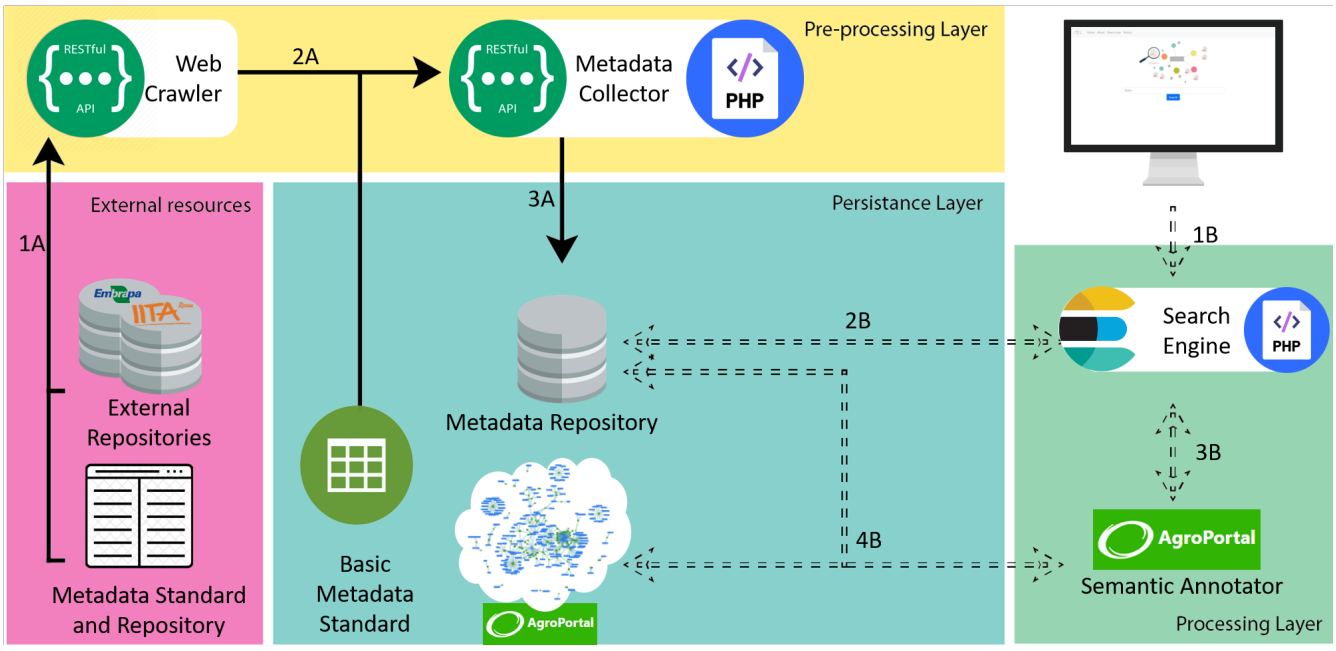

Figure 4. Implementation view of the SSM platform.

```
Glycine maxArray
(
    [0] => soybeans
)
soya beansArray
(
    [0] => soybeans
)
soybeansArray
(
    [0] => soy beans
    [1] => soyabeans
)
Glycine maxArray
(
    [0] => soybeans
    [1] => soybean
    [2] => Glycine max; cv. Wye
)
Glycine maxArray
(
    [0] => soya bean
    [1] => soybean
    [2] => soybeans
    [3] => glycine max
    [4] => Glycine max (L.) Merr.
)
```

Figure 5. Output example from Algorithm 1 where the input is the concept "*soybeans*". Query expasion performed from an agriculture ontology.

Algorithm 1: Query Expansion Algorithm - Expands user-provided terms to those found in ontologies

input : $Queryterms$
output : $annotated[label]$

$url \leftarrow$ "$receive\_query\_url\_for\_ontology$"
$json \leftarrow$ "$url\_contents$"
foreach $entry \in json$ **do**
  if $synonym \in json$ **then**
    $label \leftarrow$ "$json[n](label)$"
    $synonym \leftarrow$ "$json[n](synonym)$"
    $annotated[label] \leftarrow$ "$synonym$"
  else
    $return$ "0"
  end
end
return $annotated[label]$

The "record retrieval" module checks whether a term exists in any metadata record in the Metadata Repository, returning the records found. Algorithm 2 presents how this is processed.

Algorithm 2: Algorithm that checks Metadata Repository records against query terms and their expanded terms.

input : $QUERY$
output : $annotated[label]$

if $QUERY$!= $\emptyset$ **then**
  $query \leftarrow$ "$QUERY$"
  $response \leftarrow$ "$client->Search(query)$"
  if $QUERY$!= 0 **then**
    while $response$ **do**
      $return\ response$
      $response \leftarrow$ "$client->response+1$"
    end
  else
    $return\ \emptyset$;
  end
else
  $return\ \emptyset$;
end
return $response[n]$

Algorithm 2 for metadata search works as follows:

- *Input:* Terms - original ones and semantically related ones;
- Validation verifies if user input is empty or contains only stopwords;
- Checks terms against Metadata Repository records;

- Iterate over all relevant metadata records to create the output list;
- *Output:* list of metadata records found;

4. Evaluation

4.1 Instantiating SSM for an use case in Agriculture

We conducted a case study in the agriculture domain to evaluate our proposed architecture and implementation. In our case, 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 agricultural data. This corresponds to the first version of the implementation of our architecture (cf. Figure 1), using the technologies as presented in Figure 4.

Given our application domain, our software was tailored to connect to two repositories: the International Institute of Tropical Agriculture (IITA) data repository[13] 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 São Paulo[14]. 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.

We used two different metadata harvesting. Metadata from EMBRAPA was harvested directly from the central metadata repository of the network of public repositories of the state of São Paulo[15]. 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.

Preparing the Metadata Repository. We created the Metadata Repository (cf. Figure 1) for the case study as follows. First, our 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 our Web Interface system.

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 a 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). In the querying process, the system returns results from an "exact match" query. In the conception of our platform, if the user judges inadequate, (s)he can demand for the system returning from the semantic query.

Metasearch processing. Metasearch processing 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, keywords are input to the *ElasticSearch* engine[16]. For an "extended query", keywords are first forwarded to Agroportal for semantic processing, and the expanded result is then input to *ElasticSearch*.

AgroPortal.[17] Its implementation plays the role of the "Semantic Annotator" in the current version of our architecture. AgroPortal is a platform to identify, host and use vocabularies and ontologies in agro-informatics applications. This is widely used in implementations involving semantic processing in the agriculture domain.

In more detail, in extended queries, input keywords are forwarded to Agroportal. The latter, in turn, returns an expanded set of keywords, based on encoded ontology 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[18], 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 in our system, 2377 metadata records were retrieved. After stopword removal implementation, 10 metadata records were returned. Figure 7 presents a partial screen copy of results for this query. This figure shows that results provide metadata records (under our basic template) containing date, title, description, and source of the original metadata information.

4.2 Search scenarios

System evaluation consisted in preparing a few scenarios to check the effects of query pre-processing and query expansion on the amount and quality of data returned for a given query. We discuss about which query parameters better meets user goals.

Table 1. Search scenarios

	Without ontologies	With ontologies
No Prep	Scenario1	Scenario2
Preprocessing	Scenario3	Scenario4

[13] https://data.iita.org/about
[14] https://metabuscador.uspdigital.usp.br
[15] https://metabuscador.uspdigital.usp.br
[16] https://www.elastic.co/pt/what-is/elasticsearch
[17] http://agroportal.lirmm.fr/about
[18] https://nltk.org

Table 1 presents four scenarios implemented in our evaluation. For instance, Scenario 1 corresponds to the most basic functionality in which the initial query is not preprocessed, and there is no semantic expansion.

AgroPortal contains a large set of ontologies. We select those with the highest coverage in terms of vocabulary available.

This enables us to cover a wider spectrum of agriculture research.

We used the entire set of AgroPortal ontologies in our semantic processing because there may be concepts unique to a single ontology. We emphasize that the AgroPortal has as one of its available ontologies the Agrovoc[19], which has over 50 thousand English terms, and their correspondent to other languages.

Our hypothesis in testing the use cases is that Scenario 4 would be the one that would best suit the user requirements. The use of ontologies can help in finding key relevant and semantically correlated terms.

Our scenarios were tested based on two user queries:

1. Photos showing antrachnose (a fungus) in coffee produced in Brazil;

2. Comparing situations of planting manioc in areas with dry climate conditions.

4.3 Evaluation results

We clarify that our implementation performing semantic processing also performs the basic (non-semantic) queries. Our queries were tested in semantic processing scenarios.

In query 1 of evaluation, the use of the AgroPortal retrieved 66 concepts, of which 23 were directly related and the other were ancestor concepts of *anthracnose, coffee, Brazil*. Figure 6 presents the results for these three terms, where the colors indicate term repetition - of concepts found in various ontologies within results, before execution of the module B1 in Figure 2. In this case, the ontology that returned the highest number of results for *Anthracnose* was "Soy Ontology".

The second query retrieved 51 concepts, of which 29 were ancestor concepts of *cassava, comparisons, planting, dry climate*. The *FoodOn* ontology was the one that retrieved the most concepts to the term *cassava*.

Term expansion followed the steps discussed in Figure 2. Our Ontology Search process explored the AgroPortal. For ranking purpose of the results, we used the $Cvalue - h$ score of AgroPortal. This provides relevance ranking for ontology processing, thereby creating the lists $L1$ and $L2$ discussed in Section 3.2. Figure 7 presents for query 1 in our evaluation, the number of metadata records retrieved. It is relevant to note the following:

- the query itself, without stopwords removed - all stopwords are considered valid terms resulted in 2847 metadata records;

- the query constructed from $L1$ and associated terms returned 15 metadata records;

- the query constructed from $L2$ and associated terms returned 73 metadata records;

These points show that the recognition of terms in the ontology and consequently their change is beneficial once the possible relevant records to the user are retrieval. Comparing the amount of records retrieval highlights that this amount was satisfactory compared to the original set.

The expansion of query 1 resulted in "Photos of Reaction to Colletotrichum dematium infection in coffea arabica produced in Brazil". The replaced terms scored as follows: *Reaction to Colletotrichum dematium infection* (3.000); *coffea arabica* (3.000); and *Brazil* (3.322).

Each term obtained this score once it relates to the original concept and has related meanings. For example, the concept "Reaction to Colletotrichum dematium infection" that is a synonym to the "anthracnosis" concept, exists in the "Soy Ontology" ontology. Also exists a distance of 1 level to the original concept and the missing grammatical similarity. These items, when combined, compose the score of the related concept.

The score of each term used in the new query is obtained by its relationship in the ontology with the original term. It impact if the term is a more generic, specific, synonymous or other term. This means that the score can vary. In our example, queries ranged from 3.000 to 9.187. Once this score is obtained, it is summed with all terms to obtain a score for that combination of terms.

Analogously, for $L2$, considering ancestor terms, the sentence was transformed into "Photos of Reaction to glomerella infection in coffea produced in South America" with the following weights: Reaction to glomerella infection (9.187); coffea (9.187); and South America (3.200).

The query 2 in our evaluation, the number of metadata records retrieved was the following:

1. Without stopword removal – 2478 records;

2. with stopword removal - 101 records for concepts directly connected; and 2141 using ancestor terms;

The query 2 was transformed into "Comparisons of planting manihot esculenta in arid climate locations", where term replacement with the top ranks were: manihot esculenta (8.000); and arid climate (6.000). If, additionally, ancestor terms are adopted, the query becomes "Comparisons of substrate cassava food product in dry climate locations", with weights: substrate (3.200); and cassava food product (9.187).

5. Discussion

This research proposed a platform for metadata scientific search. We presented a solution for heterogeneity problems in scientific data retrieval querying records from different scientific data repositories. The present study refers to an extension

[19] http://agroportal.lirmm.fr/ontologies/AGROVOC/

Photos of **anthracnose** in **coffee** produced in **Brazil**

ANTHRACNOSE=

anthracnosis, Reaction to Colletotrichum dematium infection, Reaction to Colletotrichum infection, Fungal Disease Resistance Traits, Disease Resistance, Reaction to Colletotrichum truncatum infection, Reaction to colletotrichum infection, Fungal Disease Resistance Traits, Disease Resistance, Reaction to Colletotrichum coccodes infection, Reaction to Colletotrichum infection, Fungal Disease Resistance Traits, Disease Resistance, Reaction to Glomerella cingulata infection, Reaction to Glomerella infection, Fungal Disease Resistance Traits, Disease Resistance, Reaction to Glomerella glycines infection, Reaction to Glomerella infection, Fungal disease Resistance Traits, Disease Resistance, anthracnose, fungal disease, disease, biotic stress, anthracnose

COFFEE=

Coffea arabica, Viridiplantae, Fukaryota, organism, coffee(liquid drink), coffee beverage, coffee based beverage product, nonfermented plant derived beverage product, steeped beverage product, Coffea arabica, Coffea, Coffeeae, Ixoideae, coffee, coffee, coffee, fruit, plant organ, carposphere part, phyllosphere, coffee, coffee, coffee, Coffea arabica, Gardenieae complex, Ixoroideae,Rubiceae, Coffea, Coffeeae, Ixoroideae, Rubiaceae

BRAZIL=

Brazil, Country, South America, Continent, GAZ_00000448, geographical region, Brazil, Brazil, Brazil

Figure 6. Results from Query 1 - terms in color are those repeated in many ontologies.

Photos of to in produced in -> 2487

Photos of Reaction to Colletotrichum dematium infection in coffea arabica produced in Brazil

Photos Reaction Colletotrichum dematium infection coffea arabica produced Brazil -> 15

Photos of Reaction to glomerella infection in coffea produced in South America

Photos Reaction glomerella infection coffea produced South America -> 73

Figure 7. Number of metadata records returned to the user for the first query, with and without semantic expansion.

of our previous research [3]. In our current contribution, we presented more in-depth information regarding the functioning of the scientific data search records in our conceptual architecture. We presented the data flow in the architecture and how it was instantiated in a case study in the agriculture domain by implementing search scenarios for the evaluation of our software tool.

We went through the whole development process, which considered from design to algorithmic specification and implementation. We tested our solution using a large ontology portal. 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.

In the evaluation conducted, we found that the preprocessing play a key role in the obtained results. We observed the need to process queries before directing them to the search engine. Cleaning the user's query causes irrelevant information to be discarded.

Our semantic query expansion procedure was able to produce alternative queries to the original one. Scenarios with query cleaning were the best among them. In particular, the scenario with cleaning the original query and using ontologies for query enrichment proved to be the most improved. It presented an adequate amount of metadata records to the user.

Our proposed and implemented architecture and algorithms can be applied to different knowledge areas. As implementation aspects, it is enough that the access to ontologies is modified to other areas. For instance, for biomedical area, we can use BioPortal [20] instead of the AgroPortal.

6. Conclusion

Metasearch still requires further studies to exploit the full possibilities of ontological aspects for semantically enabled search engines. In this article, 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.

As future research, we plan to study how metadata standards and queries can take versioning of data into account. In this sense, search results to a query would be as "time series" of data. This is hard to design (and implement) because it requires deciding on which kind of metadata timestamp to

[20] https://bioportal.bioontology.org

consider - such as “deposit time”, “creation time”, etc. We also plan to improve the way results are shown to end users. This is relevant for the user search experience. Further studies will involve researchers in agricultural sciences for helping us to express requirements and validate results.

Author Contributions

- Gustavo Caetano Borges: solution conception, experiment design, data collection, data analysis, software implementation, hypothesis testing, generalization and interpretation, writing and revision.
- Julio Cesar dos Reis: solution conception, experiment design, generalization and interpretation, writing and revision.
- Claudia Bauzer Medeiros: solution conception, experiment design, hypothesis testing, generalization and interpretation, writing and revision.

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).

References

[1] WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, v. 3, n. 1, p. 160018, dec 2016. Disponível em: ⟨http://www.nature.com/articles/sdata201618⟩.

[2] GOTTARDI, T.; MEDEIROS, C. B.; REIS, J. D. Understanding semantic search on scientific repositories: Steps towards meaningful findability. In: *1st virtual workshop on Research data management for Linked Open Science-DaMaLOS*. [S.l.: s.n.], 2020.

[3] BORGES, G. C.; REIS, J. C. dos; MEDEIROS, C. B. Addressing search in scientific open data repositories: A semantic metasearch platform. In: SBC. *Anais do XV Brazilian e-Science Workshop*. [S.l.], 2021. p. 81–88.

[4] BREEDING, M. Plotting a new course for metasearch. *Computers in Libraries*, v. 25, n. 2, p. 27–29, 2005.

[5] SIMIONATO, A. C. Mapeamento dos metadados para dados científicos. In: *XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XVIII ENANCIB)*. [S.l.: s.n.], 2017.

[6] KAISER, K. A. et al. Metadata: The accelerant we need. *Information Services & Use*, IOS Press, n. Preprint, p. 1–11, 2020.

[7] PIERRE, M. S.; LAPLANT, W. P. J. Issues in Crosswalking Content Metadata Standards. *National Information Standards Organization - White Papers*, 1998. Disponível em: ⟨https://groups.niso.org/publications/white_papers/crosswalk/⟩.

[8] COSTA, M.; BRAGA, T. Repositórios de dados de pesquisa no mundo. *Cadernos BAD*, v. 0, n. 2, p. 80–95, 2016. Disponível em: ⟨https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1585⟩.

[9] SANCHEZ, F. A.; Da Silva, N. B. P.; VECHIATO, F. L. Padrões de metadados para representação e organização da informação em repositórios de dados de pesquisa. *Informação & Tecnologia*, v. 5, n. 1, p. 37–51, feb 2019. Disponível em: ⟨http://periodicos.ufpb.br/index.php/itec/article/view/38350⟩.

[10] YAN, Q. et al. Community metadata ISO 19115 adaptor. *28th International Conference on Computers and Their Applications 2013, CATA 2013*, p. 213–218, 2013.

[11] SANTO, J. do E.; PAULA, E. V. de; MEDEIROS, C. B. Exploring Semantics in Clinical Data Interoperability. In: SPRINGER INTERNATIONAL PUBLISHING. *Advances in Conceptual Modeling*. [S.l.], 2019. p. 201–210.

[12] IZQUIERDO, Y. T. et al. Keyword Search over the COVID-19 Data. In: *Anais XXXV SBBD*. Porto Alegre, RS, Brasil: SBC, 2020. p. 205–210. Disponível em: ⟨https://sol.sbc.org.br/index.php/sbbd/article/view/13642⟩.

[13] ÁVILA, R. et al. Ligações Semânticas Utilizando Predicados SKOS. In: *SBBD*. [S.l.: s.n.], 2017. p. 88–99.

[14] GAVANKAR, C. et al. A comparative study of semantic search systems. *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, p. 1–7, 2020.

[15] JONQUET, C.; SHAH, N. H.; MUSEN, M. A. The open biomedical annotator. *Summit on translational bioinformatics*, v. 2009, p. 56–60, 2009. Disponível em: ⟨https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041576/⟩.

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