Proceedings of the Federated Conference on Computer Science and Information Systems pp. 271–280
DOI: 10.15439/2017F88
ISSN 2300-5963 ACSIS, Vol. 11

Research on Proposals and Trends in the Architectures of Semantic Search Engines: A Systematic Literature Review

Jorge Morales
Escuela de Posgrado, Maestría en Informática,
Grupo de Investigación en Reconocimiento
de Patrones e Inteligencia Artificial Aplicada,
Pontificia Universidad Católica del Perú
Lima, Peru
Email: jorge.moralesv@pucp.pe

Andrés Melgar
Departamento de Ingeniería,
Sección de Ingeniería Informática,
Grupo de Investigación en Reconocimiento de
Patrones e Inteligencia Artificial Aplicada,
Pontificia Universidad Católica del Perú
Lima, Peru
Email: amelgar@pucp.edu.pe

***Abstract*—Semantic web technologies have gained some spotlight in recent years, mostly explained by the spread of mobile devices and broadband Internet access. As once envisioned by Tim Berners-Lee, semantic web technologies have fostered the development of standards that enable, in turn, the emergence of semantic search engines that give users the information they are looking for. This paper presents the results of a systematic literature review that focuses on understanding the proposals on the semantic search engines from an architectural point of view. From the results it is possible to say that most of the studies propose an integral solution for their users where their requirements, the context and the modules that comprise the search engine have a great role to play. Ontologies and knowledge also play an important role in these architectures as they evolve, enabling a great myriad of solutions that respond in a better way to the users' expectations.**

***Index Terms*—Semantic web, semantic search engines, ontologies, knowledge, knowledge representation, software architecture, systematic literature review.**

I. Introduction

SEMANTIC search is one of the hottest fields in recent years that have gained attraction. This is explained because search is one of the most used features in the Internet[1] and it is evolving in ways that can give users more meaningful data than before. We have witnessed the arrival of digital assistants on smartphones, tablets and computers, the presence of suggestions in social media, when buying online or when interacting with other people. These are proofs that what we search for, what our intentions are and how we like this information to be presented are becoming more important every day.

This panorama was envisioned by Tim Berners-Lee in 2001 [1], when the web was different and was starting to evolve from static contents to dynamic ones. Since then the Word Wide Web Consortium (W3C) has developed a myriad of standards in order to make that vision a reality. Several researches have been and are being carried out which demonstrate that semantics can be applied to search so that computer systems can understand the intentions, meanings and purpose of what the users want, and deliver the results they expect.

Nowadays we can see how search engines have improved their algorithms to make search results closer and useful to what users want to find. Google's Hummingbird algorithm, for instance, was developed to deal with the new needs of search, understanding the words typed by the user and returning meaningful results[2]. This kind of optimizations are also implemented in other search engines and products (e.g. Microsoft's Bing and Cortana digital assistant), and even traditional relational databases like SQL Server have some sort of semantic search capabilities built-in[3].

In the light of this current situation, it is important to understand how these applications are built, how they connect each other, and how they work to deliver what users are looking for. That is the main motivation for this research: to know and understand the architectures of these semantic search engines, how they look like and how they are changing our present.

This paper is organized as follows. Section II presents the literature review methodology. Section III presents the identification of the need for this study. Section IV presents the review protocol that this study followed. Section V presents the results obtained after the execution of the review protocol. Section VI discusses the findings of this study in order to give answer to the research questions. Finally, in section VII the conclusions and future work are discussed.

[1]Pew Research Center, "Search and email still top the list of most popular online activities", available at: http://www.pewinternet.org/2011/08/09/search-and-email-still-top-the-list-of-most-popular-online-activities/. [Online; accessed 18-July-2016]

[2]Danny Sullivan, "FAQ: All About The New Google "Hummingbird" Algorithm", available at: http://searchengineland.com/google-hummingbird-172816. [Online; accessed 10-December-2016]

[3]Microsoft Developer Network, "Semantic Search (SQL Server)", available at: https://msdn.microsoft.com/en-us/library/gg492075.aspx. [Online; accessed 10-December-2016]

II. Methodology

A systematic literature review is conducted as the methodology for this research to obtain the evidence needed to understand how the architectures of semantic search engines are formulated and how they work, which is the main objective of this research. As mentioned by Kitchenham in [2], "a systematic literature review (often referred to as a systematic review) is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest".

A systematic literature review has three main stages: planning the review, conducting the review and reporting the review [2]. There are several activities inside of those phases that involve iteration, especially those regarding the developing of the review protocol in the planning phase, or when conducting the review [2].

III. Identification of the need for a review

As stated in the introduction of this research, the main purpose is to identify how the architectures of semantic search engines have been and are being proposed and, as a result, a background will be constructed to summarize existing knowledge in this field and future research activities can be suggested [2].

There were no previous researches in the architecture of semantic search engines, as it is presented in subsection IV-B1. Therefore, the need to carry out this research was justified.

IV. Review protocol

One of the most important steps in any systematic literature review is the development of the review protocol. This protocol specifies the context of the review, the research questions, the criteria to use in the study selection, the quality assessment of the studies, the data extraction strategy, as well as the strategy to use when reporting the results.

A. Research questions

The research questions are subject to change while the review protocol is being developed [2]. As a result, the following questions went through several changes during the development of this systematic review. These research questions cover the main point of interest: to understand how a semantic search engine works and what the main building blocks are that allow them to work. With this in mind, the research questions are as follows:

- RQ1: What modules of the architecture of a semantic search engine are the most used across implementations?
- RQ2: What are the evaluation methods for validating and/or verifying the architecture of a semantic search engine?
- RQ3: What are requirements that an architecture of a semantic search engine needs to comply with?
- RQ4: What role do ontologies play in the architecture of a semantic search engine?
- RQ5: What role does knowledge play in the architecture of a semantic search engine?

TABLE I
Keywords identified from research questions

RQ1	module, architecture, semantic search engine, implementation
RQ2	evaluation, method, validation, verification, architecture, semantic search engine
RQ3	requirement, architecture, semantic search engine
RQ4	role, ontology, architecture, semantic search engine
RQ5	role, knowledge, architecture, semantic search engine

B. Search strategy

First a preliminary search was carried out, its purpose and results are presented here. Then the search terms are listed, as well as the query strings to be used. The search resources and the search process are explained afterwards. Finally, the search process documentation is mentioned.

1) Preliminary search: The preliminary search was carried out in Scopus to identify what the current studies looked like, what subjects they were talking about, and if there was any new point of interest that can be added to the research. The search string was as follows: *TITLE-ABS-KEY (architecture) AND TITLE-ABS-KEY("semantic search")*. This search was carried out on June 8th 2016, and 219 articles were found.

The first 100 articles, ordered by publication year, were picked. From these 100 articles, 55 of them were found to be related to the main subject of this study, whereas 36 were somewhat related. The other 9 articles were not related to the main subject at all. In order to classify the articles as related, somewhat related or not related, titles, abstracts and keywords were analyzed and compared to the main subject and the research questions presented in subsection IV-A.

From those 55 articles found to be related to the main subject of this study, the following concepts were found to be mentioned constantly in their abstracts and were added to the research questions:

- Ontologies, either as part of the architecture of a semantic search engine or as the core of the proposed engine. 23 articles were found to be related to ontologies.
- Knowledge, either as part of the architecture as a knowledge base or as a knowledge technique to be used in the proposed semantic search engine. 25 articles were found to be related to this concept.

2) Deriving search terms: As a first step, the search terms are derived from the research questions. In table I the main keywords are listed per each question.

In table II the synonyms for the keywords found in table I are presented. These synonyms were taken from the Thesaurus of the Oxford Dictionaries [4]. There were also words that were added because they were related to the first keywords - these words were identified in the exploratory search that was explained in subsection IV-B1.

[4]Thesaurus of the Oxford Dictionaries: http://www.oxforddictionaries.com/thesaurus/. [Online; accessed 6-July-2016]

TABLE II
SYNONYMS AND RELATED WORDS IDENTIFIED FOR KEYWORDS IN TABLE I

module	layer, component
evaluation	assessment, appraisal
method	procedure, technique, approach
architecture	system architecture
semantic search engine	semantic search system, semantic search platform, semantic search tool
implementation	implantation, application, approach
requirement	need, requisite

TABLE III
EXPRESSIONS TO USE FOR QUERIES

RQ1	(module OR layer OR component) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool)) AND (implementation OR application OR approach)
RQ2	(evaluation OR assessment OR appraisal) AND (method OR procedure OR technique OR approach) AND (validation OR verification) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
RQ3	(requirement OR need OR requisite) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
RQ4	role AND ontology AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
RQ5	role AND knowledge AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))

The expressions to use for querying the studies per each question are as stated in table III. These expressions were then customized according to the syntax of each search resource.

3) Search resources: The sources that are used for this research are the following: ACM Digital Library, IEEE Xplore, Scopus and ScienceDirect, as they have a broad set of articles in the computer science field.

4) Search process: Firstly, an initial, preliminary search was carried out in order to identify potential new terms that can enrich the keywords derived from the research questions. This also allows the identification of any new research question that can be of interest. This was presented and discussed in subsection IV-B1.

After that, a primary search phase is proposed to filter the articles found in the search resources. For this phase, duplicate articles are identified, and the inclusion and exclusion criteria are applied to the studies. These criteria will be applied to the title, abstract and keywords of each study. The criteria to be used for this phase are presented in subsection IV-C1. In case of ambiguity, the full text of the article is retrieved.

Lastly, a secondary search phase is proposed in order to identify the final articles that can answer the research questions. The full text of each article will be retrieved, the quality assessment criteria will be checked again, applying the inclusion/exclusion criteria as well as the quality assessment checklist presented in subsection IV-C2, paying special attention to the introduction, the architecture modules if applied, and the conclusions of each study. After this phase, the final articles will have been identified, ready to answer the research questions.

C. Study quality assessment criteria

The intention behind assessing study quality is to identify the primary studies that provide direct evidence about the research questions [2]. This quality assessment will be carried out to determine the relevance and identify reliable evidence of the selected studies to answer those questions.

In that way, the inclusion and exclusion criteria are presented in this section, as well as the quality assessment checklist to be used when selecting the studies in the search process. If the quality of a study does not satisfy the quality assessment criteria, it will be removed from the analysis given its weak evidence.

1) Inclusion/exclusion criteria for study selection: The inclusion and exclusion criteria, which can be refined during the search process, are defined in the systematic review protocol to minimize the bias effect that is likely to appear while conducting the review. For a study to be included in the systematic review, it will have to satisfy the first condition, and either the second, third or fourth conditions. In the case of studies for research questions 4 and 5, either the fifth or sixth condition must be fulfilled:

1) The study must be written in English.
2) The study proposes an architecture for a semantic search engine as a solution for a problem.
3) The study discusses about an architecture for a semantic search engine either in a conceptual or implemented way.
4) The study explains in greater or lesser detail the layers or modules the architecture includes.
5) In the case of a study that needs to answer RQ4, the study must provide an explanation of the role of the ontology within the architecture.
6) In the case of a study that needs to answer RQ5, the study must provide an explanation of the role of knowledge within the architecture.

The following exclusion criteria is meant to identify those studies that will not be included in the systematic review:

1) Those that do not focus on proposing an architecture of a semantic search engine, or where the semantic search engine is not the main subject in the study.
2) Those that do not include an explanation of the layers or modules that the architecture of a semantic search engine should have.
3) Those that are either books, conference proceedings, or secondary or tertiary studies.

2) Quality assessment checklist: Checklists are a way to assess the quality of the studies and therefore their importance as evidence to answer the research questions. They are also useful in order to decrease the effect of bias when reviewing

the studies [3]. Note that this assessment is in terms of relevance of evidence to answer the research questions and not to criticize the work of any researcher [3].

The questions for the following checklist are based on the ones presented in Zarour et al. [4] for their systematic review. Those were rephrased according to the needs of this research.

- QA1: Is the main subject of the study well defined?
- QA2: Is the presented architecture in the study clearly explained?
- QA3: Is the context where the study was carried out well described?
- QA4: Are the presented conclusions clearly stated?

QA1 is stated like this to identify whether the aims of the study are clearly defined. In QA2, the architecture of the semantic search engine explained by the study is analyzed to determine whether its purpose and components are presented clearly. QA3 is concerned with the background where the semantic search engine is working, so the architecture makes sense to the problem or situation that tries to solve or improve. Finally, QA4 considers the previous answers so the conclusions of the study are presented clearly and in line with the architecture and its context. Future work is also taken into account.

Each of the questions given in the checklist will be answered according to the following scale: Yes (1), No (0), Partially (0.5). In order to select a study, it needs to have a score greater than or equal to 3. This checklist will be applied to the results obtained after the primary search is carried out.

It is worth mentioning that the third question proposed by Zarour et al. about the threats to validity was not included because, from the preliminary search carried out before, there was no evidence of experimental or quantitative studies.

D. Data extraction strategy

After the primary studies have been selected and their quality assessed, the data will be extracted. The data extraction forms and the strategy to be adopted for recording the data are given in the sections below.

1) Data extraction form: Data extraction forms are meant to contain all the information that is necessary for answering the review questions and addressing the study quality criteria. The data extraction form for this systematic review is presented in table IV.

2) Data extraction procedures: In order to have a centralized storage for the execution of the review protocol and the extracted articles, a specialized software for systematic reviews was used. The name of this tool is StArt (State of the Art through Systematic Review), developed and maintained by the Laboratory of Research on Software Engineering (LaPES) that belongs to the Computing Department of the Federal University of São Carlos (DC/UFSCar) in Brazil[5]. It allows the management of the steps needed for carrying out a systematic review, giving a great support when executing the review protocol and searching for the articles.

[5]StArt (State of the Art through Systematic Review), available at: http://lapes.dc.ufscar.br/tools/start_tool/

TABLE IV
DATA EXTRACTION FORM

Field	Description	RQ
Id	Sequential number	General
Extraction date		General
Authors		General
Title		General
Study type	Journal article or a conference article	General
Search resource name	Name of the search resource where the study was found	General
Publication year		General
Institution	Researchers' institution or institutions	General
Country		General
Problem to be solved	Brief description of the main problem the architecture tries to solve	General
Architecture type	Whether is conceptual or concrete	General
Application field	Field where the architecture has been applied, or if it is a general purpose architecture	General
Architecture's modules	List of modules that the architecture is comprised	RQ1
Architecture's patterns applied	List of any pattern that the architecture applies	RQ1
Verification method	List of any methods used to verify the architecture	RQ2
Validation method	List of any methods used to validate the architecture	RQ2
Requirements	List of requirements the architecture fulfills	RQ3
Ontologies used	List of ontologies the architecture is using	RQ4
Ontology role	Brief description of the role the ontologies play within the architecture	RQ4
Knowledge role	Brief description of the role that knowledge plays within the architecture	RQ5

StArt allows the management of the articles found when retrieving them from the search resources. The review protocol is entered in this tool, including the inclusion and exclusion criteria, quality assessment checklist and the data extraction form fields provided in subsection IV-D1. With this information, and after the primary phase of the research is accomplished, the selected studies will be identified in the tool so that the secondary search phase can be carried out.

For the secondary search phase, the selected studies from the primary search phase are exported from StArt to an Excel file for further revision. As stated in subsection IV-B4, in this phase the inclusion/exclusion criteria and the quality assessment are applied, completing the respective columns in the Excel file. The resulting studies then will be used for answering the research questions. This way the data collected

TABLE V
NUMBER OF STUDIES FOUND PER RESEARCH QUESTION AND DIGITAL SOURCE

Research question	ACM	IEEE	Scopus	Science Direct	Total
RQ1	3	5	24	4	**36**
RQ2	2	0	0	0	**2**
RQ3	11	22	48	3	**84**
RQ4	2	5	8	0	**15**
RQ5	0	2	8	0	**10**
Total	**18**	**34**	**88**	**7**	**147**

from the studies is consolidated in one place, gathering both the extraction form questions and the quality assessment checklist questions.

V. EXECUTION

As mentioned in subsection IV-B4 about the search process, the first search phase comprises the identification of duplicates, as well as determine whether the articles found using the search queries presented in subsection IV-B2 can fulfill the inclusion and exclusion criteria. In this section, the documentation of the search process is presented, as well as any incidence or change that came up when executing the review protocol.

A. Searches in the search resources

The searches in each of the search resources were carried out from September 17th to October 4th. The years covered by the searches were from 2002 to 2016. In table V the number of studies per each search resource and per each research question is listed. These studies were the input to start the primary search.

147 studies were found in the search resources. The results obtained were exported in the BibTeX format, taking special attention to the authors, title, abstract and keywords fields when exporting the results. Other data, such as journal title or country, were selected if available in the search resource.

B. Primary search

In this phase, the studies found from the search resources are filtered based on the inclusion and exclusion criteria detailed in subsection IV-C1.

After being obtained from the search resources, the BibTeX files were imported into StArt and grouped by search resource. When a BibTeX file is imported, each of the studies specified in the file are analyzed by StArt in order to identify possible duplicates. Each duplicate found is then highlighted in blue across all of the previous results already imported in StArt. It is also possible to specify duplicates manually. This option was used after all BibTeX files were imported, ordered by title. 54 studies were found to be duplicates.

The next step after identifying duplicates was to read carefully the title, abstract and keywords of each of the studies, and apply the inclusion and exclusion criteria. This step took a while to accomplish because of the number of studies considered for the systematic review. 43 articles were rejected after applying the selection criteria. Table VI summarizes the previous steps.

TABLE VI
SUMMARY OF THE PRIMARY SEARCH

Search resource	Duplicates	Rejected	Accepted	Total found
ACM	5	9	4	18
IEEE	8	7	19	34
ScienceDirect	6	1	0	7
Scopus	35	26	27	88
Total	**54**	**43**	**50**	**147**

TABLE VII
SUMMARY OF THE SECONDARY SEARCH

Search resource	Accepted	Rejected	Unavailable	Total
ACM	3	1		4
IEEE	10	9		19
Scopus	16	5	6	27
Total	**29**	**15**	**6**	**50**

C. Secondary search

In this phase, the studies selected from the primary search phase are filtered out using the inclusion and exclusion criteria detailed in subsection IV-C1 and applying the quality assessment checklist presented on subsection IV-C2. This phase also helped modify the data extraction form fields in order to add or update them accordingly to any new point of interest that can help answer the research questions.

The 50 accepted studies found in the primary search, along with the duplicated and rejected studies, were exported to the Microsoft Excel format from StArt to continue with the secondary search phase. The duplicated studies were used to help identify the research question those studies had assigned, so that the selected studies can answer those research questions as well.

This phase took long to complete, starting from November 23rd 2016 to January 22nd 2017. This is because each study was reviewed thoroughly. Some studies were not available when this phase started until the authors kindly answered back with the full text of their studies after contacting them via email. Some authors were not available to contact and, as a result, their studies were not considered as part of this research. The results of this phase are presented in table VII.

The studies accepted after concluding the secondary search are presented in next list. Each study is presented with the research questions it needs to answer.

- [5], [6], [7], [8], [9] for answering RQ1 only
- [10], [11], [12] for answering RQ1, RQ3
- [13] for answering RQ1, RQ3, RQ4
- [14] for answering RQ1, RQ4
- [15] [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27] for answering RQ3 only

TABLE VIII
QUALITY ASSESSMENT CHECKLIST RESULTS

Id	Quality question	Yes	No	Partially
QA1	Is the main subject of the research well defined?	29	0	0
QA2	Is the presented architecture in the study clearly explained?	21	0	8
QA3	Is the context where the study was carried out well described?	22	0	7
QA4	Are the presented conclusions clearly stated?	18	0	11

- [28] for answering RQ3, RQ4, RQ5
- [29] for answering RQ4 only
- [30] [31], [32] for answering RQ4, RQ5
- [33] for answering RQ5 only

VI. SYNTHESIS AND ANALYSIS OF DATA

The following subsections present the most relevant data obtained from the selected studies. First, the general facts are analyzed, and then the research questions are answered.

A. General facts from selected studies

1) Quality assessment checklist results: The quality assessment checklist results are presented in table VIII. These show that the main subject of each study was found to be clearly stated. The second question tries to identified if the architecture was clearly presented and explained in the study; in this case, 8 studies were found to have gaps while explaining the architecture of the semantic search engine, or whether the architecture's modules were not explained thoroughly.

Third quality question checks whether the context is clearly presented and described, so that the semantic search engine can be a solution or propose a solution to resolve the problem identified. 7 studies were found that gave some light on the context for their architecture proposal, without making a deeply explanation of it. For the last quality question, it can be seen that a high number of studies presented not so well defined conclusions, mostly due to simple or obvious statements, mentioning previous concepts or lacking future works recommendations.

2) Application fields and problems to be solved: From the semantic search engines revised it was found that:

- 7 studies propose general purpose engines, that is, that can be applied to any field: [5], [6], [23], [19], [18], [24], [31].
- 3 studies propose solutions for digital documents: [15], [16], [22].
- 2 for medicine: [26], [28].
- 1 study for each of the following fields: military [7], distance education [8], multimedia content [9], biology [10], biomechanics [11], audiovisual content repositories [12], source code control systems [13], culture [14], education [17], web services [21], reporting [20], Russian museums [27], environment [25], transport services [29], academic library [30], innovation processes [32], and government to government cooperation [33].

Regarding the problems to be solved by the proposed semantic search engines, it can be mentioned, as the application fields, they are diverse in nature and they could not be categorized without discarding important details from them. However, most of the studies try to propose a solution for a previous unresolved problem, to propose a new alternative for users, or to improve search results.

3) Architecture types: The architecture type field is intended to identify whether the proposed architecture is conceptual or concrete. The former refers to the studies that formulates an architecture without implementing the actual semantic search engine whereas the latter refers to actual search engines implemented following the proposed architecture. From the results obtained, it can be mentioned that 21 studies proposed a concrete semantic search engine. The other 8 studies proposed a conceptual semantic search engine.

B. Answering research questions

In the next subsections the research questions are answered, as well as some discussing is added where needed. The fields used in the data extraction form, presented in table IV, are explained better, according to the results obtained during the secondary search phase.

1) RQ1: Modules most used by semantic search engines implementations: The aim of this research question is to identify what modules are the most used in the proposed architectures of semantic search engines. Although the word "implementations" can be understood as something that needs to be built or constructed, for this research is also covering conceptual architectures.

For this question, two fields were proposed in order to retrieve the data from the studies:

a **Architecture's modules**: this field aims to get the list of modules, components or tiers that the architecture has. If the study has an explanation of what the module is about, that is also taken into account.

b **Architecture's patterns applied**: this field aims to identify what architectural patterns are presented in the proposed architecture.

There are 10 studies found to answer this question. There are several modules identified that are common across the proposed architectures. These are listed as follows, highlighting the most relevant studies on each module found:

- **Extractor components**, such as crawlers used by [9] and [11], or extraction systems in [6], which navigates within raw data and store it for further processing. These can also make some sort of filtering, based on system's needs or requirements [13].
- **Storage support**, such as a database used by [9] and [8], an indexer used in [13] and [7], or tables as mentioned by [5], that can store the data and knowledge of the system. These storage elements are related to other key components, such as ontologies.

- **Reasoning components**, for example ontologies or inference engines. These are responsible for generating the answers based on the user queries and the knowledge stored in the systems. As it can be seen, usually the ontologies are customized for the field or domain where they will be applied (e.g. [10] and [14]). It is worth mentioning the work of Çelik et al. [8] where inference rules based on ontologies are proposed for the reasoning component of their semantic search engine for a Learning Management System (LMS).
- **User interfaces**, usually as web forms (e.g. [8]) where users formulates their search query. It is worth noticing the case of [14], which proposes a guided user interface, whereas others are plug-ins that need to be installed in another application in order to be available to the user [13].

In the case of the architecture patterns identified, the majority of the studies reported that a multitier (N tier) architecture was applied in the proposed solution. This leads to design the modules as layers that are loosely coupled, customized for a specific functionality. There are two specialized cases: in [5] a peer-to-peer design is proposed because of the distributed nature of the engine, where each node has its indexer and processes documents that are available for other nodes through web services. In [13], a client application (i.e. a plug-in for a developer's integrated development environment - IDE) is designed to be used by users, which displays the search results (mainly source code files).

2) RQ2: Evaluation methods for validating and/or verifying the architecture of a semantic search engine: The aim of this research question is to identify what evaluation methods exist for validating and verifying an architecture. In this case, validation is related to whether the system fulfills its requirements; verification is related to whether the system was developed right [34].

For this question, two fields were proposed in order to retrieve the data from the studies:

a **Verification method**: this field aims to get any verification method proposed by the study.
b **Validation method**: this field aims to get any validation method proposed by the study.

For this research question unfortunately there was no study found that fulfilled the search criteria and the quality assessment checklist. Even though there was no study identified, it can be said that not finding studies for this research question constitutes an opportunity for a future work. This is discussed further in the conclusions.

3) RQ3: Requirements an architecture of a semantic search engine complies with: The aim of this research question is to identify what kind of requirements an architecture needs to comply with. Although requirements are closed to the field or domain where the semantic search engine is working, the purpose of this research question is to identify any common underlying requirement that an architecture of a semantic search engine needs to fulfill independent of that field or domain.

For this question, one field was proposed in order to retrieve the data from the studies:

a **Requirements**: this field aims to get the list of requirements that the architecture of a semantic search engines needs to comply with. The requirements or needs were identified from the study, whether they were explicitly or implicitly mentioned.

This field is meant to gather both functional and non-functional requirements. This was done this way in order to understand the requirements in their context, and taking into account that usually requirements are not classified and presented in these two categories. That implied to read the full text of the selected studies thoroughly.

There are 18 studies found to answer this question. The following is a set of common requirements identified from those studies:

a **Precision on results**, mentioned by [13], [15], [10], [16], and in some way it is also mentioned by [21], [11], [23], [24]. This requirement is related to find the most relevant results based on the user's search query. The purpose of the semantic search is to improve results based on the user's intention and the context of the search query, so it does not come as a surprise that an architecture of a semantic search engine must have precision on results as one of its requirements.
b **Existence and maintenance of ontologies**, mentioned by [13], [16]. Although this is explicitly mentioned by few articles, it has a great impact because of the important role that ontologies play in an architecture (as it is described in subsection VI-B4). Almost all selected studies rely on ontologies to make the search engine work. Ontologies are the base to learn new concepts, share knowledge and make possible that search agents can retrieve information even when new concepts were not previously defined [35].
c **Usability**, mentioned by [13], [17], [23], [27]. This is concerned with how user-friendly users find the search engine, how easy it is to use and if it is accessible through common ways, such as smartphones and tablets. In the case of [27], richer representation takes a special meaning because of data the search engines needs to display, i.e. Russian museum art collections.
d **Evolution of knowledge base as new documents appears**, mentioned by [15], [16]. This is pretty close to the previous ontology-related requirement, as knowledge and ontologies are related. In this case, a knowledge base needs to accept new concepts as new information becomes available. In the case of [12], it is even proposed that the system should be able to cover various domain models.
e **Handle structured, unstructured and heterogeneous data sources**, mentioned by [15], [10], [16], [25], [26], [27]. This requirement is related to the diverse sources a semantic search needs to deal with. As shown in the work of Fernandez et al. [36], heterogeneous sources,

heterogeneous knowledge bases and heterogeneous ontologies can help getting answers for natural language queries, which are used in [26]. For structured and unstructured data, such as what we can find in the Web, crawlers and annotation mechanisms help coping with those, so semantic search can be done on those kinds of data [36].

f **Use of ontologies for suggesting or guiding the user search**, mentioned by [28], [18], [19], [11], [22], [23], [24]. This requirement is about having the help of ontologies while the user writes his/her query. This help can be presented as suggestions of additional or related terms [19], or it can use user's preferences in order to retrieve relevant results, as proposed by [24] or [18].

g **Use of natural language**, mentioned by [20], [21], [26]. This requirement is related to the usage of natural language queries that can express users' intentions in a much freer way. This implies that the search engine needs to process and translate the user's query properly, using techniques such as word-sense disambiguation. Then the query can be consumed by the domain ontologies so a match can be found against the knowledge base.

h **Handle large amount of data**, mentioned by [26]. This requirement, although mentioned by one study, is worth to be pointed out because new search engines will need to have a broader action range, such as in the Internet of Things as proposed in the work of Wang et al. [37]. However, the search engine proposed by Ślęzak et al. is oriented to the biomedical literature field, which is small when compared to other broader Internet-based solutions.

It is worth mentioning that, although is not stated literally on the previous requirements identified from the studies, for [12] having a decoupled system is important, as it keeps the engine core independent from the data and knowledge layers.

4) RQ4: The role of ontologies in the architecture of a semantic search engine: The aim of this research question is to identify what role ontologies play within the architecture of a semantic search engine. As it was seen in RQ1 and RQ3, ontologies have a strong presence in the proposed architectures. With this research question, what is sought is to unveil the functionalities ontologies perform.

For this question, two fields were proposed in order to retrieve the data from the studies:

a **Ontologies used**: this field aims to identify what kind of ontologies are proposed in the study.

b **Ontology role**: this field aims to get any description of the role that the ontologies perform within the proposed architecture.

There are 7 studies found to answer this question. The ontologies used and the ontology roles identified are as follows:

a **Domain ontologies are mostly used**, which seems to be a pattern across architectures. That is an expected scenario because a domain ontology can give specialized results and further customization, satisfying users' need in a better way. Even those that make use of general purpose ontologies (e.g. WordNet), as mentioned by Kerschberg et al. [31], at the end they resort to use domain ontologies in order to represent better user concepts, or to represent several domains within the same engine, as mentioned by [14].

b **Ontology roles are diverse**, but most of the selected architecture use them as a way to classify and express relationships among key concepts - that is the case of [29], [28], [31] and [32]. Two cases are special: in [13], Durão et al. mention that the domain ontology is used for reasoning processing, in order to identify relevant source code documents and suggest related terms to improve future user queries. In [30], although it is not further discussed, Jamgade and Karale mention that the domain ontology is used for building ontotriples (or ontology triples), a way to express concepts by a subject, a property and an object [38]. These ontotriples are then used for queried the knowledge base to retrieve the relevant documents.

5) RQ5: The role of knowledge in the architecture of a semantic search engine: The aim of this research question is to identify what role knowledge plays within the architecture of a semantic search engine. As it was seen in RQ1, RQ3 and RQ4, knowledge has a relevant role in the proposed architecture, mostly by means of a knowledge base. Most of the studies have already answered RQ4 before.

For this question, one field was proposed in order to retrieve the data from the studies:

a **Knowledge role**: this field aims to get any description of the role that knowledge performs within the proposed architecture.

There are 5 studies found to answer this question. The following are the aspects found in those studies:

a **Knowledge sharing should be a key feature**, so that the semantic search engines proposed in these studies should allow knowledge sharing by means of the domain ontology they have implemented, such as a document ontology [33] or an innovation process ontology [32].

b **Knowledge bases help getting better search results**, and in doing so ontologies play an important role. As it is already noted before, knowledge and ontologies usually work together in order to retrieve better and relevant results [30]. On the other hand, in [28], Mendonça et al. use a knowledge base to help on the document annotation process so that users can query those documents, and it can be used to help users creating their queries, which leads to get better search results.

c **Knowledge is gathered from heterogeneous sources**, which enriches the results a user can get. In order to accomplish this, a set of agents were proposed by Kerschberg et al. so that those diverse sources can be queried [31]. This takes into account the set of general

and domain specific ontologies the system uses, as already mentioned by that study.

VII. Conclusions and future work

The goal of this review is to identify how the architectures of semantic search engines work, how the proposals are designed and what problems they were and are solving. Most of the studies, as depicted previously, propose a concrete implementation for their architectures, so those systems were and are working now in a myriad of application fields. As there was no previous study that summarized this subject, no time range filter was set when searching for the studies.

It can be seen that most of the studies try to propose a solution for a previous unresolved problem, to design a new alternative for users, or to improve search results. To measure whether those proposals represent an improvement, some studies present comparison results or performance benchmarks as is the case of the work of Amanqui et al. [10], Thangaraj and Sujatha [22] or Dong et al. [29]. However, as the purpose of this systematic review is not related to that kind of experiments, this can be considered as a good starting point for a future work.

As for the modules that a semantic search engine comprises, it can be said that reasoning components such as ontologies and inference engines constitute key modules present across the studies. Domain ontologies in particular are a fundamental piece in a semantic search engine as they allow addressing the needs of a specific domain and user requirements [31]. The use of ontologies fosters reusability, as new concepts are identified and added to the ontology, making its maintenance crucial as it evolves over time [38].

Likewise, it was identified that ontologies and knowledge play together a key role in the architectures reviewed. One of the key roles for knowledge is knowledge sharing that is achievable through the implementation of the ontologies that search engines rely on [39]. This brings benefits to search results, improving search engines' precision and recall which, along with usability and the ability to handle unstructured and heterogeneous sources, constitutes some of the most important requirements that the architecture of a semantic search engine needs to fulfill as it is designed and developed.

Finally, although there was no validation and verification methods identified for the architectures of semantic search engines, this can be seen as an opportunity for future work by proposing validation and verification mechanisms already in use in other software engineering application fields. As mentioned by Abowd et al. in [40], there are many benefits of architectural evaluation methods, such as a better understanding and documentation of the system, clarification and prioritization of requirements, and early detection of problems in the architecture, which boosts architecture quality.

Acknowledgment

The authors of this review thank the support of the "Programa Nacional de Innovación para la Competitividad y Productividad", Peru, under the contract 124-PNICP-PIAP-2015.

References

[1] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 28–37, 2001. doi: 10.1038/scientificamerican0501-34. [Online]. Available: http://dx.doi.org/10.1038/scientificamerican0501-34

[2] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[3] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008. [Online]. Available: http://dx.doi.org/10.1002/9780470754887

[4] M. Zarour, A. Abran, J.-M. Desharnais, and A. Alarifi, "An investigation into the best practices for the successful design and implementation of lightweight software process assessment methods: A systematic literature review," *Journal of Systems and Software*, vol. 101, pp. 180 – 192, 2015. doi: 10.1016/j.jss.2014.11.041. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2014.11.041

[5] T. Luu, F. Klemm, I. Podnar, M. Rajman, and K. Aberer, "Alvis peers: A scalable full-text peer-to-peer retrieval engine," in *Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, ser. P2PIR '06. New York, NY, USA: ACM, 2006. doi: 10.1145/1183579.1183588. ISBN 1-59593-527-4 pp. 41–48. [Online]. Available: http://doi.acm.org/10.1145/1183579.1183588

[6] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker, "Searching and browsing linked data with swse: The semantic web search engine," *Journal of Web Semantics*, vol. 9, no. 4, pp. 365–401, 2011. doi: 10.1016/j.websem.2011.06.004. [Online]. Available: http://dx.doi.org/10.1016/j.websem.2011.06.004

[7] K. Schutte, F. Bomhof, G. Burghouts, J. Van Diggelen, P. Hiemstra, J. Van 'T Hof, W. Kraaij, H. Pasman, A. Smith, C. Versloot, and J. De Wit, "Goose: Semantic search on internet connected sensors," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8758, 2013. doi: 10.1117/12.2018112. [Online]. Available: http://dx.doi.org/10.1117/12.2018112

[8] D. Çelik, A. ElçI, and E. Elverici, "Finding suitable course material through a semantic search agent for learning management systems of distance education," *Proceedings - International Computer Software and Applications Conference*, pp. 386–391, 2011. doi: 10.1109/COMPSACW.2011.71. [Online]. Available: http://dx.doi.org/10.1109/COMPSACW.2011.71

[9] M. Ponnada and N. Sharda, "Model of a semantic web search engine for multimedia content retrieval," *Proceedings - 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007; 1st IEEE/ACIS International Workshop on e-Activity, IWEA 2007*, pp. 818–823, 2007. doi: 10.1109/ICIS.2007.135. [Online]. Available: http://dx.doi.org/10.1109/ICIS.2007.135

[10] F. K. Amanqui, K. J. Serique, S. D. Cardoso, J. L. D. Santos, A. Albuquerque, and D. A. Moreira, "Improving biodiversity data retrieval through semantic search and ontologies," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, vol. 1, Aug 2014. doi: 10.1109/WI-IAT.2014.44 pp. 274–281. [Online]. Available: http://dx.doi.org/10.1109/WI-IAT.2014.44

[11] T. Dao, T. Hoang, X. Ta, and M. Ho Ba Tho, "Knowledge-based personalized search engine for the web-based human musculoskeletal system resources (hmsr) in biomechanics," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 160–173, 2013. doi: 10.1016/j.jbi.2012.11.001. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2012.11.001

[12] T. Bürger and G. Güntner, "Smart content factory - semantic knowledge based indexing of audiovisual archives," *IET Seminar Digest*, vol. 2005, no. 11099, pp. 367–371, 2005. doi: 10.1049/ic.2005.0757. [Online]. Available: http://dx.doi.org/10.1049/ic.2005.0757

[13] F. A. Durão, T. A. Vanderlei, E. S. Almeida, and S. R. de L. Meira, "Applying a semantic layer in a source code search tool," in *Proceedings of the 2008 ACM Symposium on Applied Computing*, ser. SAC '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1363686.1363952. ISBN 978-1-59593-753-7 pp. 1151–1157. [Online]. Available: http://doi.acm.org/10.1145/1363686.1363952

[14] E. Borini, R. Damiano, V. Lombardo, and A. Pizzo, "Dramasearch. character-mediated search in cultural heritage," *Proceedings - 2009 2nd Conference on Human System Interactions, HSI '09*, pp. 554–561, 2009. doi: 10.1109/HSI.2009.5091038. [Online]. Available: http://dx.doi.org/10.1109/HSI.2009.5091038

[15] A. M. Khattak, J. Mustafa, N. Ahmed, K. Latif, and S. Khan, "Intelligent search in digital documents," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/WIIAT.2008.208. ISBN 978-0-7695-3496-1 pp. 558–561. [Online]. Available: http://dx.doi.org/10.1109/WIIAT.2008.208

[16] A. M. Khattak, N. Ahmad, J. Mustafa, Z. Pervez, K. Latif, and S. Y. Lee, "Context-aware search in dynamic repositories of digital documents," in *2013 IEEE 16th International Conference on Computational Science and Engineering*, Dec 2013. doi: 10.1109/CSE.2013.59 pp. 338–345. [Online]. Available: http://dx.doi.org/10.1109/CSE.2013.59

[17] D. Çelik, E. Elverici, A. Elçi, and N. Inan, "Educational activity finder for children with pervasive developmental disorder through a semantic search system," in *2012 IEEE 36th Annual Computer Software and Applications Conference*, July 2012. doi: 10.1109/COMPSAC.2012.84. ISSN 0730-3157 pp. 482–487. [Online]. Available: http://dx.doi.org/10.1109/COMPSAC.2012.84

[18] N. Guelfi, C. Pruski, and C. Reynaud, "Experimental assessment of the target adaptive ontology-based web search framework," in *2010 10th Annual International Conference on New Technologies of Distributed Systems (NOTERE)*, May 2010. doi: 10.1109/NOTERE.2010.5536622. ISSN 2162-1896 pp. 297–302. [Online]. Available: http://dx.doi.org/10.1109/NOTERE.2010.5536622

[19] S. Movva, R. Ramachandran, S. Graves, and H. Conover, "Customizable search engine with semantic and resource aggregation capability," in *2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*, July 2008. doi: 10.1109/CECandEEE.2008.115. ISSN 2378-1963 pp. 376–381. [Online]. Available: http://dx.doi.org/10.1109/CECandEEE.2008.115

[20] A. Vasilateanu, N. Goga, and A. Moldoveanu, "Semantic report search engine - questor," in *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*, Oct 2014. doi: 10.1109/ICSTCC.2014.6982404 pp. 134–139. [Online]. Available: http://dx.doi.org/10.1109/ICSTCC.2014.6982404

[21] M. E. Kholy and A. Elfatatry, "Intelligent broker a knowledge based approach for semantic web services discovery," in *Evaluation of Novel Approaches to Software Engineering (ENASE), 2015 International Conference on*, April 2015. doi: 10.5220/0005455300390044 pp. 39–44. [Online]. Available: http://dx.doi.org/10.5220/0005455300390044

[22] M. Thangaraj and G. Sujatha, "An architectural design for effective information retrieval in semantic web," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8225–8233, 2014. doi: 10.1016/j.eswa.2014.07.017. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.07.017

[23] S. Paiva, M. Cabrer, and A. Solla, "Precision: A semantic search guided-based system," *Systems Theory: Perspectives, Applications and Developments*, pp. 209–228, 2014.

[24] G. Besbes, H. Baazaoui-Zghal, and H. Ghezela, "Fuzzy ontology-based system for personalized information retrieval," *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pp. 286–293, 2014.

[25] U. Bügel, M. Schmieder, B. Schnebel, T. Schlachter, and R. Ebel, "Leveraging ontologies for environmental information systems," *IFIP Advances in Information and Communication Technology*, vol. 359 AICT, pp. 364–371, 2011. doi: 10.1007/978-3-642-22285-6_40. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-22285-6_40

[26] D. Ślęzak, A. Janusz, W. Świeboda, H. Nguyen, J. Bazan, and A. Skowron, "Semantic analytics of pubmed content," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7058 LNCS, pp. 63–74, 2011. doi: 10.1007/978-3-642-25364-5_7. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-25364-5_7

[27] D. Mouromtsev, P. Haase, E. Cherny, D. Pavlov, A. Andreev, and A. Spiridonova, "Towards the russian linked culture cloud: Data enrichment and publishing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9088, pp. 637–651, 2015. doi: 10.1007/978-3-319-18818-8_39. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18818-8_39

[28] R. Mendonça, A. F. Rosa, J. L. Oliveira, and A. Teixeira, "Ontology-based health information search: Application to the neurological disease domain," in *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, June 2013. ISSN 2166-0727 pp. 1–6.

[29] H. Dong, F. K. Hussain, and E. Chang, "A service search engine for the industrial digital ecosystems," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 6, pp. 2183–2196, June 2011. doi: 10.1109/TIE.2009.2031186. [Online]. Available: http://dx.doi.org/10.1109/TIE.2009.2031186

[30] A. N. Jamgade and S. J. Karale, "Ontology based information retrieval system for academic library," in *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*, March 2015. doi: 10.1109/ICIIECS.2015.7193106 pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICIIECS.2015.7193106

[31] L. Kerschberg, H. Jeong, Y. Song, and W. Kim, "A case-based framework for collaborative semantic search in knowledge sifter," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4626 LNAI, pp. 16–30, 2007. doi: 10.1007/978-3-540-74141-1_2. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74141-1_2

[32] M. Jurczyk-Bunkowska and I. Pawełoszek, "The concept of semantic system for supporting planning of innovation processes [koncepcja semantycznego system wspomagania planowania procesów innowacji]," *Polish Journal of Management Studies*, vol. 11, no. 1, pp. 79–89, 2015.

[33] F. Corradini, F. De Angelis, F. Paoloni, A. Polzonetti, and B. Re, "A case study of a semantic search engine for g2g collaboration based on intelligent documents," *Proceedings of 4th International Conference on e-Government, ICEG 2008*, pp. 499–506, 2008.

[34] B. Boehm, "Software risk management," in *European Software Engineering Conference*. Springer, 1989. doi: 10.1007/3-540-51635-2_29 pp. 1–19. [Online]. Available: http://dx.doi.org/10.1007/3-540-51635-2_29

[35] M. Uschold, "Where are the semantics in the semantic web?" *AI Mag.*, vol. 24, no. 3, pp. 25–36, Sep. 2003.

[36] M. Fernandez, V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells, "Semantic search meets the web," in *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, ser. ICSC '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/ICSC.2008.52. ISBN 978-0-7695-3279-0 pp. 253–260. [Online]. Available: http://dx.doi.org/10.1109/ICSC.2008.52

[37] W. Wang, S. De, G. Cassar, and K. Moessner, "Knowledge representation in the internet of things: Semantic modelling and its applications," *Automatika – Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 54, no. 4, pp. 388 – 400, October 2013. doi: 10.7305/automatika.54-4.414. [Online]. Available: http://dx.doi.org/10.7305/automatika.54-4.414

[38] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006. [Online]. Available: http://dx.doi.org/10.1007/b97353

[39] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993. doi: 10.1006/knac.1993.1008. [Online]. Available: http://dx.doi.org/10.1006/knac.1993.1008

[40] G. Abowd, L. Bass, P. Clements, R. Kazman, and L. Northrop, "Recommended best industrial practice for software architecture evaluation." DTIC Document, Tech. Rep., 1997.