Semantic Search to Foster Scientific Findability: A Systematic Literature Review

Thiago Gottardi, Claudia Bauzer Medeiros, Julio Cesar Dos Reis

Institute of Computing, University of Campinas – UNICAMP
Av. Albert Einstein, 1251, Campinas, 13083-852, Brazil
{gottardi,cmbm,jreis}@ic.unicamp.br

Abstract. One of the main goals of the Open Science movement is to leverage scientific collaboration through, among others, promoting the sharing and reuse of research outputs, such as publications, data and software. Sharing is enabled by public and accessible scientific repositories where these outputs are managed throughout their lifecycle. In this context, finding these digital artifacts has become a key problem. Semantic search mechanisms have risen as a means to solve this issue. However, implementing and integrating them into scientific repositories presents many challenges. This article presents a systematic literature review of research efforts on mechanisms for supporting search for scientific papers, data and processes. Our investigation is based on extracting and analyzing the entire contents of nine digital libraries using the associated search engines – in alphabetical order: ACM Digital Library, arXiV, Engineering Village, IEEE Xplore, SBC OpenLib, Springer Link, Scopus, Wiley Online Library and Web of Science. After retrieving a combined amount of 5012 documents, we identified 2054 unique papers that were used as a basis for our analysis. Our findings provide, among others, a new categorization of literature on search and discuss unexplored gaps, thereby contributing to advancing research on semantic search mechanisms to support Open Science.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Open Science, Semantic Search, Scientific collaboration

1. INTRODUCTION

The sharing of research results has become a key enabler for Open Science [Woelfle et al. 2011], thereby fostering advancement of science through reuse of research outputs. Though there are many definitions for Open Science, there is a consensus that it should support at least three mechanisms for sharing of knowledge: open publications, open data, and open processes and methods, all made available in public repositories.

A major obstacle for effective reuse is *findability* of data - and thus the institution of FAIR principles for data sharing and reuse [Wilkinson et al. 2016], extensible to papers and 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 mechanisms that help search for research outputs. These mechanisms assume that researchers look for data, papers or processes of interest, and their authors. To avoid constant enumeration of these four parameters, we simply refer to them as *classes* – of search parameters.

Search mechanisms are cumbersome, and often require lengthy efforts to identify artifacts of interest. Several research solutions were proposed to alleviate the search process – such as the use of metadata standards, consensual vocabularies, or annotations. Semantic search mechanisms have risen as a means

Work partially supported by FAPESP (#2013/08293-7, #2019/19389-1, and #2017/02325-5), and CNPq (#428459/2018-8 and #305110/2016-0), The opinions expressed here are not necessarily shared by funders.

to improve the quality of search results. 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, data, and software in public repositories.* Besides these purposes, we discuss the availability of search for authors and their affiliations.

Our research is concerned with investigating search mechanisms that are best suited to ensure meaningful findability. This has led us to come up with the following research questions: RQ1) What are the integrated semantic search and semantic mapping approaches in the literature? RQ2) How are the proposed search systems characterized, and what are their software architectures? RQ3) How can we further characterize the search platforms? RQ4) What are the platforms' objectives and the classes of data handled by the corresponding approach? In this work, we present a systematic literature review to answer these questions. During our review investigation, we realized that researchers adopt different meanings for both "integrated semantic search" and "semantic mapping". As a consequence, our results also include a discussion of these meanings. To the best of our knowledge, there are no systematic literature reviews in the context of semantic search and its integration to scientific repositories; rather, reviews cover associated issues. Our systematic mapping focuses on publications concerning semantic search mechanisms. These mechanisms are applied to public repositories that contain publications, data or processes – from now on called *scientific repositories.*

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. Our systematic mapping was conducted using contents from the following repositories: ACM Digital Library[1], arXiV[2], Engineering Village[3], IEEE Xplore[4], SBC OpenLib[5], Springer Link[6], Scopus[7], Wiley Online Library [8] and Web of Science[9]. From these repositories, we identified 2054 unique documents, describing a variety of search mechanisms and approaches. We provide a quantitative summarization, and a qualitative categorization and descriptions of the objectives and class of objects employed in the corresponding approaches.

This work presents two major contributions – the systematic review itself, and its discussion; and the presentation of a few major open problems concerning semantic search mechanisms with open science in mind. Our results indicate that most semantic search approaches lack several factors to fully meet findability – *e.g.*, flexibility in search parameters, or support to multiple domains. Our analysis points out that there are still many research challenges on the use, design and implementation of mechanisms for semantic search on open scientific repositories. In particular, we discuss how they can be enhanced to provide a more meaningful range of results. As such, we provide insights into steps towards semantic search efforts to meet the demands of the open science movement.

A preliminary version of this work was published as a short paper [Gottardi et al. 2020a], subsequently extended to a full workshop paper [Gottardi et al. 2020b], which discussed additional analyses. The two previous versions and the present paper adopt the same methodology for systematic review. All the rest reported in this paper has not been published before (including most numerical results and qualitative analysis). Furthermore, they only concentrated on answering RQ4, since we did not have enough data to satisfactorily answer the other questions. The search engines covered in the previous papers were IEEE Xplore and Scopus (short paper) and arXiV (full workshop paper), with analysis performed up to August 2020. Here, we extend the study to six additional engines, with analysis

[1] `http://dl.acm.org`
[2] `http://arxiv.org`
[3] `http://www.engineeringvillage.com`
[4] `http://ieeexplore.ieee.org`
[5] `http://sol.sbc.org.br`
[6] `http://link.springer.com`
[7] `http://www.scopus.com`
[8] `http://onlinelibrary.wiley.com`
[9] `http://www.webofknowledge.com`

performed up to February 2021 – ACM DL, Engineering Village, SBC OpenLib, Springer Link, Wiley Online Library, Web of Science. While the previous papers covered analysis of a total of 324 unique documents, this paper analyzes 2054 unique documents. This numeric difference brought novel insights that are described in this work. A more thorough enumeration of the differences between this version of our work, and the two previous versions appears in Section 5.

The rest of this article is organized as follows. Section 2 presents the SM (Systematic Mapping) procedure, describing the stages performed in our literature review. Section 3 reports on the quantitative results gathered from the SM process, which are then qualitatively discussed in Section 4. Section 5 describes key related studies and compares existing results with our contributions. Finally, Section 6 concludes this article.

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 of [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. We employed iterative selection techniques to better provide quantitative analyses on the selected studies, a technique similar to screening, as used in recent mapping studies [Hummel et al. 2021]. Further documentation on our entire process is available online [Gottardi 2021].

The following terms are used throughout the rest of this paper, following terminology adopted in systematic mapping work [Kitchenham and Charters 2007]: (a) a *Primary Study* is any written study that investigates a research question; (b) a *Secondary study* is a work that reviews primary studies – *i.e.*, ours is a secondary study; (c) a *Source* is a digital repository that is searched to retrieve studies; (d) a *Search engine* refers to dedicated search tools made available by the Sources to retrieve studies. For brevity sake, we often refer to "a search engine" when we actually mean "an engine to process a search request on the corresponding Source".

2.1 The Planning Phase

Planning is the first phase of the systematic mapping process; its output is a document named "Protocol". Table I presents an excerpt of the protocol we specified to perform our systematic mapping. Each row displays a protocol item and corresponding description – objective, questions, intervention, results, Source selection criteria, and study selection criteria. In more detail, a *Source*, in the protocol, is a repository in which documents relevant to our work will be "selected" – e.g., the ACM Digital Library is an example of a "Source", and any document retrieved from it that meets our selection criteria is a *study*. For instance, a conference paper on semantic search mechanisms published in ACM DL is a study of interest to our analysis.

The protocol defines the criteria for studies to be considered at each Source – Inclusion (I1 to I3) and excluded – Exclusion (E1 to E4 criteria). I1 was planned to select all papers that involve any kind of search or query on databases. I2 was planned to include papers discussing any kind of integration scheme, *e.g.*, integration of different datasets. I3 involves selecting studies that present any kind of semantic mapping approach, *e.g.*, through use of annotation, metadata or ontologies that are used to map digital artifacts (data, documents, software) to semantic predicates. We point out that we, on purpose, used in I1 and I2 terms that have more than one interpretation - namely "integration" and "mapping", so that we could select a larger set of studies to analyze: the disambiguation of these terms is discussed as part of the results.

Exclusion criteria are associated to studies that should not be used in our quantitative and qualitative analyses. E2 through E4 are self-explanatory – E2 and E3 concerned unrelated studies, such as those that were retrieved because of appropriate keywords, but in unrelated meanings, *e.g.*, the search for something that is not in a database. E1 ("not a valid document or inaccessible") involves results that are not linked to actual documents, *e.g.*, conference calls and references to papers that do not exist. "Inaccessible" means that we could not access the entire contents of the primary study.

In most cases, we excluded as non-valid all documents that were not scientific studies, *e.g.*, a call for papers or a report on a conference. We also excluded documents that might not be full-fledged papers or articles, *e.g.*, posters, or talk summaries. Last but not least, book chapters were also excluded as non-valid because we have identified that these kinds of documents are often partial, *e.g.*, when a single study is sliced into several chapters. They also often repeat content of articles and papers written by the same authors.

Table I. Protocol Definition. Composed by Protocol items and their descriptions.

Protocol Item	**Item Description**
Objective	Identify existing approaches to integrating semantic search mechanisms on scientific production.
Research Question	What are the approaches and techniques that perform semantic search on scientific production?
Intervention	Identify and categorize related primary studies.
Results	Quantitative data on frequency distribution within categories. Qualitative data on approaches that integrate semantic search on scientific databases.
Source Selection Criteria:	Source must include indexed studies on Computer Science, Mathematics, Engineering, Medicine or Biology[11]. Extraction of papers from source must allow Boolean operators. Source must be accessible to us.

Study Selection Criteria:	Inclusion I1 - Scientific Database Search approach;
	Inclusion I2 - Approach involves Integration;
	Inclusion I3 - Application of Semantic Mapping;
	Exclusion E1 – Not a valid document or inaccessible;
	Exclusion E2 – Unrelated to computing/databases;
	Exclusion E3 – Does not discuss search issues;
	Exclusion E4 – Not primary study.

Data Collection Form	F1 Contains Integrated Search (boolean); *if(F1) then:* F1.1 Integration Type (nominals) F2 Contains Semantic Mapping (boolean); *if(F2) then:* F2.1 Semantic Mapping Type (nominals) F3 Contains Software Architecture (boolean); *if(F3) then:* F3.1 Identified Software Architecture (nominals) F4 Contains Class of Scientific Data (boolean); *if(F4) then:* F4.1 Class of Scientific Data (nominals) F5 Contains Objectives for Scientific Data (boolean)); *if(F5) then:* F5.1 Objectives for Scientific Data (nominals) *if(F1∧F2) then:* F6 Descriptive Summary for Detailed Study (text)

2.2 The Conduction Phase

The Conduction Phase concerns the actual systematic mapping; it is composed by the "Selection" and "Extraction" activities, which must be performed according to the protocol. In a nutshell, during the Conduction phase, all results from the search engines are analyzed; the duplicate documents are linked; invalid documents are removed; the relevant studies are selected and those most applicable studies are qualitatively analyzed as part of the Extraction activity. Figure 1 presents this process.

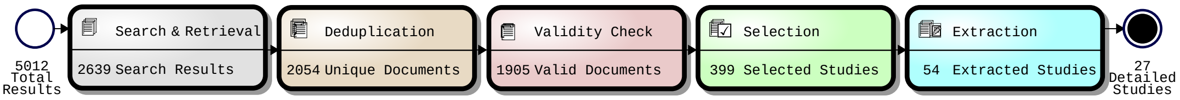

Fig. 1. Overview of the Conduction Phase.

The Conduction phase begins by running queries on the search engines, which return search *results*. The combination of all search results is referred to as "Total Results". Next, we *deduplicate* these results to find unique documents – since a given study may appear in more than one Source. Afterwards, we identify the valid documents – namely, those that correspond to publications (having title, author and abstract). Deduplication and validity check activities involve both automatic and manual checking.

The *Selection* and *Extraction* activities are performed manually. *Selection* involves selecting the (already validated) primary studies of interest. To this end, Inclusion and Exclusion criteria are applied according to titles and abstracts; the outcome of this activity is the set of *selected studies.* We provide quantitative analyses on all selected studies (cf. Section 3). The *Extraction* activity consists in reading the studies completely and writing a detailed summary. Here, we add further details to the information collected during the selection activity.

2.2.1 *Search Strategy.* The invocation of the search engines was divided into different sessions, which were conducted to retrieve three different categories of studies. The first category is focused on approaches that include any type of semantic search aspect, including semantics, ontologies, metadata or annotations. The second category comprised all types of data retrieval and search approaches, regardless of whether using semantics or applied to scientific repositories. The third category focused on synonyms for scientific, research and studies. "Research" was eventually removed as a search parameter for being far too common. "Study packing" was added since it has been used to refer to documentation of systematic reviews.

Table II shows the four search strings applied to title, abstract and keywords of each study in the Sources. Each string was constructed by joining the basic keywords defined in the protocol. Afterwards, each string was used as sessions to retrieve studies from the Sources. Strings were calibrated by identifying whether a set of previously known papers were present in the results. For instance, the "research" keyword was removed after calibration, since it is ambiguous and extremely common. While our focus was on semantic search, our strategy for selecting studies for our systematic review included other search approaches for completeness' sake.

Selection was executed through the end on the following sources, listed by order of session execution: Elsevier Scopus; IEEE Xplore; Cornell University arXiV; Wiley Online Library; ACM Digital Library; Elsevier Engineering Village; Clarivate Web of Science; Springer Link; SBC OpenLib. The initial search sessions were executed on February 17, 2020 (Scopus only) and updated throughout February 1^{st}, 2021 (comprising all sources).

SBC Open Lib and arXiV were repeatedly searched using the indicated search strings, but many of the search attempts returned no results. It is possible that these sources require custom calibration of search strings. Calibrating the string for specific search engines may also cause a search bias. Search on Springer Link excluded text previews, because these would add several unrelated results that were not accessible by us. We point out that in arXiV authors are able to revoke access to their pre-prints, which can decrease the number of search results throughout search history. We initially planned to include Google Scholar[12] as one of our Sources. However, the results of selecting studies from it was canceled after a few search sessions, since it was not possible to calibrate or complete the selection. Google Scholar often duplicates studies, presents an imprecise total number of search results and mixes non studies with studies. Though Google Scholar was considered by Gusenbauer [Gusenbauer 2019] as the most comprehensive academic search engine, in a more recent article, Gusenbauer *et al.* [Gusenbauer and Haddaway 2020] argued against using this source to conduct systematic reviews. We also discarded using the University of Trier/Schloss Dagstuhl DBLP[13] as a source, because it does not provide abstracts. Subsection 4.1 presents further discussions on possible bias and limitations.

2.2.2 *Study Selection.* After deduplication, the review was conducted manually by exhaustively analyzing studies returned by each Source according to the inclusion and exclusion criteria. Data was collected on the selected studies for quantitative analyses by using a *Data collection form* (*c.f.* Table I); This form was manually filled for each study.

2.2.3 *Study Extraction.* During the Extraction phase, all studies were qualitatively summarized by manually evaluating their full texts completely. Studies that are accepted for the extraction activity

[12] `http://scholar.google.com`

[13] `https://dblp.uni-trier.de`

Table II. Definition of search strings. The Name column identifies each string: *I* for Semantic Integration, *Q* for Semantic Query, *R* for Semantic Retrieval and *S* for Semantic Search. String column contains the resulting string.

Name	**String**
I	(("semantic information retrieval" OR "ontology information retrieval" OR "metadata information retrieval" OR "meta data information retrieval") AND ("scientific" OR "study pack" OR "study packing"))
Q	(("semantic query" OR "ontology query" OR "metadata query" OR "meta data query") AND ("scientific" OR "study pack" OR "study packing"))
R	(("semantic retrieval" OR "ontology retrieval" OR "metadata retrieval" OR "meta data retrieval") AND ("scientific" OR "study pack" OR "study packing"))
S	(("semantic search" OR "ontology search" OR "metadata search" OR "meta data search") AND ("scientific" OR "study pack" OR "study packing"))

must have all fields of the *Data collection form* (*c.f.* Table I) completely filled in this process. Table III presents an example of a filled *Data collection form* (see [Gottardi 2021] for all the forms). Extracted Studies were summarized by writing a descriptive text, as further discussed in Subsection 3.3.

Table III. Example of a Data Collection Form created for a reviewed study

	Reference: [Sengloiluean and Khuntong 2020] – Journal Article			
Activity	**Inclusion**	**Value**	**Exclusion**	**Value**
Selection	I1	True	E1	False
	I2	True	E2	False
	I3	True	E3	False
			E4	False
	Field	**Value**	**Dependent Value**	
Selection and Extraction	F1	True	Multiple Data Source Integration	
	F2	True	Automatic Semantic Mapping	
	F3	True	Presents Prototype / Multiple Database	
	F4	True	Documents (teaching)	
	F5	True	Access/Search	
Extraction	F6	An approach for integration of learning resources stored in heterogeneous databases by using mapping ontologies.		

2.3 Reporting Phase

Reporting consists on writing the results of the analysis conducted during the extraction phase. This involved both quantitative and qualitative summarization of the studies. The quantitative analysis took the form of a descriptive statistical analysis. The qualitative analysis was performed by writing textual descriptions for each study and clustering them according to their form data. Both analyses are described in Section 3.

3. RESULTS OF THE SYSTEMATIC MAPPING

First we show the sets of studies retrieved by the search strings, followed by an analysis of the results. The distinct kinds of results (e.g., valid results, extracted results) are those described in Figure 1 that provides an overview of the Conduction Phase (cf. Section 2). We point out that our analysis considered studies that approached semantic search mechanisms from multiple angles – either by directly proposing such mechanisms or by proposing architectures or infrastructure to enable search.

3.1 Results from Invoking Search Engines

Table IV presents the number of "Results" returned by each search session grouped by their "Search String" (*cf.* Table II) and "Source". Sources are indicated in alphabetical order, respectively in

Table IV. Number of Results from each Source. Row Acc. Sum indicates the accumulated sum of all search result numbers; Max. contains the maximum search result numbers; Unique indicates the number of documents after their duplicates were discarded at the Deduplication phase; Valid is the number of valid documents from the Validity Check.

	Search String	ACM DL	arXiV	EV	Xplore	Scopus	Springer L	SBC OL	Wiley OL	WoS	Total
Acc. Sum	I	44	0	39	15	31	15	0	22	10	5012
	Q	406	3	163	90	218	213	0	136	32	
	R	74	0	145	42	135	61	0	214	40	
	S	361	0	396	119	571	707	0	518	192	
Max.	I	44	0	20	5	11	15	0	11	5	2639
	Q	204	3	83	31	76	213	0	69	16	
	R	74	0	73	14	45	61	0	114	20	
	S	361	0	199	42	196	354	0	184	96	
Unique	I	39	0	9	4	6	10	0	9	1	**2054**
	Q	195	3	44	25	67	207	0	63	2	
	R	63	0	34	5	35	56	0	117	2	
	S	299	0	41	17	189	301	0	197	14	
Valid	I	39	0	9	4	6	10	0	7	1	**1905**
	Q	187	3	41	25	58	206	0	49	2	
	R	62	0	34	5	33	54	0	80	2	
	S	294	0	41	17	174	298	0	150	14	

columns: ACM Digital Library, arXiV, Engineering Village, IEEEXplore, Scopus, SBC OpenLib, Springer Link, Wiley Online Library and Web of Science. The number of unique documents per Source is affected by the search execution order (cf. Subsubsection 2.2.1). Since uniqueness requires eliminating duplicate documents – e.g., that were retrieved from another source – if a document from source A is also subsequently retrieved from source B, it is eliminated from B's results.

Following the search sessions, *cf.* Figure 1, we retrieved an overall 5012 documents (accumulated sum, first row of Table IV), with a maximum number of 2639 results in a search session – *cf.* second row. We ended up with a total of **2054** unique documents (remaining documents after excluding duplicates), of which **1905** are valid documents, *i.e.* those that have author and title. All unique documents were processed during the selection activity.

3.2 Results of the Selection Phase

After performing the selection activity, we obtained the number of studies according to each inclusion or exclusion criterion. Table V presents the results for the first and second selection phases. This table includes numbers of input and output study and columns for the number of studies as processed by the inclusion and exclusion criteria. We highlight that many of the documents were excluded by more than one exclusion criterion, which explains why the result of subtracting the exclusions from the inputs does not match the output numbers. Since there were different meanings for integration and semantic mapping, all of the meanings were accepted as part of selection. Subsection 3.3 discusses the different meanings we observed.

Table V. Results of the First and Second Phases of the Selection Activity. The last row shows between parentheses the number of studies that passed without any exclusion criteria.

Step	Input	E1 (Not Article/Paper)	E2 (Unrelated)	E3 (No Search)	Output
First	**2054**	504	118	307	**1029**

Phase	Input	I2 (Integration)	I3 (Semantic Mapping)	I2 ∩ I3	I2 ∪ I3	E4 (Non Primary)	Output
Second	**1029**	363 (209)	91 (50)	54 (27)	399 (273)	69 (0)	**27**

During the selection phase, we selected studies that contain search approaches and excluded those that matched any exclusion criterion. This phase included studies with either integration or semantic mapping (I2 ∪ I3), by providing grouped analyses on the set of studies that present integration, as well as another analyses on those that involve semantic mapping. For full description presented as the extracted studies, we present the studies that include both criteria (I2 ∩ I3).

3.3 Results of the Extraction Phase

Table VI and VII present the results of the Extraction phase, which includes **both** semantic mapping and integration. There were 54 studies in this phase. As some studies were excluded for being non primary, **27** studies remained (one per row). Tables VI and VII cite each study by its reference ("Ref") column, with a short "Descriptive Summary", and their categorization, as follows: "**Integration**" refers to the different ways through which each search approach integrated data: "Layer" stands for a semantic layer built on top of another database; "Multi" stands for the integration of multiple databases; "Existing" refers to annotating existing data to be enriched with semantics, "Tools" refers to a set of tools provided by the authors concerning the domain and/or search, *i.e.*, integration of data search results may be performed by an external semantic layer, or by integrating underlying databases, or indirectly via semantic annotations or by using integrated tools. "**Semantic Mapping**" includes the process executed to map semantics to data: either by "Manual" definitions or by "Auto" (automatic) definitions. We identified if the approach is "Strict" or "Fuzzy" where applicable. "**Software Architecture**" cites the referenced software architectures. "**Class**" indicates the type of the data handled by the approach according to the three main axes of Open Science: "Data" for scientific data; "Document" for Papers, articles and other documents; Methods, workflows, software and other processes for data handling are listed as "Process"; "Citation" for citation of authors and documents; "CFP" for call for papers announcements (any call for contributions is included); "Conference" involves data generated from conferences, including information, slides, reports, recordings and videos; "Funding" includes funding calls and reports generated from projects that receive funding; "Institution" combines information and reports from institutions that produce scientific data or papers. "**Objective**" involves data usage intent of each approach. "Access" refers to data access, including search and retrieval; "Discover" refers to the discovery of new conclusions based on existing data; "Review" is the activity of surveying and aggregating data from other studies. "Generate" includes automatic or manual synthetic data generation used to test database systems. Subsection 3.7 provides further analyses on these categories.

The subsections that follow show results grouped according to the research questions. The numbers of total studies are based on the selected studies. The charts show how these number of studies vary from 1984 to 2021 – the publication years of the oldest and newest selected study, respectively.

3.4 RQ1: Integrated Semantic Search and Semantic Mapping Approaches

Semantic search has been applied to different scientific fields. As shown in Section 3, we identified 209 selected studies that involve some sort of "integration". We identified four 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 – identified in 131 selected studies. The second meaning was to take existing data and study how to add semantics to these data, identified in 84 selected studies. The third meaning relates to how adding a semantic layer to existing search engines, identified in another 35 selected studies. This semantic layer is closely related to semantic mapping. The fourth meaning refers to the definition of integrated tools to be used for search, as identified in another 38 selected studies. These integration concerns are related to the software architecture of the approaches (further discussed in Subsection 3.5).

We were unable to find a generic proposal that was tested on multiple scientific fields – namely, an integrated approach to semantic search combining arbitrary domains. Rather, studies are motivated by or solely tested on a specific knowledge domain, usually life sciences. This simplifies semantics issues, since there is no need for, e.g., checking context-sensitiveness, as different contexts may involve different meanings for a single word. In this sense, a related question appears: "how generic are

Table VI. Detailed Studies of Extraction Phase. Columns 3 to 7 correspond to data on our research questions

Ref.	Descriptive Summary	Integration	Semantic Mapping	Software Architecture	Class	Objective
[Budak Arpinar et al. 2006]	A framework for exploiting analytics from geospatial data. Automatic capability for extracting information from metadata.	Existing	Auto	Prototype & Web	Data	Discovery
[Xiaoming et al. 2007]	A semantic model for annotating scientific data (material research) scattered over several databases.	Existing & Layer	Manual	Grid	Data	Access
[Zhizhin et al. 2007]	Search engine for enviromental data discovery. Includes fuzzy mapping for natural languages that facilitates queries.	Multi	Loose	Desktop & Grid	Data	Discovery & Access
[Pirrò et al. 2008]	Peer to Peer architecture for collaborative research.	Multi	Auto & Loose	P2P & Web	Document	Review & Access
[Kraines 2008]	A framework and web application for sharing scientific production with semantic reasoning capability for new knowledge discovery.	Existing & Layer	Strict	Web	Document	Discovery & Access
[Neri 2009]	Query rewriting based on formal ontologies that map different schemas.	Multi	Auto & Strict	Prototype	Data	Access
[Kumazawa et al. 2009]	Ontology proposal for the discovery of problems and solutions in sustainability science.	Layer	Manual	Prototype	Data	Discovery & Access
[Adams and Janowicz 2011]	An approach to create mapping ontologies to integrate geospatial databases using machine learning.	Multi & Tools	Loose	Prototype	Data	Access
[Adamusiak et al. 2011]	API for programming ontology search and integration.	Multi & Tools	Manual	API	Data	Access
[Deus et al. 2012]	API for integrating biological data. Programmatic rules to map to existing ontology.	Multi & Layer	Auto & Strict	API & Web	Data	Discovery & Access
[de la Villa et al. 2012]	A tool that automatically builds concept maps from medical knowledge bases.	Multi	Auto	Prototype	Data	Discovery & Access
[Chua and Kim 2012]	A system that annotates text to allow cross-ontology integration.	Multi	Auto	Prototype	Document	Access
[Thomas et al. 2013]	A search engine for genetics that automatically analyzes documents and maps their relationships.	Existing	Auto	Web	Document	Access
[Luo et al. 2013]	Dynamic semantic mapping between ontologies for Grid computing.	Multi & Layer	Auto	Grid	Data	Access
[Khattak et al. 2013]	A (semi)automatic crawler to build an ontology from document repositories that allows semantic search.	Existing & Layer	Auto & Loose	Grid	Document	Access
[Brinkley et al. 2013]	A mapping ontology that links specific phenotypes to genotypes for two different species. Allows to discover cause and effect of mutation/malformations.	Existing	Manual	Web	Data	Discovery & Access
[Muresan and Klavans 2013]	An approach for automatically building terminologies from health care text.	Multi	Auto	Web	Document	Discovery

the proposed mechanisms?". Several studies argued that, since their proposal was based on specific ontologies, changing the ontology would provide appropriate support to other domains. However, domain specificity hinders generality. The challenge here is to find an adequate compromise between domain-specific (and thus more limited) mechanisms, and generic (and thus potentially less effective) semantic search. There is thus an open challenge concerning the balance between a domain-specific and a generic semantic search.

Similar to "integration", there are different meanings for "Semantic Mapping". Figure 2 presents plots showing the evolution of these different meanings in the amount of selected studies over time.

Table VII. Detailed Studies of Extraction Phase (Cont. of Table VI)

Ref.	Descriptive Summary	Integration	Semantic Mapping	Software Architecture	Class	Objective
[Zheng et al. 2014]	A Middleware that adds semantic query capability for biomedical scientific data. Rules are written by the user.	Layer	Manual	Middleware	Data	Access
[Annane et al. 2016]	Proposes mapping between ontologies of different languages to allow multilingual queries.	Existing	Manual	Prototype	Data	Access
[Gil et al. 2017]	A distributed architecture for software artifact catalog with semantic search capability. Mapping is crowdsourced.	Multi & Layer	Manual	Web	Software	Access
[Portilla Herrera et al. 2017]	An architecture proposal (not functional) for semantic annotation of existing scientific documents to allow semantic search.	Layer	Auto	Prototype & Web	Document	Access
[Djokic-Petrovic et al. 2017]	A platform that allows integrated search in different bioinformatics data sources.	Multi	Manual	Web	Data	Access
[Jonquet et al. 2018]	Combines metadata annotations from 805 ontologies and 23 vocabularies to create an unified vocabulary for agronomy.	Multi	Manual	Web	Data	Access
[Omar et al. 2019]	Semantic mapping for Semantic web on top of relational databases.	Layer	Manual	Prototype & Web	Data	Access
[Seifer et al. 2019]	A semantic reasoning layer extension for Scala, which allows to integrate ontology search into programming.	Existing & Layer	Manual	API	Data & Software	Access
[Ahmed and Afzal 2020]	An approach to map sections from scientific articles to metadata, which is used by search engines.	Existing	Auto	Prototype	Citation & Document	Discovery & Access
[Sengloiluean and Khuntong 2020]	An approach for integration of learning resources stored in heterogeneous databases by using mapping ontologies.	Multi	Auto	Prototype	Document	Access

In general, “semantic mapping” refers to metadata fields added to the actual data with the purpose of enriching data with semantic information. Our work identified 50 selected studies concerning semantic mapping, and identified three categories of this mapping. The first category, corresponding to 24 selected studies, is the “Manual Definition”, in which metadata is manually specified by authors or curators. Since manual definition is time consuming, new approaches to automate these efforts were reported. A second category, which we named “Automatic Definition”, in which metadata is automatically added by software, presented in 23 selected studies. Automatic definitions present challenges – *e.g.*, when algorithms add incorrect metadata. As part of efforts to address this issue, researchers proposed what we name here “Loose Mechanisms”, which are variations of automatic metadata definitions. We identified in eight selected studies. Loose mechanisms include the usage of fuzzy logic to include loosely related results based on their metadata, which are often ordered by relevance; in another category, the “automatic definitions” are automatically linked to related results by matching equal metadata attributes. We highlight that fuzzy definitions may lack precision, i.e., they may not follow a strict formal definition. It is interesting that in this context, 3 studies have been identified advocating “Strict Definitions”. For example, the application of formal definitions to avoid ambiguity within the semantic search.

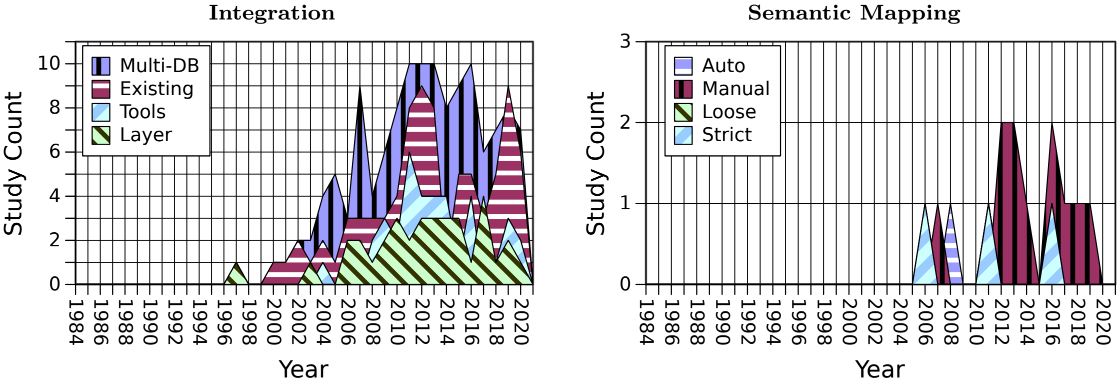

Fig. 2. Distribution of Integration and Semantic Mapping Meanings.

3.5 RQ2: Proposed Search Systems Characterization and Software Architectures

Different software architectures have been adopted while designing integrated semantic search engines. Figure 3 shows the evolution of software architectures, grouping whether the studies involve integration or semantic mapping. We identified 211 selected studies that propose an integrated implementation. Most of the studies (55 within selected) are based on multiple database composition, *i.e.*, the authors integrate several databases by implementing a single query system. We point out that multiple database systems are used interchangeably with integration concerns throughout these studies. This category of system appears in many situations, including large scale computing systems, *e.g.* clusters and grids, being slowly replaced by the emergence of services and cloud computing, which is represented by 37 and 13 selected studies, respectively. A total of 79 selected studies indicate the use of web-based systems, often advocating that this implementation is adequate for the mainstream community. We found several prototype proposals (reported by 40 papers) although we could not check the actual architectures of these prototypes because 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 seven studies reported this approach, which may indicate that this presents an implementation challenge to be followed up. We identified sensor network systems, represented by 12 selected studies. This category includes *things* from the Internet of Things, an increasing trend, as portrayed at the integration plot. This indicates a need for platforms that can cope with the rate of data capture provided by these sensors. Another recent trend, while only represented by 6 selected studies, is the emergence of artificial intelligence for both semantic mapping and integrated semantic search during the past year.

3.6 RQ3: Further Characterization of the Search Platforms

As part of the software architecture identification, we collected further details on how software was defined, including the metadata format and whether this format is standardized. We noted down the research field because it could affect its software domain. Metadata formats were summarized into seven categories, as presented in Figure 4. Ontology is the most referenced category (108 selected studies), which is not surprising because it is part of the search string. Knowledge models were mentioned in many of the selected studies (99), often surpassing the number of ontology studies in the analyzed publication years. This suggests these models were used for searching separately from ontologies. Linked data (including Linked Open Data) (20 selected studies) reached a peak in selected studies in 2013, but has been fading since then. Text corpora (31 selected studies) have been referenced in a broad range of years. Studies in this context often describe searches based on synonyms and natural language processing. Formal methods (18 selected studies) were explored since the oldest selected study from our analysis and continue being explored steadily through recent years. Besides

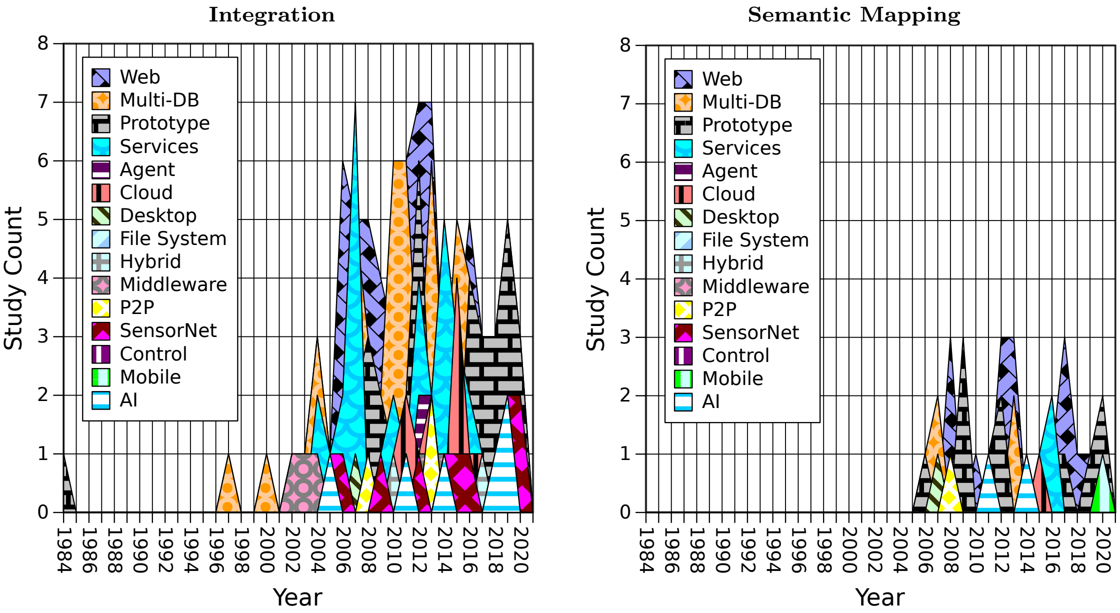

Fig. 3. Distribution of Employed Software Architecture Categories.

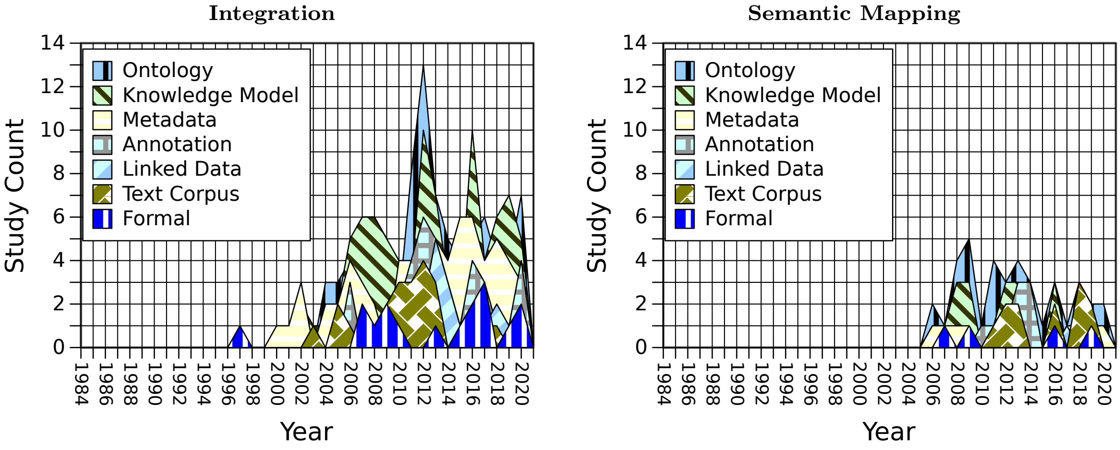

Fig. 4. Distribution of Scientific Metadata types.

these well defined categories, we found the use of annotations (45 selected studies) and metadata (68 selected studies), where the studies declared to use some kind of annotation or metadata, but they did not clarify how these metadata were formed.

Besides focusing on search systems, 166 selected studies mentioned scientific data tool proposals (cf. Figure 5). Most of these studies (107 selected studies) declare these proposals as analysis tools, while a considerable portion mentions creating tools to construct a semantic layer for existing data (35 selected studies). Crawler systems were reported, which focus on extracting data from external sources (34 selected studies). 43 selected studies proposed metric or ranking systems to compare data similarities. These studies are often related to recommender systems (16 selected studies).

We collected data domains of the reported software tools and platforms as areas, represented in Figure 6. We point out that it is not simple to categorize research topics (areas) because many

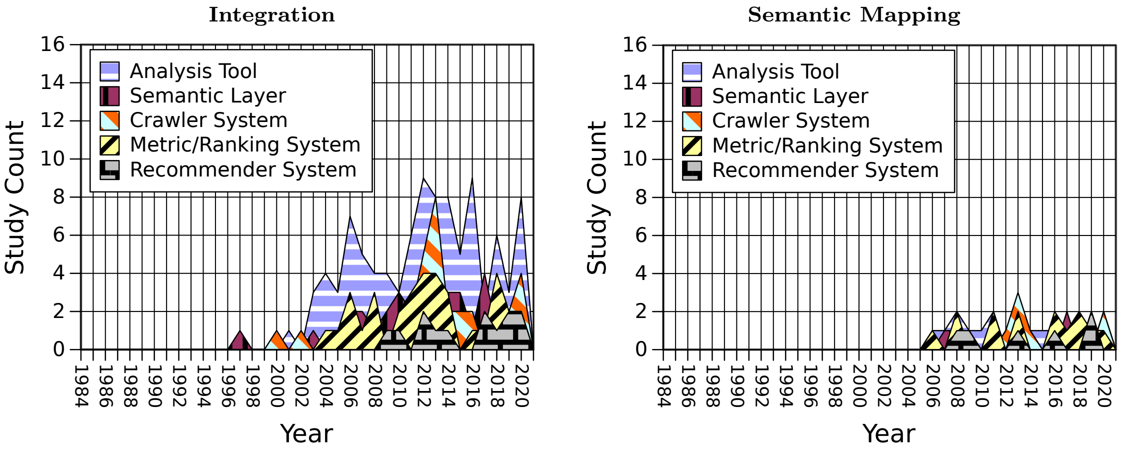

Fig. 5. Distribution of Scientific Data Tool Types.

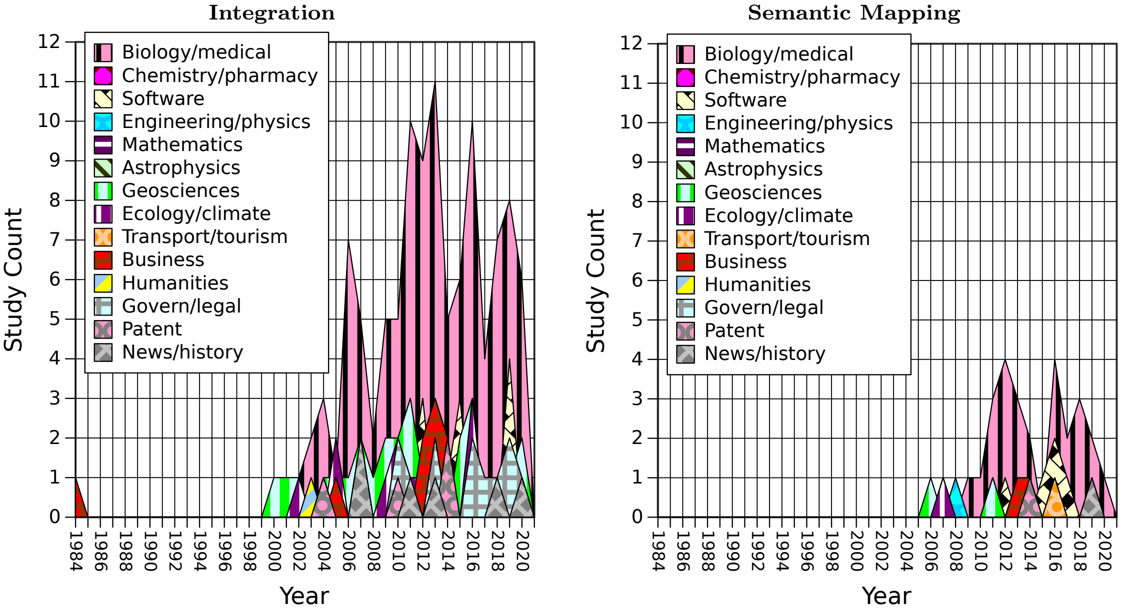

Fig. 6. Distribution of Publication Areas of the Selected Studies.

fields are closely related. For instance, it was not possible to completely separate general physics from engineering. "Software" involved studies reporting software repositories (18 selected studies), but the repositories could also store software for different domains. We identified that the majority of the selected studies was related to biology and medicine (122 selected studies). Geological and geographical data was also a common domain, with 26 selected studies. We noticed the increasing trend of selected studies (14 in total) that focus on government and legal domain.

3.7 RQ4: Objectives and Class Distributions

Figure 7 presents four dominant classes of search parameters declared in 254 selected studies, as follows: (a) Science Data: including text notes, spreadsheets, images, videos, recordings (195 selected studies); (b) Documents: including articles and theses (54 selected studies); (c) Processes: involving workflows;

methods, hypotheses, comparison metrics, software (56 selected studies); and Author names and their affiliations (11 selected studies). Though the latter is not directly included in the three Open Science axes, it is a frequent parameter of search mechanisms. Additional classes appear in a few selected studies, namely Citation (10), CFP (1), Conference (5) and Funding (1). Although not considered as part of Open Science axes, these classes were made available for specific search purposes. In particular, "Citation" goes through citation bases; "CFP" denotes that the mechanism was proposed to search for "call for papers"; "Conference" looks for data from conferences that have been held; "Funding" is interested in discovering funding agencies/organizations and "Institution" is concerned with finding specific institutions. Considering the total number of processes and software repositories, we identified that a subset is not for scientific software (18 selected studies). Our results indicate that most studies only focus on a single object class. Indeed, out of 254 studies, only 52 deal with more than one class and only three selected study involved more than two classes. In this sense, another research challenge is the design of (semantic) search mechanisms that allow combining distinct kinds of search parameters - documents, data, processes and authors.

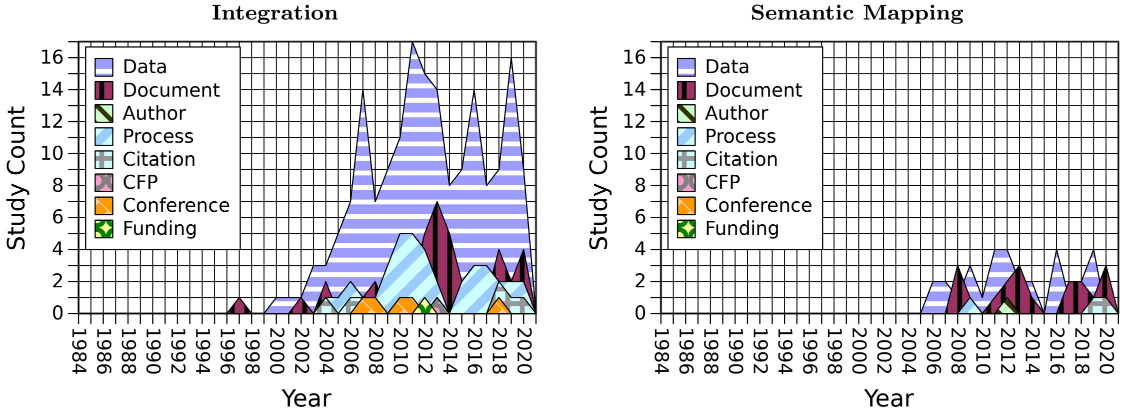

Fig. 7. Distribution of Scientific Data Classes.

There were 260 selected studies that declared explicit objectives for using the scientific data. Figure 8 presents these objectives. The most common reported objective is to access the stored data in addition to retrieving search results; and to notify users when new results appear (184 selected studies). 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 (123 selected studies). A slightly less frequent objective is data management, where existing data, documents and authors are registered to foster report creation (54 selected studies). A still less common objective found – 9 selected studies focused on simulations. They may be used, for instance, to generate data for experiments and observations, validate data or extrapolate findings.

Another selected study focused on using 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. 3 selected studies discussed reproduction or replication, where 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 summarize existing documents, aggregating their quantitative data and qualitative descriptions and comments into into a new (non-primary) study.

Table VIII shows the distinct objectives we identified for the mechanisms (rows), and the eight object classes (columns) - four of which are the most common, resulting in combinations that employ the class (subject) with the action for an objective (verb) – e.g., cell (1,1) indicates that there were 139 studies that were concerned with some sort of access to scientific data. Cells present the numbers

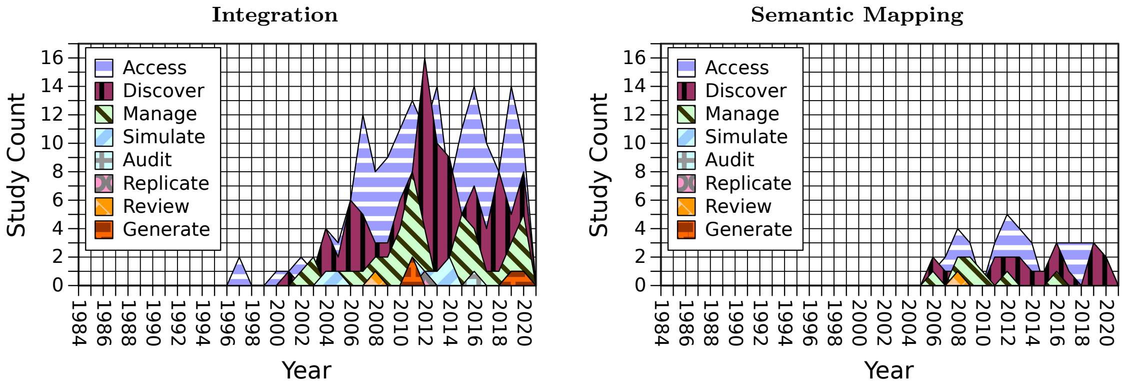

Fig. 8. Distribution of Scientific Data Objectives

grouped by integration and semantic mapping: e.g., this total of 139 studies is the result of the union of sets that contain 130 and 25 studies, which were selected for integration and semantic mapping, respectively. The table includes their frequency and descriptions.; Each cell 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.

4. DISCUSSION AND OPEN CHALLENGES

This section discusses our results addressing open opportunities and study limitations. Our analysis shows that some objectives have been relatively unexplored (cf. the numeric values in the cells of Table VIII). This indicates lines for open research on situations that would benefit from semantic search. Additional objectives were identified that are unrelated to semantic search – *e.g.*, those that mention issues s with "prediction" or means to export data. Prediction is considered part of "Discovery" studies (Row 2 of Table VIII), which were focused on prediction. In these cases, they "Predict" or estimate new data from existing data. Data export is an issue treated within the "Access" objective: a)retrieve workflow objects within packages; or b) access results of experiments; or c) take data results and explore them using software tools. "Teaching" is an objective associated with search and includes support to prepare learning material for students at all levels (for findability), and to help them find the material. "Visualization" combined to semantic search could lead to better comprehension for both the semantic queries as well as the results from the semantic search. Our literature analysis showed promising directions that can be further explored. For instance, our analysis of studies on data management tools, shows that they do not address the opportunity to include strategic decisions, using past scientific data to help researchers to plan future research efforts. Another gap in the reviewed investigations in our view is how to support the design of public "Policies" based on evidence to effectively aid in the creation of new edicts and bills that are backed by scientific production. We could not find any study on semantic search to extract specific data and metadata from "Internal" content available in documents and data, *e.g.*, article sections or images.

New infrastructures should be planned to support future requirements on semantic search. Regardless of how complex are those unexplored opportunities and data classes, open challenges could foster the design and implementation of new infrastructures. This can help novel research methodologies providing faster responses to open scientific questions as required by emerging topics.

We stress that all documents and raw data collected during this study are openly available, thereby supporting reproducibility, auditing, and extensions to our study [Gottardi 2021].

Table VIII. Objectives and Classes. Provided with Frequency, Descriptions and Resulting Combinations.

		Class							
		Scientific Data	Document	Process	Authors	Citation	CFP	Conference	Funding
Objective	Access	**139**= (130 ∪ 25): Search, query, access, recommend and/or retrieve science data.	**39**= (33 ∪ 13): Search, query, access, recommend and/or retrieve papers, articles, journals, reports, magazines, etc.	**36**= (34 ∪ 4): Search, access, recommend and/or retrieve scientific workflows.	**10**= (9 ∪ 1): Search and find or recommend authors and related authors.	**7**= (7 ∪ 1): Search for cited references and their counts.	**0**= (0 ∪ 0): Search for calls for papers.	**2**= (2 ∪ 0): Search or access past conference data, documents, and videos.	**1**= (1 ∪ 0): Search for funding calls.
	Discover	**93**= (84 ∪ 15): Discover conclusions using aggregated science data.	**25**= (21 ∪ 7): Discover conclusions and related documents using existing documents.	**28**= (27 ∪ 1): Discover combined workflows.	**4**= (4 ∪ 0): Discover what authors collaborate on research efforts.	**6**= (5 ∪ 2): Discover or suggest relevant citations for a given topic.	**1**= (1 ∪ 0): Discover or suggest relevant call for papers for a given topic.	**1**= (1 ∪ 0): Discover or suggest relevant conferences; suggest conference events.	**1**= (1 ∪ 0): Discover or suggest relevant funding calls on a given topic.
	Manage	**42**= (36 ∪ 6): Manage known science data, also their sources and bases.	**8**= (6 ∪ 2): Manage known document references/citations. Manage documents being written.	**17**= (14 ∪ 3): Manage known workflows and assess their usage.	**3**= (2 ∪ 1): Manage known authors, relationships, contributions and their roles.	**1**= (1 ∪ 0): Manage citations of documents and their relevance.	**0**= (0 ∪ 0): Manage calls for existing and future conferences.	**1**= (1 ∪ 0): Manage conference participants and schedule events.	**0**= (0 ∪ 0): Manage funding requests and finances.
	Simulate	**8**= (8 ∪ 0): Simulate experiments and compare against existing data for validation.	**0**= (0 ∪ 0): Simulate document publications and acceptance.	**3**= (3 ∪ 0): Simulate workflow usage and outcomes.	**0**= (0 ∪ 0): Simulate author contributions and outcomes.	**0**= (0 ∪ 0): Simulate citation bibliometrics based on models.	**0**= (0 ∪ 0): Simulate call for papers and its impact on paper submissions.	**0**= (0 ∪ 0): Simulate conferences and the attendance of participants.	**0**= (0 ∪ 0): Simulate funding calls and the outcome of its investment.
	Generate	**4**= (4 ∪ 0): Generate [synthetic] scientific data for using or testing software and workflows.	**0**= (0 ∪ 0): Generate scientific documents from data.	**0**= (0 ∪ 0): Generate custom scientific workflows.	**0**= (0 ∪ 0): Generate authors or pseudonyms.	**0**= (0 ∪ 0): Generate [synthetic] citation bibliometrics.	**0**= (0 ∪ 0): Generate call for papers for conferences.	**0**= (0 ∪ 0): Generate conference schedule and information	**0**= (0 ∪ 0): Generate funding calls and requests.
	Replicate	**3**= (3 ∪ 0): Replicate studies based on existing science data and compare the outcomes.	**0**= (0 ∪ 0): Replicate (or plagiate) existing documents and their structures.	**0**= (0 ∪ 0): Replicate existing workflows and compare their outcomes.	**1**= (1 ∪ 0): Imitate author roles.	**0**= (0 ∪ 0): Copy citations from similar studies.	**0**= (0 ∪ 0): Copy call for papers from past conferences.	**0**= (0 ∪ 0): Imitate similar successful conferences	**0**= (0 ∪ 0): Copy funding calls from similar requests.
	Audit	**1**= (1 ∪ 0): Audit data for validation and verification; protect from corruption and false data; blame manipulators.	**0**= (0 ∪ 0): Audit documents to verify authorship and protect documents from corruption.	**0**= (0 ∪ 0): Audit execution of workflows. Audit who can edit the workflow.	**0**= (0 ∪ 0): Audit roles and authorship to protect authors' curricula from corruption and false data.	**0**= (0 ∪ 0): Audit paper citations for legit citations and valid references.	**0**= (0 ∪ 0): Audit CFP from past conferences to compare the call versus outcome.	**0**= (0 ∪ 0): Audit review processes and schedules of conferences.	**0**= (0 ∪ 0): Audit funding reports and requests to verify accomplishments.
	Review	**0**= (0 ∪ 0): Review and compare data sets of science data to aggregate results.	**2**= (2 ∪ 1): Support for literature reviews, secondary studies.	**1**= (1 ∪ 0): Review workflows and methods and compare their efficiency.	**0**= (0 ∪ 0): Review existing author roles and contributions.	**1**= (1 ∪ 0): Review paper citations (snowballing)	**0**= (0 ∪ 0): Review call for papers and aggregate their requests.	**0**= (0 ∪ 0): Review conferences to write reports.	**0**= (0 ∪ 0): Review funding requests and scholarships to write reports.

4.1 Search Limitations

Number of search engines. Nine search engines were selected and invoked. To the best of our knowledge, this number is higher than usual for secondary studies in Computer Science. We point out that we left Google Scholar and DBLP out for the reasons described in Section 4. The first is not trustworthy for quantitative analyzes whereas the second does not make publication abstracts available. *Exclusion of partial documents.* Partial documents, e.g., book chapters, were excluded because they cannot simply be counted as complete studies. One could argue some chapters might also be considered complete studies. However, this would require specific metrics and duplicated verification that would create new issues with the review method. *Calibration of search strings.* Search strings were the same for all engines, and defined so as to collect a large number of studies. Our explored search strings were not changed per search engine to avoid biased results. It is possible that if slightly different strings were used per search engine, we could collect more relevant studies, but this modification would potentially cause search bias.

4.2 Study Selection and Data Extraction

Study selection and data extraction were performed by the first author, and all authors of this paper agreed on the method and protocol, which were strictly enforced. Data extraction was carefully executed considering all selected studies. Additional parameters might be used in the future to potentially aggregate more data from existing studies.

4.3 Threats to Validity

We describe some of the threats identified to the validity of our work, and describe the strategies employed to mitigate them. Validity is categorized in internal, external, construction and conclusion, with subtopics for each identified risk.

Internal Validity: This concerns "the extent to which the design and conduction of the study are likely to prevent systematic error" [Kitchenham and Charters 2007]. – Incomplete analysis. To mitigate this threat, all studies were verified three times according to the protocol. We recorded and made available all timings involving the review process for each document to support future audits, as part of the individual data collection Forms - cf. [Gottardi 2021]. – Errors in tools used for study selection and extraction. To mitigate this threat, we used tools and methods reported as reliable by related secondary studies (cf. Section 5). – Duplication of studies affecting quantitative analyses. To mitigate this threat, plagiarized (including self-plagiarized) studies were carefully linked to ensure that studies were only counted once during quantitative analyses regardless of whether they were repeatedly published. We excluded partial documents, e.g. chapters, because they do not represent a complete study and often represent small portions of a single study scattered in different small publications, which could affect our results. *External Validity:* This concerns the applicability of the results of the study outside the study itself [Kitchenham and Charters 2007]. – Representativeness of the set of collected studies and the search engines used. To mitigate the threat of missing relevant studies, nine search engines were searched completely, involving hundreds of thousands of documents.

Construction Validity - concerning the quality of the methodology adopted: – Suitability of protocol and method to perform a systematic review. To ensure the appropriateness of our review, we followed protocols and methods duly documented by literature specialized on systematic reviews.

Conclusions Validity: – Measure reliability (system and metrics used to count studies per category). The only metrics adopted was counting instances, common in meta-analysis studies. All results were produced automatically using a database system and a spreadsheet tool. – Insufficient sample size. First, all engines were invoked exhaustively to mitigate this risk. We recall that we excluded book chapters from our analysis to avoid considering non self-contained studies. Nevertheless, we provide the number of chapters and partial documents discarded for reference and auditing.

5. RELATED WORK

Related work concerns secondary studies on semantic search mechanisms. We identified a total of 107 secondary studies as part of our search results, of which 69 were selected during the selection phase. Within the latter, 24 were related to the context of semantic search. However, none of these related studies were conducted to answer the same questions as our secondary study.

From all secondary studies, those presenting some common elements to ours are the following. Xu *et al.* [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.* [Zhang et al. 2019] studied approaches to identify requirements for metadata search in the context of scientific data management. Additionally, Gustafsson *et al.* performed a secondary study to identify how Semantic Web Technologies have been employed to share knowledge among medical clinicians [Gustafsson et al. 2006]. Mulwad [Mulwad 2011] reviewed best practices for inferring meaning associated to data tables. The author argumented how these practices can be employed to be used in Semantic Web and Linked Open Data scenarios.

Karimi *et al.* [Karimi et al. 2019] analysed different approaches that employ thesauri and ontologies for semantic search. Additional examples of loosely related systematic reviews include Nguyen and Chowdhury [Nguyen and Chowdhury 2013], who performed a systematic review to create a knowledge map of digital libraries. Figueroa *et al.* [Figueroa et al. 2015] presented a review on the progress of linked data technology, while Havukkala [Havukkala 2009] wrote a discussion regarding solutions of semantic web retrieval for bio-technology, chemistry and related patents. In a similar context, Urdidiales, *et al.* [Urdidiales-Nieto et al. 2017] wrote a survey on the search of web services for the integration of biological databases. Gacitua *et al.* [Gacitua et al. 2019] provided a systematic review on semantic web technologies and discussed application to data warehouses or other industrial uses.

Comparison to previous versions. We compare this version of our paper with two previous publications of ours on the same issue – a preliminary study published as a short paper [Gottardi et al. 2020a], subsequently extended to a full workshop paper [Gottardi et al. 2020b]. These past versions followed the same methodology; also, they were focused on RQ4 because they were based upon a smaller amount of documents. The *major* differences are the following: 1) the present study thoroughly analyzes results concerning RQ1 through RQ4, whereas the previous papers concentrated on RQ4. 2) Our systematic mapping was updated throughout February 2021, including six new search engines. The present version analyzes 2054 unique studies as compared to 324 reported in the two previous versions. 3) There is now qualitative information on 27 extracted studies, as opposed to the 11 found in the previous versions. 4) It includes more data classes for the identified primary studies, e.g. article citations, conferences and funding. 5) It includes more data usage objectives for the identified primary studies, e.g. synthetic data generation. 6) We now include completely new analyses on the research field of the selected primary studies; in this work, we analyze the usage of scientific data in Legal, Government and Patent definition. 7) We now analyze software architectures and related techniques in our systematic review. 8) Another alternative meaning was added in the discussion of the word "integration" in studies (tool integration). 9) We further provided a differentiation between semantic mapping and integration in the quantitative analysis.

6. CONCLUSIONS

Open Science relies on collaboration through sharing research outcomes, usually grouped into three classes - articles, data and processes. Effective sharing presupposes findability, which requires understanding research efforts on search mechanisms – their goals, approaches and underlying mechanisms. We presented a systematic literature review on semantic search mechanisms applicable to scientific repositories. Our study analyzed and synthesized 2054 documents as a result of processing the entire contents of ACM Digital Library, arXiV, Engineering Village, IEEE Xplore, SBC OpenLib, Springer Link, Scopus, Wiley Online Library and Web of Science. We presented both quantitative and qualitative results, providing insights and pointing out open research issues to be addressed. The full set of results, detailed methodology, graphic plots and analysis datasets are provided in [Gottardi 2021]. Our analysis effectively confirms that most search mechanisms for open science results are based on finding publications, data or processes. There is also considerable research on author findability, and a few studies on other parameters, e.g. funding agencies or institutions. Our study can contribute to research in refining semantic search mechanisms to improve findability in the Open Science context.

REFERENCES

Adams, B. and Janowicz, K. Constructing geo-ontologies by reification of observation data. In *ACM GIS 2011*. GIS '11. ACM, New York, NY, USA, pp. 309–318, 2011. http://doi.org/10.1145/2093973.2094015.

Adamusiak, T., Burdett, T., Kurbatova, N., Joeri van der Velde, K., Abeygunawardena, N., Antonakaki, D., Kapushesky, M., Parkinson, H., and Swertz, M. A. Ontocat – simple ontology search and integration in java, r and rest/javascript. *BMC Bioinformatics* 12 (1): 218, May, 2011. http://doi.org/10.1186/1471-2105-12-218.

Ahmed, I. and Afzal, M. T. A systematic approach to map the research articles' sections to imrad. *IEEE Access* vol. 8, pp. 129359–129371, 2020. http://doi.org/10.1109/ACCESS.2020.3009021.

Annane, A., Emonet, V., Azouaou, F., and Jonquet, C. Multilingual mapping reconciliation between english-french biomedical ontologies. In *WIMS 2016*. ACM, 2016. http://doi.org/10.1145/2912845.2912847.

Brinkley, J., Borromeo, C., Clarkson, M., Cox, T., Cunningham, M., Detwiler, L., Heike, C., Hochheiser, H., Mejino, J., Travillian, R., and Shapiro, L. The ontology of craniofacial development and malformation for translational craniofacial research. *AJMG* 163 (4): 232–245, 2013. http://doi.org/10.1002/ajmg.c.31377.

Budak Arpinar, I., Sheth, A., Ramakrishnan, C., Lynn Usery, E., Azami, M., and Kwan, M.-P. Geospatial ontology development and semantic analytics. *Transactions in GIS* 10 (4): 551–575, 2006. http://doi.org/10.1111/j.1467-9671.2006.01012.x.

Chua, W. W. K. and Kim, J.-j. Semantic querying over knowledge in biomedical text corpora annotated with multiple ontologies. In *ACM BCB 2012*. BCB '12. ACM, New York, NY, USA, pp. 400–407, 2012. http://doi.org/10.1145/2382936.2382987.

de la Villa, M., Aparicio, F., Maña, M. J., and de Buenaga, M. A learning support tool with clinical cases based on concept maps and medical entity recognition. In *ACM IUI 2012*. ACM, New York, NY, USA, pp. 61–70, 2012. http://doi.org/10.1145/2166966.2166978.

Deus, H., Zhao, J., McCusker, J., Fox, R., Prud'hommeaux, E., Malone, J., Das, S., Miller, M., Adamusiak, T., Rocca Serra, P., and Marshall, M. Translating standards into practice - one semantic web api for gene expression. In *SWAT4LS '11*. ACM, London, UK, 2012. http://doi.org/10.1145/2166896.2166900.

Djokic-Petrovic, M., Cvjetkovic, V., Yang, J., Zivanovic, M., and Wild, D. J. Pibas fedsparql: a web-based platform for integration and exploration of bioinformatics datasets. *J. of Biomedical Semantics* 8 (1): 42, Sep, 2017. http://doi.org/10.1186/s13326-017-0151-z.

Figueroa, C., Vagliano, I., Rocha, O. R., and Morisio, M. A systematic literature review of linked data-based recommender systems. *CPE* 27 (17): 4659–4684, 2015. https://doi.org/10.1002/cpe.3449.

Gacitua, R., Mazon, J. N., and Cravero, A. Using semantic web technologies in the development of data warehouses: A systematic mapping. *WIREs Data Mining and Knowledge Discovery* 9 (3): e1293, 2019. http://doi.org/10.1002/widm.1293.

Gil, Y., Garijo, D., Mishra, S., and Ratnakar, V. Ontosoft: A distributed semantic registry for scientific software. In *e-Science*. IEEE, Baltimore, MD, USA, pp. 331–336, 2017. http://doi.org/10.1109/eScience.2016.7870916.

Gottardi, T. Support data and documentation on meta-analysis on semantic search, 2021. https://doi.org/10.25824/redu/89NUBJ.

Gottardi, T., Medeiros, C. B., and Reis, J. D. Semantic search on scientific repositories: A systematic literature review. In *SBBD 2020*. SBC, Salvador, Brazil, 2020a. http://doi.org/10.5753/sbbd.2020.13653.

Gottardi, T., Medeiros, C. B., and Reis, J. D. Understanding semantic search on scientific repositories: Steps towards meaningful findability. ZBMed, Athens, Greece, 2020b.

Gusenbauer, M. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118 (1): 177–214, Jan, 2019. https://doi.org/10.1007/s11192-018-2958-5.

Gusenbauer, M. and Haddaway, N. R. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods* 11 (2): 181–217, 2020. http://doi.org/10.1002/jrsm.1378.

Gustafsson, M., Falkman, G., Lindahl, F., and Torgersson, O. Enabling an online community for sharing oral medicine cases using semantic web technologies. In *The Semantic Web - ISWC 2006*. Springer, Berlin, Heidelberg, pp. 820–832, 2006.

Havukkala, I. Ontologies and semantic mining for bio-technology and chemistry data and patents. In *PaIR 2009*. PaIR '09. ACM, pp. 41–42, 2009. http://doi.org/10.1145/1651343.1651354.

Hummel, P., Braun, M., Tretter, M., and Dabrock, P. Data sovereignty: A review. *Big Data & Society* 8 (1): 1–17, 2021. http://doi.org/10.1177/2053951720982012.

Jonquet, C., Toulet, A., Dutta, B., and Emonet, V. Harnessing the power of unified metadata in an ontology repository: The case of agroportal. *Journal on Data Semantics* 7 (4): 191–221, Dec, 2018. http://doi.org/10.1007/s13740-018-0091-5.

Karimi, E., Babaei, M., and Beheshti, M. The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems. *JIPM* 34 (4): 1579–1606, 2019. https://jipm.irandoc.ac.ir/article-1-3463-en.html.

Khattak, A. M., Ahmad, N., Mustafa, J., Pervez, Z., Latif, K., and Lee, S. Y. Context-aware search in dynamic repositories of digital documents. In *IEEE CSE*. IEEE, Sydney, Australia, 2013. http://doi.org/10.1109/CSE.2013.59.

Kitchenham, B. and Charters, S. Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE 2007-001, Keele University and Durham University Joint Report, UK, 2007. http://www.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf.

Kraines, S. B. Active computer-mediated sharing and discovery of scientific knowledge through ontologies and logical inference. In *Knowledge Management*. Vol. 7. World Scientific, Singapore, pp. 195–206, 2008. http://doi.org/10.1142/9789812837578_0017.

Kumazawa, T., Saito, O., Kozaki, K., Matsui, T., and Mizoguchi, R. Toward knowledge structuring of sustainability science based on ontology engineering. *Sustainability Science* 4 (1): 99, Feb, 2009. http://doi.org/10.1007/s11625-008-0063-z.

Luo, Y., Yu, Z., Zhuang, Y., and Zheng, Z. Dynamic mapping processing between global ontology and local ontologies in grid environment. *ITJ* 12 (12): 2454–2459, 2013. http://doi.org/10.3923/itj.2013.2454.2459.

Mulwad, V. Dc proposal: Graphical models and probabilistic reasoning for generating linked data from tables. In *ISWC 2011*. Springer, Berlin, Heidelberg, pp. 317–324, 2011.

Muresan, S. and Klavans, J. L. Inducing terminologies from text: A case study for the consumer health domain. *ASI* 64 (4): 727–744, 2013. http://doi.org/10.1002/asi.22787.

Neri, M. A. Knowledge integration through semantic query rewriting. In *Proceedings of the 9th WSEAS International Conference on Applied Computer Science, ACS '09*. WSEAS, Morioka City, Iwate, Japan, pp. 229 – 234, 2009.

Nguyen, S. H. and Chowdhury, G. Interpreting the knowledge map of digital library research (1990–2010). *Journal of the ASIST* 64 (6): 1235–1258, 2013. http://doi.org/10.1002/asi.22830.

Oakley, A., Gough, D., Oliver, S., and Thomas, J. The politics of evidence and methodology: lessons from the eppi-centre. *Evidence & Policy* 1 (1): 5–32, 2005. http://doi.org/10.1332/1744264052703168.

Omar, Y. M. K., el Moneim, A. A., and Mohamed, K. Building a framework for mapping rdbms to rdf with semantic query capabilities. In *ITMS 2019*. IEEE, Riga, Latvia, pp. 1–5, 2019. http://doi.org/10.1109/ITMS47855.2019.8940733.

Pirrò, G., Ruffolo, M., and Talia, D. Advanced semantic search and retrieval in a collaborative peer-to-peer system. In *UPGRADE'08*. ACM, Boston, MA, USA, pp. 65–71, 2008. http://doi.org/10.1145/1384209.1384222.

Portilla Herrera, N. A., Gomez, F. L., Bucheli, V. A., and Pabón, O. S. Semantic annotation and retrieval of scientific documents in a big data environment. In *LACNEM 2017*. IET, Valparaiso, Chile, pp. 1–6, 2017. http://doi.org/10.1049/ic.2017.0032.

Seifer, P., Leinberger, M., Lämmel, R., and Staab, S. Semantic query integration with reason. *The Art, Science, and Engineering of Programming* 3 (3): 1–28, 2019. http://programming-journal.org/2019/3/13.

Sengloiluean, K. and Khuntong, R. Ontology-based semantic integration of heterogeneous data sources using ontology mapping approach. *J. of Theoretical and Applied Information Technology* 98 (22): 3489–3502, 2020. http://www.jatit.org/volumes/Vol98No22/13Vol98No22.pdf.

Thomas, P., Starlinger, J., and Leser, U. Experiences from developing the domain-specific entity search engine gene view. *Lecture Notes in Informatics (LNI)* vol. P-214, pp. 225–239, 2013.

Urdidiales-Nieto, D., Navas-Delgado, I., and Aldana-Montes, J. F. Biological web service repositories review. *Molecular Informatics* 36 (5-6): 1600035, 2017. http://doi.org/10.1002/minf.201600035.

Wilkinson, M., Dumontier, M., Aalbersberg, J., Appleton, G., and et al. The fair guiding principles for scientific data management and stewardship. *Nature Data* 3 (1): 1–9, 2016. http://doi.org/10.1038/sdata.2016.18.

Woelfle, M., Olliaro, P., and Todd, M. H. Open science is a research accelerator. *Nature Chemistry* vol. 3, pp. 745–748, October, 2011. http://doi.org/10.1038/nchem.1149.

Xiaoming, Z., Changjun, H., Qian, Z., and Chongchong, Z. Material scientific data integration for semantic grid. In *SKG 2007*. IEEE, Xi'an, China, pp. 414–417, 2007. http://doi.org/10.1109/SKG.2007.95.

Xu, H., Sun, L., Zou, M., and Meng, A. A survey of scientific metadata schema. *Applied Mechanics and Materials* vol. 411-414, pp. 349–352, 2013. http://www.scientific.net/AMM.411-414.349.

Zhang, W., Byna, S., Niu, C., and Chen, Y. Exploring metadata search essentials for scientific data management. In *2019 IEEE HiPC*. IEEE, Hyderabad, India, pp. 83–92, 2019. http://doi.org/10.1109/HiPC.2019.00021.

Zheng, S., Wang, F., and Lu, J. Enabling ontology based semantic queries in biomedical database systems. *International Journal of Semantic Computing* 8 (1): 67–83, 2014. http://doi.org/10.1142/S1793351X14500032.

Zhizhin, M., Kihn, E., Lyutsarev, V., Berezin, S., Poyda, A., Mishin, D., Medvedev, D., and Voitsekhovsky, D. Environmental scenario search and visualization. In *ACM GIS 2007*. GIS '07. ACM, New York, NY, USA, 2007. http://doi.org/10.1145/1341012.1341047.