Semantic Search on Scientific Repositories: A Systematic Literature Review

Thiago Gottardi[1], **Claudia Bauzer Medeiros**[1], **Julio Cesar Dos Reis**[1]

[1]Institute of Computing, University of Campinas – UNICAMP
Av. Albert Einstein, 1251, Campinas, 13083-852, Brazil

{gottardi,cmbm,jreis}@ic.unicamp.br

Abstract. *Open Science has been recognized as one of the most important movements for leveraging scientific collaboration, helping scientists produce high quality research through sharing and reuse. It is usually defined as a combination of three factors: open access, open data and open processes, and relies on the corresponding publication of papers, data and software in repositories that can be publicly accessed on the Web. However, finding relevant papers, data and software has become one of the associated problems. Many search mechanisms – in particular semantic search – have risen as a means to solve this issue. Nevertheless, implementing these mechanisms and integrating them into scientific repositories presents many challenges. This paper presents a systematic literature review of research efforts on mechanisms for supporting search for scientific papers, data and processes, based on extracting and analyzing the entire contents of Scopus and IEEE Xplore.*

1. Introduction

The sharing of research results has become a key enabler for Open Science (Woelfle et al., 2011), thereby enabling advancement of science through reuse of such results. Open Science relies on a combination of three major factors: open publications, open data, and open processes, all made available in public repositories. A major obstacle to effective reuse is the difficulty to find relevant papers, or data, or processes (which include, among others, software and workflows). We identified that these three factors, together with authors, constitute the four most important parameters considered by search mechanisms. For simplicity, we refer to these parameters as *classes*.

Search mechanisms vary widely in approaches and purposes. Our main concern is with semantic search mechanisms that serve Open Science purposes, namely *supporting search for scientific papers, scientific data, and scientific software in public repositories*. Which mechanisms, however, are best suited to help researchers in their work? Are there still pressing research issues that need to be considered? Indeed, few studies are concerned with literature review on semantic search issues. To the best of our knowledge, there are no systematic literature reviews on the context of semantic search and its integration to scientific repositories; rather, surveys cover associated issues. For instance, Xu et al. (2013) presented a study on semantic search by providing a survey on schemas for metadata associated to scientific publishing; Zhang et al. (2019) studied approaches to identify the requirements for metadata search in the context of scientific data management. The work of Karimi et al. (2019) analysed different approaches that employ

thesauri and ontologies for semantic search. As an example of a loosely related systematic review, Niknia and Mirtaheri (2015) presented a review on the progress of linked data technology for geoscience web portals.

This paper presents a systematic mapping of literature on semantic search mechanisms on public repositories containing papers, data or processes – from now on called *scientific repositories* in this paper. A systematic mapping is a method that allows to present empirical data from a broad subject of interest (Oakley et al., 2005), thereby structuring a research area. After searching all documents from IEEE Xplore and Scopus as inputs, we processed 387 documents (of which 297 are unique studies), providing a quantitative summarization, as well as as a qualitative categorization and descriptions of the objectives and class of objects employed in the corresponding approaches. We thus present two major contributions – the systematic review itself, and its discussion; and the presentation of a few major open problems concerning semantic search mechanisms for open science.

2. Applying the Systematic Mapping Methodology

Our literature review follows the structure of a systematic mapping (Oakley et al., 2005) and was executed according to guidelines (Kitchenham and Charters, 2007). These guidelines involve three sequential phases: (1) Planning; (2) Conducting and (3) Reporting. This section briefly outlines the Systematic Mapping Methodology, and how we applied it to analyze publications on semantic search mechanisms.

Planning is the first phase of the systematic mapping process; its output is a document named "Protocol". All documentation on the entire process is available online[1], which includes this protocol with research questions and extra documents. The Conduction Phase is composed by the "Selection" and "Extraction" activities, which must be performed by following the protocol. The Selection phase defined which studies must be selected from the Sources, based on their titles and abstracts. The Extraction phase involved completely reading the documents for extracting data as planned by the protocol.

We initially planned to carry out our search processes through nine different Sources. However, the results of selecting studies from some sources had to be cancelled after a few search sessions, since it was not possible to calibrate or complete the selection. Selection was executed through the end on the following Sources: IEEE Xplore (IEEE) and Elsevier Scopus (Scopus). The initial search sessions were executed on February 17, 2020 and updated throughout April 23, 2020.

During the Extraction phase, all studies were qualitatively summarized by manually evaluating their full texts completely. The extraction form was filled manually for each study. The form contained the following fields: (I) Existence of Integrated Search (boolean); (II) Existence of Semantic Mapping (boolean); (III) Identified Software Architecture (nominal); (IV) Identified Objectives for Scientific Data (nominal); (V) Identified Class of Scientific Data (nominal);

3. Results, Discussion and Open Challenges

Following the search sessions, We ended up with a total of 323 documents, of which 297 are unique studies. An analysis of the results is reported in this section.

[1] `http://tiny.cc/gottardi-semantic-review`

3.1. Integrated Semantic Search and Semantic Mapping

The primary research question of the Protocol is related to the existence of "Integrated Semantic Search" in the literature. Semantic search has been applied to different scientific fields. The term "integration" has been used loosely in the text, and can be found in many contexts; as shown in the previous Section, 75 total studies that involve some sort of "integration". We identified three different meanings for this term in the context of semantic search. The first meaning was how to connect multiple databases that include semantics with the intent to search them jointly – this was identified in 38 studies. The second meaning was to take existing data and study how to add semantics to this data, identified in 33 studies. The third meaning relates to how to add a semantic layer to existing search engines, identified in another 33 studies. This semantic layer is closely related to semantic mapping.

Moreover, we were not able to find a generic proposal that was tested on multiple scientific fields – namely, an integrated approach to semantic search combining arbitrary domains, *i.e.*, studies are motivated by or solely tested on a specific scientific domain, usually life sciences. Thus, a related question is: "how generic are the proposed mechanisms?". Many argue that, since their proposal is based on specific ontologies, changing the ontology would provide appropriate support to other domains. However, domain specificity hinders generality. The challenge is to balance between a domain-specific and a generic semantic search.

Similar to "integration", there are different meanings of "Semantic Mapping". In general, semantic mapping refers to metadata fields added to the actual data with the intent to enrich the data with semantic information. Our work identified 17 studies concerning semantic mapping, and identified three categories of this mapping. The first category, corresponding to 8 studies, is the "Manual Definition", in which metadata is manually specified by authors or curators, Since this represents a complex increase on work efforts, new approaches to automate these efforts were reported. In this sense, we identified a second category, which we named "Automatic Definition", in which metadata is added automatically by computers, presented in 9 studies. Automatic definitions also present challenges – *e.g.*, when algorithms add incorrect metadata. As part of efforts to address this issue, researchers created what we name "Fuzzy Mechanisms", which are variations of automatic metadata definitions, and which we identified in three studies. Fuzzy mechanisms are those that use metadata to sort results by relevance, including loosely related - as opposed to "automatic definitions" in which only return directly related. We highlight that fuzzy definitions may lack precision. No identified study advocates "Strict Definitions", for example, the application of formal definitions to avoid ambiguity within the semantic search.

Different software architectures have been adopted while designing integrated semantic search engines. We identified 76 studies that propose an integrated implementation. Most of the studies (24 in total) are based on multiple database composition, *i.e.*, the authors integrate several databases by implementing a single query system. This category of system appears in many situations, including large scale computing systems, *e.g.* clusters and grids, slowly been replaced by the emergence of cloud computing, which is represented by 5 papers. A total of 20 studies indicate the use of web-based systems, often advocating that this implementation is adequate for the mainstream community. We also

found many prototype proposals (reported by 11 papers); we could not check the actual architecture of these prototypes, since they were not described. Semantic integration can be added as a layer to existing databases. Therefore, we expected studies suggesting middleware software solutions to support this kind of integration. However, only one study reported this attempt, which may indicate that this presents an implementation challenge to be followed up.

There are four main classes of search parameters declared in 63 studies: (a) Science Data: including text notes, spreadsheets, images, videos, recordings (41 studies); (b) Documents: including articles and theses (12 studies); (c) Processes: involving workflows; methods, hypotheses, comparison metrics, software (21 studies); and Author names and their affiliations (3 studies). Though the latter is not directly included in the three Open Science axes, it is a frequent parameter of search mechanisms. Considering the total number of processes and software repositories, we identified that a subset are not for scientific software (9 studies). Our results indicated that most papers only focus on a single object class. Indeed, out of 63 studies, only 12 deal with more than one data type and no study involved more than two classes. Thus, another research challenge is the design of (semantic) search mechanisms that allow combining distinct kinds of search parameters - documents, data, processes and authors.

3.2. Objectives and Data Classes

The search mechanisms also had different objectives or goals, as reported on the study, involving 59 studies. The most common objective is to access the resulting data, moreover retrieving the results and to notify users when new results appear. The second most common objective is discovery of new conclusions that are not part of the original data submissions, including how to identify existing discovery aggregate data to identify and infer new conclusions (35 studies). A slightly less frequent objective is management, where existing data, documents and authors are registered and reported (23 studies). A less common objective, 3 studies focused on simulations; they may be used, for instance, to generate data for experiments and observations, validate data or extrapolate findings.

Another study focused on using the search mechanism for auditing data and conclusions; by using the collected data it is possible to identify the authors responsible for each claim, verify data, ensure correctness and detect frauds or corruption. The same study discussed reproduction or replication, where the experiments returned by the search should be reproduced or replicated to verify the findings. Finally, there were studies where the search was employed for supporting review efforts. Study reviews use existing documents and summarize them for creating new documents, aggregating their quantitative data and qualitative descriptions and comments into into a new (non-primary) study. Our review methods could eventually benefit from these search engines as well.

Table 1 shows the seven objectives and four object classes, resulting in combinations that employ the class (subject) with the action for an objective (verb). The table includes their frequency and descriptions; its columns are composed by the object class, while its rows are composed by objectives. Each cell in the middle contains the number of studies followed by its description. The descriptions are colored according to the number of identified studies. Different combinations may indicate new opportunities for the usage of the given data, though some may not be feasible. Figure 1 includes a plot for the

distribution of both class and usage objectives, showing how the number of studies vary from 1997 to 2020. These plots provide insights on periods in which objectives and object class appeared, *e.g.*, the increase on discovery studies or the rarity of workflow studies.

Table 1. Frequency and Descriptions for Objectives and Combination of Classes

		Class			
		Scientific Data	Document	Process	Authors
Objective	Access	**28**: Search, query, access, recommend and/or retrieve science data.	**10**: Search, query, access, recommend and/or retrieve papers, articles, journals, reports, magazines, etc.	**8**: Search, access, recommend and/or retrieve science data.	**2**: Search and find or recommend authors and related authors.
	Discover	**22**: Discover conclusions using aggregated science data.	**4**: Discover conclusions and related documents using existing documents.	**7**: Discover combined workflows.	**1**: Discover what authors collaborate on research efforts.
	Manage	**13**: Manage known science data, also their sources and bases.	**2**: Manage known document references/citations. Manage documents being written.	**5**: Manage known workflows and assess their usage.	**1**: Manage known authors, relationships, contributions and their roles.
	Simulate	**3**: Simulate experiments and compare against existing data for validation.	**0**: Simulate document publications and acceptance.	**1**: Simulate workflow usage and outcomes.	**0**: Simulate author contributions and outcomes.
	Audit	**1**: Audit data for validation and verification; protect from corruption and false data; blame manipulators.	**0**: Audit documents to verify authorship and protect documents from corruption.	**0**: Audit execution of workflows. Audit who can edit the workflow.	**0**: Audit roles and authorship to protect authors' curricula from corruption and false data.
	Replicate	**1**: Replicate studies based on existing science data and compare the outcomes.	**0**: Replicate (or plagiate) existing documents and their structures.	**0**: Replicate existing work-flows and compare their outcomes.	**0**: Plagiate author roles.
	Review	**0**: Review and compare data sets of science data to aggregate results.	**1**: Support for literature reviews.	**0**: Review work-flows and methods and compare their efficiency.	**0**: Review existing author roles and contributions.

An analysis of the objectives shows future challenges, including cases that would benefit from semantic search. Some objectives identified in studies that are unrelated to semantic search – *e.g.*, studies concerned with prediction, which allow to "Predict" or estimate new data from existing data. Other studies advocated the support of "Data Export", that allows users to take data results and explore them using software tools. Another challenge is to use semantic search to find "Teaching" material for students. "Visualization" combined to semantic search could lead to better comprehension for both the semantic queries as well as the results from the semantic searches. There are also other objectives we identified after analyzing recent opportunities. For instance, there are no studies beyond the current data management tools that include strategic decisions for the future research efforts. Another opportunity is support the design of public "Policies" based on evidence. A completely missing objective we identified is the lack of semantic search to extract specific data and metadata from "Internal" content available in documents and data, *e.g.*, article sections or images.

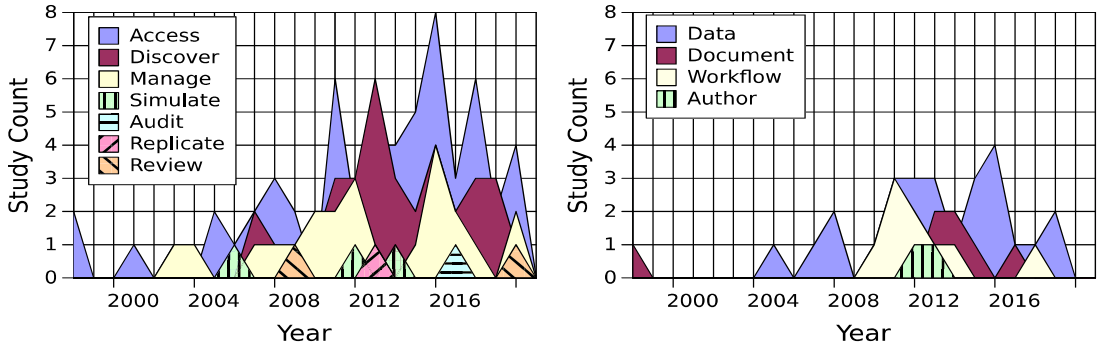

Figure 1. Objective and Class Distributions over years

4. Conclusions and Ongoing Efforts

Open Science relies on sharing research results, usually grouped into three classes - articles, data and processes. Effective sharing requires findability, and for this we must understand research efforts on search mechanisms. To this purpose, this paper presented a systematic literature review on semantic search issues – analyzing and synthesizing 297 papers as a result of processing the entire contents of IEEE Xplore and Scopus. We presented both quantitative and qualitative results, providing insights and pointing out open research issues to be addressed. The full set of results, including detailed methodology, graphic plots and analysis datasets are provided as extra documents[2]. Ongoing work concerns analysing a broader scope of studies. We also intend to search pre-prints and additional repositories, as well as provide updates. In addition, we plan to continue analysing further studies to provide additional descriptive analyses.

Acknowledgements

Work partially supported by FAPESP (#2019/19389-1, #2017/02325-5 and #2013/08293-7), and CNPq (#428459/2018-8 and #305110/2016-0).

References

Karimi, E., Babaei, M., and Beheshti, M. (2019). The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems. *Iranian Journal of Information Processing and Management*, 34(4):1579–1606.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, UK.

Niknia, M. and Mirtaheri, S. (2015). Mapping a decade of linked data progress through co-word analysis. *Webology*, 12(2).

Oakley, A., Gough, D., Oliver, S., and Thomas, J. (2005). The politics of evidence and methodology: lessons from the eppi-centre. *Evidence & Policy*, 1(1):5–32.

Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, 3:745–748.

Xu, H., Sun, L., Zou, M., and Meng, A. (2013). A survey of scientific metadata schema. *Applied Mechanics and Materials*, 411-414:349–352.

Zhang, W., Byna, S., Niu, C., and Chen, Y. (2019). Exploring metadata search essentials for scientific data management. In *2019 IEEE HiPC*, pages 83–92.

[2] `http://tiny.cc/gottardi-semantic-review`