Undefined 0 (2020) 1–0
IOS Press


# A decade of Semantic Web research through the lenses of a mixed methods approach

**Editor(s):** Christoph Schlieder, University of Bamberg, Germany
**Solicited review(s):** Yingjie Hu, University of California, Santa Barbara, USA; two anonymous reviewers

Sabrina Kirrane [a], Marta Sabou [b], Javier D. Fernández [a], Francesco Osborne [c], Cécile Robin [d], Paul Buitelaar [d], Enrico Motta [c], and Axel Polleres [a]

[a] *Vienna University of Economics and Business, Austria*
*E-mail: firstname.lastname@wu.ac.at*
[b] *Vienna University of Technology, Austria*
*E-mail:firstname.lastname@ifs.tuwien.ac.at*
[c] *Knowledge Media institute (KMi), The Open University, UK*
*E-mail:firstname.lastname@open.ac.uk*
[d] *Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland*
*E-mail: firstname.lastname@insight-centre.org*

**Abstract.** The identification of research topics and trends is an important scientometric activity, as it can help guide the direction of future research. In the Semantic Web area, initially topic and trend detection was primarily performed through qualitative, *top-down* style approaches, that rely on expert knowledge. More recently, data-driven, *bottom-up* approaches have been proposed that offer a quantitative analysis of the evolution of a research domain. In this paper, we aim to provide a broader and more complete picture of Semantic Web topics and trends by adopting a *mixed methods* methodology, which allows for the combined use of both qualitative and quantitative approaches. Concretely, we build on a qualitative analysis of the main seminal papers, which adopt a top-down approach, and on quantitative results derived with three bottom-up data-driven approaches (Rexplore, Saffron, PoolParty), on a corpus of Semantic Web papers published between 2006 and 2015. In this process, we both use the latter for "fact-checking" on the former and also to derive key findings in relation to the strengths and weaknesses of top-down and bottom-up approaches to research topic identification. Although we provide a detailed study on the past decade of Semantic Web research, the findings and the methodology are relevant not only for our community but beyond the area of the Semantic Web to other research fields as well.

Keywords: Research Topics, Research Trends, Linked Data, Semantic Web, Scientometrics

## 1. Introduction

The term scientometrics is an all encompassing term used for an emerging field of research that analyses and measures science, technology research and innovation [21]. Although the term scientometrics is a broad term, in this paper, we focus on one particular sub field of scientometrics that uses topic analysis to identify trends in a scientific domain over time [17]. Understanding topics and subsequently predicting trends in research domains are important tasks for researchers and represent vital functions in the life of a research community. Overviews of present and past topics and trends provide important lessons of how research interests evolve and allow research communities to better plan future work, whereas visions of future topics can inspire and channel the work of a research community.

Considering the critical role played by topic and trend analysis when it comes to identifying underrepresented and emerging research topics, it is not surprising that there have been a number of works from

0000-0000/20/$00.00 © 2020 – IOS Press and the authors. All rights reserved

Semantic Web researchers that take an introspective view of the community. Several papers endeavor to predict Semantic Web research topics and trends [1,2], or as the research advanced over the years, to analyse topics and trends within the community [15,19]. In parallel, several researchers [5,22,27,30,31,34] are actively working on tools and techniques that can be used to automatically uncover research topics and trends from scientific publications.

Most of the trend prediction/analysis papers in the Semantic Web area [1,2,15] adopt a *top-down* approach that primarily relies on the knowledge, intuition and insights of experts in the field. While undoubtedly these are very valuable assets, trend-papers that purely follow this approach risk focusing on major topics and trends alone while overlooking under-represented or emerging topics and trends. These shortcomings could potentially be addressed by (semi-) automatic, data-driven approaches, which identify research topics and trends in a *bottom-up* fashion from large corpora.

The primary goal of this paper is to provide a more complete picture of Semantic Web topics and trends in the last decade by relying on both top-down and bottom-up approaches. Our hypothesis being that *there is a high correlation between expert driven and data driven topic and trend analyses, however by combining both approaches it is possible to gain additional, valuable insights with respect to the Semantic Web research domain*. Starting from this hypothesis, we devise two primary research questions:

(1) Is it possible to identify the predominant Semantic Web research topics using both expert based predictions and topic and trend identification tools?
(2) What are the strengths and weaknesses of expert-driven and data-driven topic and trend identification methods?

In order to answer the aforementioned research questions we adopt a *mixed methods* research methodology [25], which involves the combination of quantitative and qualitative research methods, in order to gain better insights into Semantic Web topics and trends. Concretely, our study comprises three core tasks.

- Firstly, in a qualitative study we converge the findings of three top-down style seminal papers [1,2,15] at different points in time, into a unified *Research Landscape*.
- Secondly, we employ three alternative data-driven quantitative approaches in order to uncover topics and trends from a corpus of Semantic Web publications in a bottom-up fashion.
- Thirdly, we compare and contrast the topics derived from both the expert analysis and the data driven approaches, in order to provide a more holistic picture of Semantic Web research.

In order to enable the Semantic Web community to further build upon the results of our study, additional information about the resources described in this paper are available via `https://doi.org/10.5281/zenodo.1492693`.

The remainder of the paper is structured as follows: *Section* 2 provides an overview of existing work on automatic topic and trend analysis in the Semantic Web community. *Section* 3 describes the mixed methods methodology that guided our analysis. *Section* 4 provides a snapshot of the Semantic Web research community based on the observations of several domain specific experts [1,2,15]. This is followed by the presentation of the topic analysis of papers published in the main Semantic Web publishing venues over a 10 year period from 2006 to 2015 in *Section* 5. A discussion on the findings of our analysis is presented in *Section* 6. Finally, *Section* 7 concludes the paper and presents directions for future work.

## 2. Related Work

The analysis presented in this paper is situated within the field of *Scientometrics*, defined by Leydesdorff and Milojević [26] as the *"quantitative study of science, communication in science, and science policy"*. Although this research field is closely related to *Bibliometrics* (i.e., the application of statistical methods to books and other media of communication), and *Informetrics* (i.e., the study of the information phenomena), these terms are not necessarily synonymous [21]. In this section, we examine approaches for detecting and analyzing research topics, as a specific task within the Scientometrics landscape, with a primary focus on the contributions from the Semantic Web community.

Detecting topics that accurately represent a collection of documents is an important task that has attracted considerable attention in recent years leading to a variety of relevant approaches from different media sources, such as news articles [12], social networks [7], blogs [29], emails [28], to name but a few. A classical way to model the topics of a document is to extract a list of significant terms [6] (e.g., using tf-idf) and to cluster them [39]. Another common solution

is the adoption of probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [3] or Probabilistic Latent Semantic Analysis (pLSA) [20]. However, these generic approaches suffer from a number of limitations that often hinder their application for the task of detecting scientific topics. Firstly, they produce unlabeled bags of words that are often difficult to associate with distinct research areas. Secondly, the number of topics to be extracted needs to be known a priori. Finally, using such methods it is not possible to distinguish research areas from other kinds of topics contained in a document.

Therefore, several approaches were proposed to specifically address the problem of detecting research topics. For instance, Morinaga et al [28] present a method that exploits a Finite Mixture Model to detect research topics and to track the emergence of new topics. Derek et al [13] developed an approach that matches scientific articles with a manually curated taxonomy of topics that is used to analyse topics across different timescales. Chavalarias et al [8] propose a tool known as CorText that can be used to extract a list of n-grams from scientific literature and to perform clustering analysis in order to discover patterns in the evolution of scientific knowledge.

Topics can also be identified and analyzed with methods for bibliometric mapping, which focus on generating spatial representations of the interaction between disciplines, papers, and authors. In the last years we saw the emergence of several relevant tools, which leverage a variety of techniques, such as bibliographic coupling and co-author, co-citation, and co-word analysis. CiteSpace [9] is a long running application for identifying trends and patterns in scientific literature that can identify emerging topics by combining co-citation analysis and burst detection [24]. SciMAT[11] is an advanced science mapping analysis tool that incorporates several algorithms and measures and covers all the steps in the bibliometric mapping workflow. VOSViewer [41] is another well-known software for constructing and analyzing bibliographic networks. Jo et al [23] present a relevant approach that detects topics by combining distributions of terms with the citation graph related to publications containing these terms. A detailed comparison of several such tools can be found in [10].

Public tools for the exploration of research data usually identify research areas by using keywords as proxies (e.g., DBLP++ [14], Scival[1]), adopting probabilistic topic models (e.g., aMiner [40]) or exploiting handcrafted classifications (ACM[2], Microsoft Academic Search[3]).

However, all these solutions suffer from some limitations. For example, keywords are unstructured and usually noisy, since they include terms that are not research topics. In addition, the quality of keywords assigned to a paper varies a lot according to the authors and the venues. Probabilistic topic models produce bags of words that are often not easy to map to commonly known research areas within the community. Finally, handcrafted classifications are expensive to build, requiring multiple expertise, and tend to age very quickly, especially in a rapidly evolving field such as Computer Science.

The Semantic Web Community has also produced a number of tools and techniques that use semantic technologies for detecting and analyzing research topics. For instance, Bordea and Buitelaar [5] demonstrate how expertise topics extraction (with ranking and filtering) along with researcher relevance scoring can be used to build expert profiles for the task of expert finding. In a related work, Monaghan et al. [27] present their expertise finding platform Saffron based on the same principles, and demonstrate how it can be used to link expertise topics, researchers and publications, based on their analysis of the Semantic Web Dog Food (SWDF) corpus. The data is further enhanced with URIs and expertise topic descriptions from DBpedia and related information from the Linked Open Data (LOD) cloud. An alternative approach is adopted by the Rexplore system [31], an environment for exploring and making sense of scholarly data that integrates statistical analysis, semantic technologies, and visual analytics. Rexplore builds on Klink-2 [30], an algorithm which combines semantic technologies, machine learning and knowledge from external sources (e.g., the LOD cloud, web pages, calls for papers) to automatically generate large-scale ontologies of research areas. The resulting ontology is used to semantically enhance a variety of data mining and information extraction techniques, and to improve search and visual analytics. Hu et al. [22] demonstrate how Semantic Web technologies can be used in order to support scientometrics over articles and data submitted to the Semantic Web Journal as part of their open review process. Towards this end the authors provide external ac-

[1] http://www.elsevier.com/solutions/scival

[2] https://www.acm.org/publications/class-2012

[3] http://academic.research.microsoft.com/

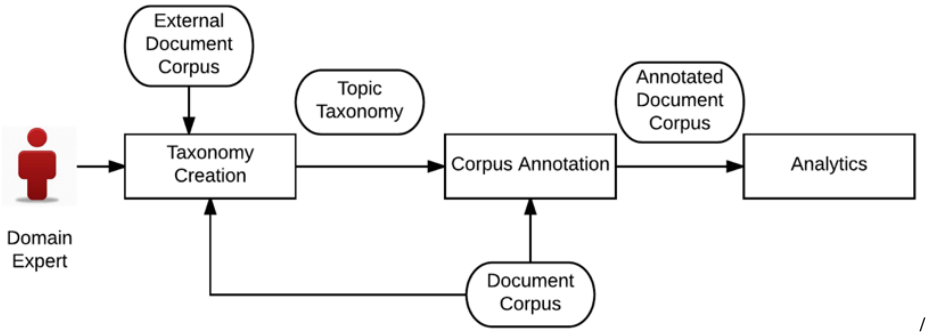



Fig. 1. Conceptual overview of topics detection approaches: main steps and data sources

cess to their semantified dataset, which is also linked to external datasets such as DBpedia and the Semantic Web Dog Food corpus. On top of this data they provide several interactive visualizations that can be used to explore the data, ranging from general statistics to depicting collaborative networks. Whereas Parinov and Kogalovsky [34] describe the Socionet research information system that focuses on linking research objects in general and research outputs in particular, the authors argue that information inferred from the semantic linkage of research objects and actors can be used to derive new scientometric metrics.

An interesting case of data-driven analysis is that reported in Glimm and Stuckenschmidt [19], looking back at the last 15 years of Semantic Web research through the lens of papers published at ISWC conferences from 2002 to 2014. The authors adopt an empirical approach to better understand the topics and trends within the Semantic Web community, in which they identify 12 key topics that describe Semantic Web research and then manually classify papers published in ISWC conference proceedings according to these topics. This work can also be categorized as a data-driven analysis of research topics and trends, which was performed completely manually.

Although data-driven approaches have been evaluated on their own, to date there is a lack of works that compare and contrast existing approaches, or indeed evaluate them with respect to expert-driven approaches. This paper fills this gap by adopting a holistic approach to topic and trend analysis, by analyzing the results of three expert-based and data-driven topic-detection approaches in the context of Semantic Web research.

## 3. Background and Methodology

In order to gain a better understanding of the topics and trends in the Semantic Web community over a ten year period from 2006-2015, we adopt a *mixed methods* approach to topic extraction and analysis, which combines both expert-based and data-driven approaches. According to Leech and Onwuegbuzie [25], the mixed methods research methodology involves the combination of quantitative and qualitative research methods in order to gain knowledge about some phenomenon under investigation. The mixed methods approach that guided the work carried out in this study is illustrated in *Figure* 2.

### *3.1. Seminal paper qualitative topic analysis*

The goal of the qualitative analysis of the seminal papers was primarily to identify research topics mentioned in [1,2,15]. The work was conducted in a two step process. In `Step 1` each paper was read by three of the authors of this paper who were each tasked with identifying technical research topics mentioned in the three seminal papers (e.g., ontology, OWL). To keep the analysis as objective as possible, the authors extracted the exact wording used in the papers instead of using synonyms more familiar to them. Following on from this, the authors grouped extracted keywords into broader topic areas (e.g., ontologies and modeling, logic and reasoning). In order to reduce any bias, in `Step 2` the results of the aforementioned analysis were discussed and aligned during a consensus workshop. Where disagreement occurred with respect to the grouping of keywords the seminal papers were consulted in order to better understand the context of the topic, such that it was possible to reach consensus as to its categorization. The final outcome of the qualitative analysis is the unified *Research Landscape*, shown in *Table* 2 and discussed in detail in *Section* 4.

### 3.2. Semantic Web publications quantitative topic analysis

Rather than using a single topic and trend identification tool in `Step 3` we elected to perform the analysis of a corpus of Semantic Web publications with three different tools (i.e., PoolParty[4], Rexplore[5], and Saffron[6]), such that we could compare and contrast the results obtained via the different tools.

*Semantic Web Venues (SWVs) corpus:* The corpus, which was analyzed by each of the tools, comprises papers from five enduring international publishing venues dedicated to Semantic Web research, namely: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), the SEMANTiCS conference, the Semantic Web Journal (SWJ) and the Journal of Web Semantics (JWS), over a 10 year period from 2006 to 2015 inclusive. These publishing venues were chosen as they are dedicated to Semantic Web research and have been running continuously for several years. Although, the SEMANTiCS conference was traditionally seen as a more business oriented event, it also has a strong academic component, with high overlap between the organizing and program committee members and the various committees and boards of the other publishing venues. The corpus contained 2,045 papers in total (1,472 conference papers and 573 journal papers). For ease of readability this corpus is simply referred to as the SWVs corpus in the rest of the paper.

*A conceptual topic extraction and analysis workflow:* Generally speaking, the typical topic extraction and analysis workflow, as depicted in *Figure* 1, is composed of the following sequential steps:

**Taxonomy creation** involves the creation of a `topic taxonomy` that guides the analysis process. In practice, this step can be achieved manually by domain experts, or automatically with the taxonomy being learned either from the `document corpus` of interest or from a larger `external document corpus`.

**Corpus Annotation** concerns the annotation of the `document corpus` in terms of the taxonomy topics. Different annotation approaches range from manually assigning each paper in a corpus to the most representative topics, annotating the document abstracts with the relevant topics, or annotating the entire text of the paper based on a topic list or hierarchy.

**Analytics** refers to various analytical activities that can be conducted over the `annotated document corpus`. For instance, document classification, trend detection, expert profiling and recommendations.

*Data-driven topic extraction and analysis tools:* Although all three tools conducted their analysis over the same corpus, each of them employ different approaches to topic extraction. An overview of the approaches adopted by PoolParty, Rexplore, and Saffron with respect to the main steps depicted in *Figure* 1 is summarized in *Table* 1 and described below:

**PoolParty** is a semantic technology suite that supports the creation and maintenance of thesauri by domain experts [38]. Although PoolParty is a commercial product, a free version, which was made available in the context of the PROPEL project[7] [16], was used to perform the analysis described in this paper. In the case of the analysis described in this paper the taxonomy was created from conference and journal metadata (i.e., call for papers, sessions, tracks, special issues etc.), which have been manually curated by experts from the Semantic Web community (i.e., conference organizing committee and editorial board members). In order to reduce the potential for bias during the taxonomy construction, the classification, which was performed in the context of the PROPEL project, was collectively performed by five Semantic Web experts. The topic frequency analysis was subsequently conducted by PoolParty over the full text of the research articles from the SWVs corpus, without any parameterization.

**Rexplore** is an interactive environment for exploring scholarly data that leverages data mining, semantic technologies and visual analytics techniques [31]. In the context of this paper, we used Rexplore for tagging research papers with relevant research topics from the Computer Science Ontology (CSO) [35], an existing ontology of research areas that was automatically generated from a large computer science corpus. The ap-

[4] PoolParty, https://www.PoolParty.biz/system-architecture/
[5] Rexplore, http://skm.kmi.open.ac.uk/rexplore/
[6] Saffron, http://saffron.insight-centre.org/
[7] PROPEL, https://www.linked-data.at/

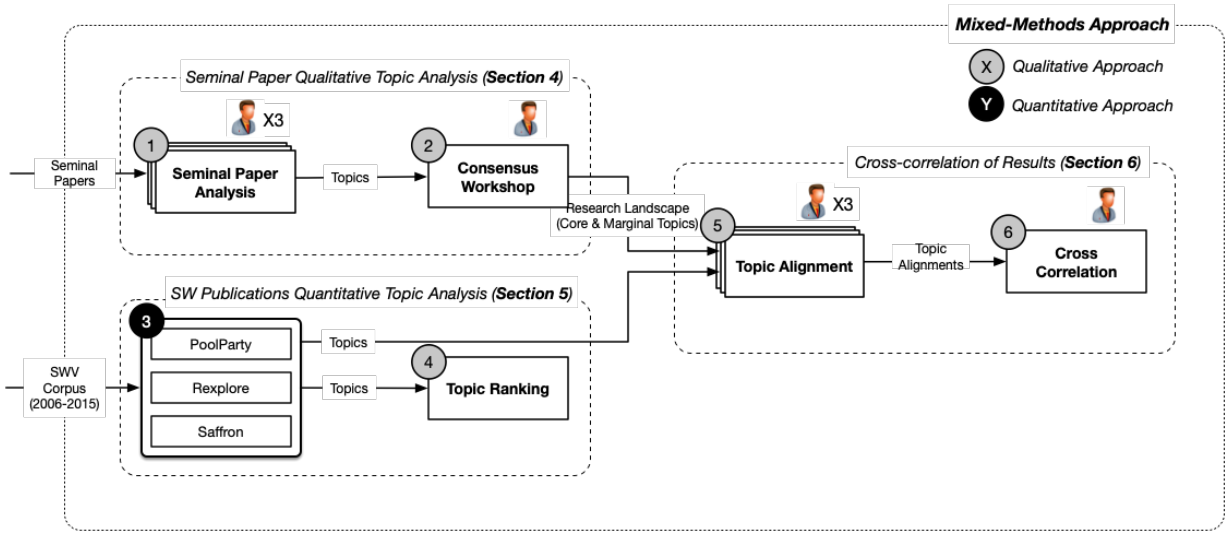



Fig. 2. Overview of the mixed methods-based methodology.

Table 1
Comparison of the methods and data sets used by various topic and trend analysis tools.

| Tool | Taxonomy Creation | Topic Taxonomy | Document Corpus | Corpus Annotation | Topic Analysis | Other Analytics |
|---|---|---|---|---|---|---|
| PoolParty | Manual | Fairly broad/deep | SWVs 2006-2015 | Automatic (full-text) | Topic frequency in text | Taxonomy extension |
| Rexplore | Automatic from broader external corpus | 17K topics in CS, 96 topic in SW, 9 levels deep | SWVs 2006-2015 Scopus 2006-2015 | Automatic (abstracts, titles, keywords) | Number of papers and citations associated with a topic | Taxonomy learning, expert profiling |
| Saffron | Automatic from the document corpus | Fairly broad/deep | SWVs 2006-2015 | Automatic (full-text) | Topic frequency and semantic relatedness | Taxonomy learning, expert finding, document classification and search |

proach for tagging the publications, which took into consideration their title, keywords, abstract and citations, is a slight variation of the method adopted by Springer Nature for characterizing semi-automatically their Computer Science proceedings [32]. The analysis involved the generation of statistical information based both on the number of papers and the citations associated with a topic. No special parameterization was used by the Rexplore in the context of this study. Rexplore was applied both on the SWVs corpus and on a more comprehensive dataset including 32,431 publications associated to the Semantic Web. The aim of this additional analysis was to assess if the set of papers published in the main venues present a different topic distribution than the set of all papers about the Semantic Web.

**Saffron** is a topic and taxonomy extraction tool whose main applications include expert finding, document classification and search [27]. In the context of this paper, we used Saffron's Natural Language Processing (NLP) techniques to extract domain-specific terms based solely on the full text of articles in the SWVs corpus, and a novel taxonomy generation algorithm that uses a global generality measure to direct the edges from generic concepts to more specific ones, in order to construct a topical hierarchy. Additional details on the algorithms used for term (topic) extraction and for extraction of a topic taxonomy can be found in [4]. The topic frequency and relatedness analysis was conducted automatically by Saffron over the SWVs corpus without the need of any additional corpora. Based on previous studies conducted by the Saffron team, in terms of parameterization the taxonomy was limited to 500 topics and topics that appear in at least 3 papers.

In `Step 4` we performed a syntactic analysis of the top forty topics extracted by each of the data-driven

tools. Both singular and plural representations of a topic were treated as the same topic. Additionally, topics with a high syntactic correlation were treated as the same, for instance *knowledge base* and *knowledge based systems*. A detailed description of the respective analysis performed by PoolParty, Rexplore and Saffron and the cross correlation of topics is presented in *Section* 5.

### 3.3. Cross correlation of results

The final stage of our analysis involved the alignment of the topics identified by Rexplore, PoolParty and Saffron with the *Research Landscape* topics emerging from the analysis of the seminal papers. In `Step 5` the output of each of the three data-driven approaches was mapped by one of the authors of this paper to the topics of the *Research Landscape*. The principles used to guide the mapping process, which involved a combination of syntactic and semantic matching, can be summarized as follows:

**Exact syntactic match:** is the most straightforward case as topics that have exactly the same label (e.g., *Linked Data*) are already aligned.

**Partial syntactic match:** refers to cases where two topics have similar but not exactly matching labels, however clearly refer to the same body of research. For instance, *Description Logics* is a subtopic of *Logic and Reasoning*.

**Semantic match:** denotes topics that have syntactically completely disjoint labels but they are semantically related. Links between syntactically different labels are often recorded in our extended *Research Landscape* document, where several keywords were assigned to a larger overlapping topic. For example, we assigned keywords such as SPARQL to the *Query Languages* topic.

**No match:** is used to represent topics identified by the data-driven approaches that are completely new and cannot be related to any of the topics of the *Research Landscape*.

In order to reduce any bias, in `Step 6` individual topic alignments were cross-checked by the two additional authors and further discussed during an analysis and cross-correlation workshop. The results of this workshop are depicted in *Tables* 3- 6 and further discussed in *Section* 6.

## 4. Seminal Papers Topic and Trend Analysis

In the Semantic Web area, a handful of well-known papers identify research topics and discuss trends within the community [1,2,15]. Some of these papers predict future topics [1,2], while others reflect on research topics in the past years or in the present [2,15].

### 4.1. The seminal papers

At the turn of the millennium (2001), Berners-Lee et al. [1] coined the term "Semantic Web" and set a research agenda for a multidisciplinary research field around a handful of topics.

Six years later, Feigenbaum et al. [15] analyzed the uptake of Semantic Web technologies in various domains as of 2007. In doing so, they provided a picture of the technologies available at that time as well as the main challenges that these technologies could solve. The authors took a reflective rather than predictive stance in their work. On the 15-year anniversary of the Semantic Web community, Bernstein et al. [2] provide their vision of research beyond 2016 by grounding their predictions in an overview of past and present research. Therefore, their paper is both reflective of past/present work and predictive in terms of future research.

Each of the vision papers mentioned above are primarily based on the expert knowledge of the authors and reflect their views, without aiming to be complete. Our objective is to use the topics identified in these seminal papers as a baseline for a comparison with the output of the three data-centric topic identification methods discussed in this paper. Note that, unlike in information retrieval research, the proposed *Research Landscape* (cf. *Table* 2) is by no means an absolute gold-standard that should be achieved, but rather acts as an intuitive comparison basis for understanding the strengths and weaknesses of expert-driven versus data-driven topic identification methods.

### 4.2. Core topics from the seminal papers

After manually annotating research topics discussed in each of the seminal papers, we aligned the identified topics across papers, and observed eleven *core research topics* that are mentioned by two or three of the seminal papers (cf. *Table* 2). All three papers agree on the following eight core research topics:

*Knowledge representation languages and standards*, such as XML, RDF and a so-called Seman-

Table 2
Research Landscape: Core and Marginal topics discussed in the seminal papers. Topics in () were only intuitively mentioned.

| | Berners-Lee et al. [1] Future | Feigenbaum et al. [15] Past (2000-2007) | Bernstein et al. [2] Past (2000-2016) | Bernstein et al. [2] Future from 2016 |
|---|---|---|---|---|
| Core topics | knowledge representation languages and standards | knowledge representation languages and standards | knowledge representation languages and standards | representing lightweight semantics |
| | ontologies and modeling, taxonomies, vocabularies | ontologies and modeling, taxonomies, vocabularies | ontologies and modeling, (PR) knowledge graphs | - |
| | logic and reasoning | logic and reasoning | logic and reasoning | - |
| | search and question answering | (ranking) | (PR) question answering systems | - |
| | (data integration) | (ontology matching) | (PR) needs-based, lightweight data integration | integration of heterogeneous data |
| | proof & trust | privacy, trust, access control | personal information, privacy | trust & data provenance (representation, assessment) |
| | databases | semantic web databases | database management systems | - |
| | decentralization | (decentralization) | vastly distributed heterogeneous data | (decentralization) |
| | (machine learning, prediction, analysis, automatic report) | knowledge extraction and discovery | latent semantics, knowledge acquisition, ontology learning | - |
| | - | query language (SPARQL) | developing efficient query mechanisms | - |
| | - | (linked data, DBpedia) | (PR) linked data (open government data), (social data) | - |
| Marginal topics | intelligent software agents | - | multilingual intelligent agents | - |
| | (Internet of Things) | - | - | high volume and velocity of data, e.g., streaming & sensor data |
| | - | (scalability, efficiency, robust semantic approaches) | - | scale changes drastically |
| | (semantic web services) | - | - | - |
| | - | visualization | - | - |
| | - | change management and propagation | - | - |
| | - | (social semantic web, FOAF) | - | |
| | - | - | - | data quality, e.g., representation, assessment |

tic Web language, were considered crucial to enabling the vision of intelligent software agents by Berners-Lee et al. [1]. Work on the development of web-based knowledge representation languages (now also including OWL) continued over the next 7 years [15]. By 2016 this was seen as a core line of research extending also to the standardisation of representation languages for services [2]. As for the future, Bernstein et al. [2] predict that knowledge representation research will focus on representing lightweight semantics, dealing with diverse knowledge representation formats and developing knowledge languages and architectures for an increasingly mobile and app-based Web.

*Knowledge structures and modeling.* Berners-Lee et al. [1] consider knowledge structures such as ontologies, taxonomies and vocabularies as essential components of the Semantic Web. Follow up papers confirm active research on the creation of ontologies [2,15]. While, Bernstein et al. [2] introduce knowledge graphs as novel knowledge representation structures.

*Logic and Reasoning.* Berners-Lee et al. [1] assumed that inference rules and expressive rule languages would enable logic-based automated reasoning on the Semantic Web. Their prediction was abundantly confirmed in follow-up papers: Feigenbaum et al. [15] reporting work on the development of inference engines for reasoning by 2007; and Bernstein et al. [2] confirming work on developing tractable and efficient reasoning mechanisms.

*Search, retrieval, ranking, and question answering.* Besides intelligent agents, Berners-Lee et al. [1] predicted that search and question answering programs would also benefit from the Semantic Web. In 2007, Feigenbaum et al. [15] indirectly refer to this topic in the context of ranking, however this research topic becomes increasingly important accord-

ing to Bernstein et al. [2] who describe work on question answering systems based on semantic markup and linked data from the Web (e.g., IBM Watson).

*Matching and Data Integration.* Ontology matching and data integration were already intuitively mentioned, but not concretely named, by Berners-Lee et al. [1]. Data integration played an important role in many commercial applications developed up until 2007 and opened up the need for change management and change propagation across integrated data sets [15]. By 2016, a new trend towards needs-based, lightweight data integration is observed [2]. For the future, Bernstein et al. [2] discuss the need to integrate heterogeneous data as part of the broader topic of data management.

*Privacy, Trust, Security, and Provenance.* Berners-Lee et al. [1] envision proofs and digital signatures as key aspects of the Semantic Web in order to enable more trustworthy data exchange and the topic of privacy was also mentioned in 2007 [15]. According to Bernstein et al. [2] future work should focus on the representation and assessment of provenance information, as part of the broader topic of data management.

*Semantic Web Databases.* Similarly to Berners-Lee et al. [1], Feigenbaum et al. [15] discuss research topics around the development of Semantic Web tools as instrumental for commercial uptake, especially ontology editors (e.g., Protégé) and Semantic Web databases (e.g., triple stores). According to Bernstein et al. [2] many of these tools evolved into commercial tools by 2016.

*Distribution, decentralization, and federation.* Berners-Lee et al. [1] envisioned that the Semantic Web would be as decentralized as possible, bringing new interesting possibilities at the cost of losing consistency. Feigenbaum et al. [15] exemplified one of these novel scenarios by mentioning FOAF as an example of a decentralized social-networking system. Bernstein et al. [2] commented on this topic briefly, confirming that modern semantic approaches already integrate distributed sources in a lightweight fashion, even if the ontologies are contradictory.

Besides the aforementioned core topics, three important topics were not predicted by Berners-Lee et al. [1], but were mentioned by the other two papers:

*Knowledge extraction, discovery and acquisition.* In 2007, Feigenbaum et al. [15] hint at this topic with terms such as machine learning, prediction and analysis. Automatic knowledge acquisition was boosted by more powerful statistical and machine learning approaches as well as improved computational resources [2]. For the future, Bernstein et al. [2] identify a need for new techniques to extract latent, evidence-based models (ontology learning), to approximate correctness and to reason over automatically extracted ontologies/knowledge structures. An increasing importance is given to using crowdsourcing for capturing collective wisdom and complementing traditional knowledge extraction techniques.

*Query Languages and Mechanisms.* By 2007, research also focused on the development of query languages, most notably SPARQL [15] and developing efficient query mechanisms [2].

*Linked Data.* By mentioning DBpedia, Feigenbaum et al. [15] intuitively pointed to the future research topic of Linked Data. This topic became well established by 2016 and a new wave of structured data available on the web (e.g., open government data, social data) further extended research on the Linked (open) Data topic [2].

### *4.3. Marginal topics from the seminal papers*

Our analysis also identified several marginal topics, mentioned by two of the seminal papers (*Table* 2), as follows:

*Intelligent software agents.* The underpinning theme of Berners-Lee et al. [1]'s vision paper was *intelligent software agents* that would provide advanced functionality to users by being able to access the meaning of Semantic Web data. Interestingly, this topic has not been mentioned until recently, when Bernstein et al. [2] discuss work on training conversational intelligent agents based on multilingual textual data on the web.

*Internet Of Things.* The application of Semantic Web to physical objects within the context of the future Internet Of Things (IoT) was intuitively mentioned by Berners-Lee et al. [1]. This topic was not mentioned by any of the follow-up papers, even thought it is considered to play an important role in the future. Indeed, Bernstein et al. [2] predict that dealing with high volume and velocity data will be necessary due to the increased number of streaming data sources from sensors and the IoT. They envision techniques for the selection of streaming data (data triage), for decision-making on streaming sensor data as well as the integration of streaming sensor data with high quality semantic data.

*Scalability, efficiency and robustness.* Feigenbaum et al. [15] position *scalability, efficiency and robust semantic approaches* as key factors needed to ad-

dress Semantic Web challenges, in particular integration, knowledge management and decision support. In turn, Bernstein et al. [2] recognize that new research is needed given that the *scale changes drastically*.

*Semantic Web Services.* Berners-Lee et al. [1] also envisioned the applicability of Semantic Web technologies for advertising and discovering web-services.

*Human-Computer Interaction.* Feigenbaum et al. [15] mention visualization as features of user-centric applications.

*Change management and propagation.* Feigenbaum et al. [15] mention or hint that *change management and change propagation* across integrated data sets is needed to accompany data integration research.

*Social semantic web.* Although predicting future trends was not their explicit goal, by mentioning FOAF Feigenbaum et al. [15] intuitively pointed to the future research topic on the *Social Semantic Web*.

*Data quality* Under the heading of data management, Bernstein et al. [2] group work on data integration, data provenance and new technologies that should allow representing and assessing *data quality*, such as task-focused quality evaluation (e.g., is a resource of sufficient quality for a task?).

### *4.4. Trends*

Although the seminal papers focus primarily on research topic identification, they also offer some hints on the way these topics evolve over time (i.e., trends).

In 2001, Berners-Lee et al. [1], used a fictitious scenario to describe a vision of a web of data that can be exploited by *intelligent software agents* that carry out data centric tasks on behalf of humans. Additionally the paper identifies the infrastructure necessary to realize this vision focusing on four broad areas of research, namely: *expressing meaning*, *knowledge representation*, *ontologies* and *intelligent software agents*.

In 2007, Feigenbaum et al. [15] reflected on the ideas presented in [1] and highlighted that although the original autonomous agent vision was far from being realized, the technologies themselves were proving to be highly effective in terms of tackling *data integration* challenges in enterprises especially in the life sciences and health care domains. Furthermore, the authors highlighted that consumers were starting to adopt *FOAF* profiles and to embrace *decentralized* social-networking. However, they also point to new *privacy* concerns when linking disparate data sources.

In 2016, discussing present research topics, Bernstein et al. [2] noted a large spectrum between two opposite research lines on expressivity and *reasoning* on the Web on the one hand and ecosystems of *Linked Data* on the other. Particularly notable is the adoption of Semantic Web technologies in several large, more applied systems centered around *knowledge graphs*, which use Semantic Web representations yet ensure the functionality of applied systems which resulted in less formal and precise representations than expected at the earlier stages of Semantic Web research. Based on these considerations, the authors predict moving from logic-based to evidence-based approaches in an effort to build truly *intelligent applications* using vast, *heterogeneous*, *multi-lingual* data.

## 5. Semantic Web Publications Topic and Trend Analysis

In this section we describe the results of topic and trend analysis by employing data-driven tools. The bottom up analysis was performed with three different tools (i.e., PoolParty, Rexplore, and Saffron) that enable users to gain insights into the various research topics that appear in research papers published at popular Semantic Web publishing venues.

### *5.1. PoolParty quantitative analysis*

The analysis conducted by PoolParty was based on a coarse grained taxonomy of 3,420 unique dictionary topics (that were crowd-sourced from experts in the community in the form of conference and journal metadata), which was generated by assigning each topic to one or more foundational technologies worked on by the community.

The chart presented in *Figure* 3 provides details on the % coverage for each of the eighteen foundations, across the five venues for the 10-year timeframe under examination. As expected, *Knowledge Representation & Data Creation/Publishing/Sharing* is the top foundation, with almost 23% of the total occurrences in all documents. This foundation includes several topics that are fundamental to the Semantic Web community (i.e., the ability to represent semantic data and to publish and share such data). Next in order of importance, the management of such knowledge (*Data Management*) and the construction of feasible systems (*System Engineering*), constitute almost 16% and 11% of the occurrences, respectively. Important functional areas such as *Searching, Browsing & Exploration*, *Data Integration and Ontology/Thesaurus/Taxonomy Man-*

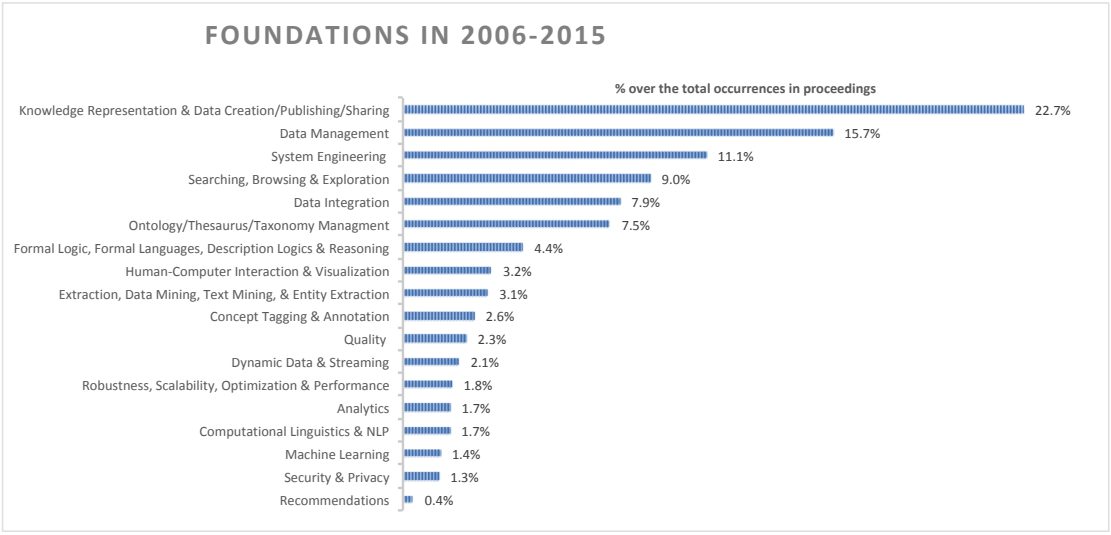



Fig. 3. PoolParty: % coverage per foundational technology across the 5 venues for the 10-year timeframe

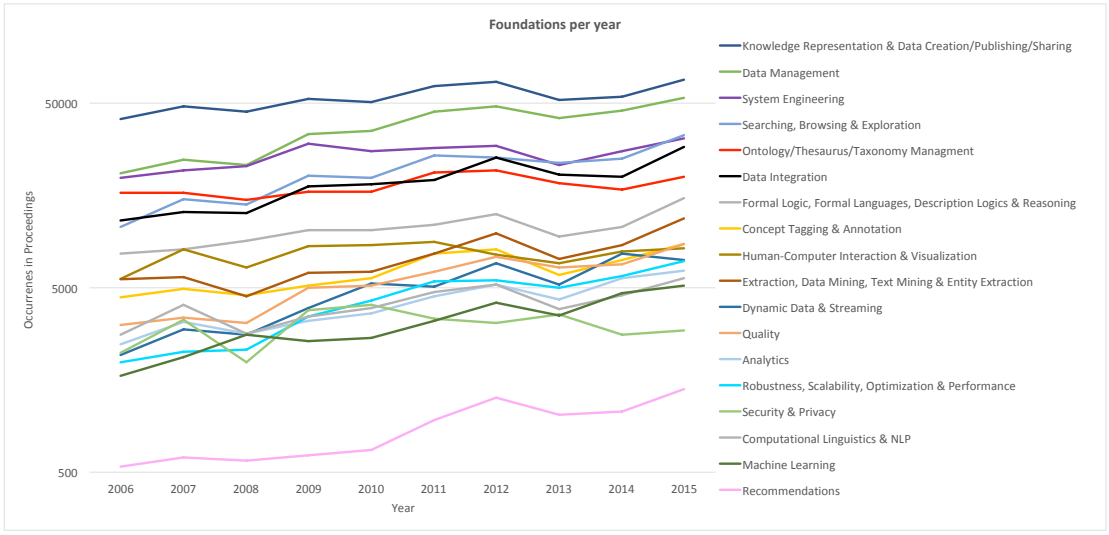
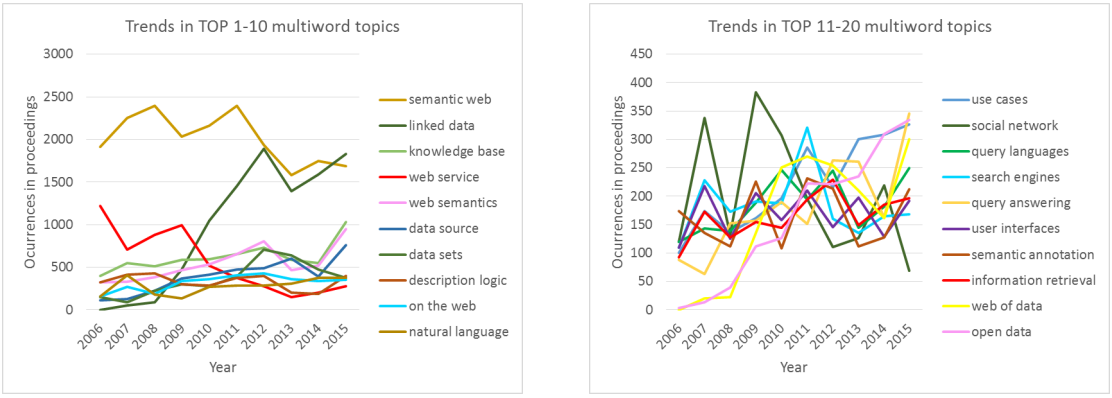



Fig. 4. PoolParty: Growth/Decline of foundational technologies across the 5 venues for the 10 year timeframe

*agement* also figure strongly in comparison to the other foundations (all of them with more than 7.5% occurrences). In contrast, very specific topics, such as *Formal Logic, Formal Languages, Description Logics & Reasoning*, and *Concept Tagging & Annotation* represent a modest 4.4% and 2.6% respectively, and cross-topics, such as *Human Computer Interaction & Visualization*, *Machine Learning*, *Computational Linguistics & NLP*, *Security & Privacy*, *Recommendations*, and *Analytics* are only marginally represented. Topics that relate to *Quality*, *Dynamic Data & Streaming*, and *Robustness, Scalability, Optimization & Performance* are also under-represented (at around 2%).

In order to gain some insights into the research trends over the last decade, *Figure* 4 depicts the growth/decline of each of the foundations over the 10-year timeframe. Although the general trend for all topics shows year on year increases, we note that *Robustness, Scalability, Optimization & Performance*, *Dynamic Data & Streaming*, *Searching, Browsing & Exploration*, and *Machine Learning* have increased by more than 200% since 2005. In contrast, *Security & Privacy*, and *Ontology/Thesaurus/Taxonomy Manage-*




(a) top 10 multi-word topics

(b) top 11-20 multi-word topics

Fig. 5. PoolParty: Growth/Decline of the (a) top 10 and (b) top 11-20 multi-word topics across the 5 venues for the 10 year timeframe.

*ment* have had marginal growth of only 30% for the same period.

*Figure* 5 focuses on the growth/decline of the top 20 multi-word topics. Interestingly, results show a sharp increase of *Linked Data* at the expense of *Semantic Web*. Note also that *Natural Language* is in the top-10 multi-word topics, even though this is a cross topic which may be more represented in a different community. Finally, the decrease in the occurrence of *Web Services* can also be seen here.

### 5.2. *Rexplore quantitative analysis*

Rexplore characterizes topics according to the Computer Science Ontology (CSO) [8] [35], which is a large-scale automatically generated ontology of research areas. Since it is interesting to compare the trends exhibited by high-tier domain conferences with the ones appearing in the full literature, we analysed both the SWVs corpus (described in *Section* 3) and a more comprehensive dataset (here labelled Full Semantic Web, FSW) containing 32,431 publications associated with the topic Semantic Web or with its 96 associated subtopics in CSO (e.g., Linked Data, RDF, Semantic Web Services) from a dump including all Scopus Computer Science papers in the interval 2006-2015.

The analysis presented here follows the Expert-Driven Automatic Methodology (EDAM) [33] for performing systematic reviews of scholarly articles. EDAM is a methodology that reduces the amount of manual tedious tasks involved in systematic reviews by 1) applying data driven methods for automatically generating an ontology of research areas, 2) revising it with domain experts, and 3) using it to annotate papers and produce relevant analytics. The papers were associated to a topic if they contained in the title, abstract, or keywords: 1) the label of the topic (e.g., "Semantic Web"), 2) a *relevantEquivalent* of the topic (e.g., "Semantic Web Tecnologies"), 3) a *skos:broaderGeneric* of the topic (e.g., "ontology matching"), or 4) a *relevantEquivalent* of any *skos:broaderGeneric* of the topic (e.g., "ontology mapping")[9]. We chose this straightforward approach instead of other more complex methods based on string similarity [36] or word embeddings [37], since it is simple to reproduce and yields the best precision, as discussed in Salatino et al. [37].

*Figure* 6 shows the main research fields addressed by the Semantic Web papers in both SWVs and FSW, ranked by the percentage of their publications in the field of Semantic Web. We excluded from this view any super and sub areas of Semantic Web that will be discussed later in detail. Unsurprisingly, the topic *Ontology* appears in about 61.2% of the papers (55.3% for FSW), followed by *Artificial Intelligence* (35.1%, 27.2%), *Information Retrieval* (32.7%, 25.2%), *Query Languages* (26.5%, 17.1%) and *Knowledge Base System* (17.5%, 12.7%). Interestingly, these five core research areas appear more often in the main venues (+7.1% in average), but they are also very important areas for the FSW dataset. Other research areas appear more prominently in one of the datasets. The *Query Language* area is much more frequent in the

[8] https://cso.kmi.open.ac.uk

[9] A detailed description of the relevant semantic relationships is available in Salatino et al. [35].

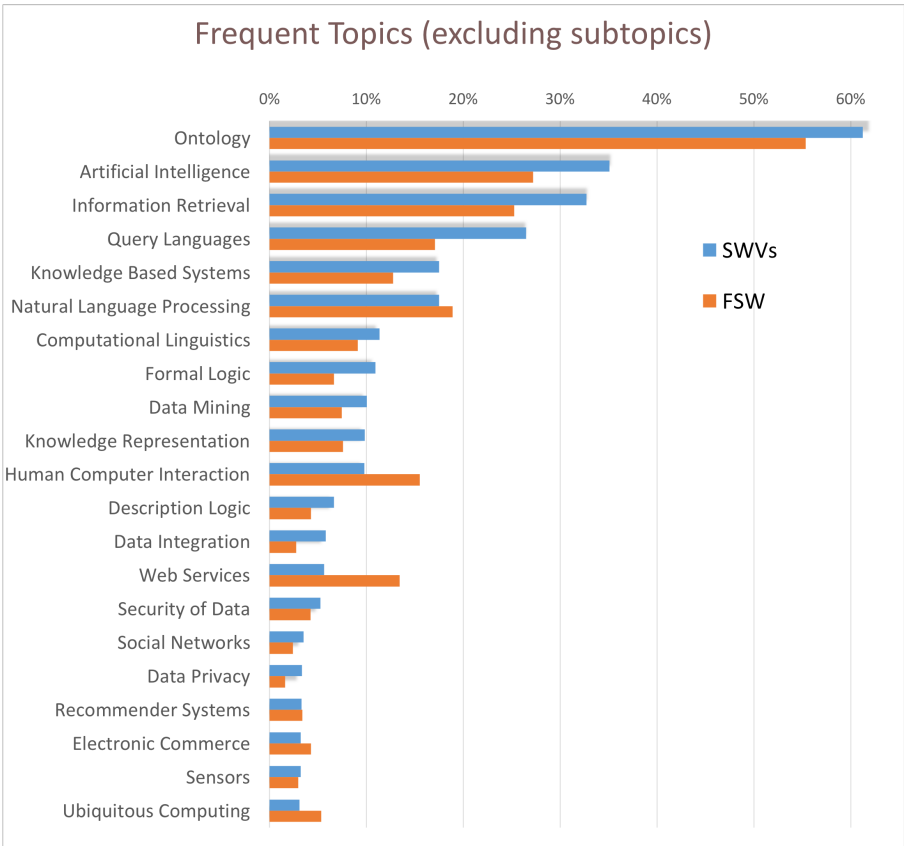



Fig. 6. Rexplore: Frequent topics (excluding subtopics) in SWVs (blue) and FSW (red).

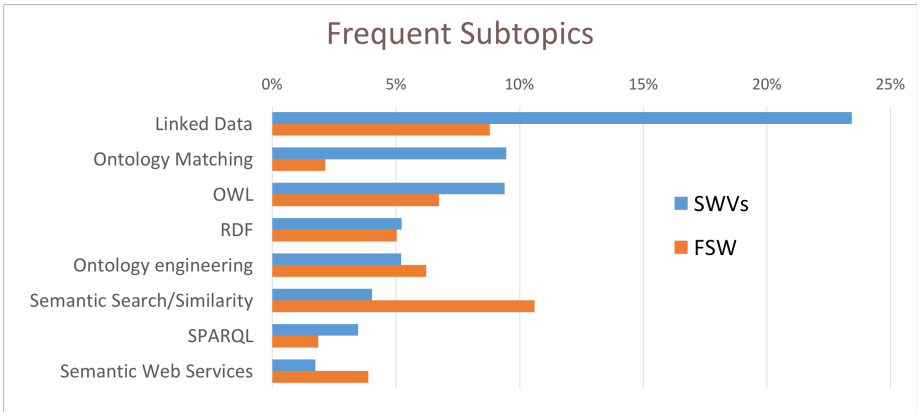



Fig. 7. Rexplore: Frequent Semantic Web subtopics in SWVs (blue) and FSW (red).

SWVs, probably due to the fact that the main venues traditionally are focused on Semantic Web query languages, such as SPARQL. *Formal Logic* has a similar behavior (10.9%, 6.6%), suggesting a stronger focus of the main venues on this topic. Conversely, other research fields appear more often in the FSW dataset. This is the case of *Natural Language Processing* (17.5%, 18.9%), *Human Computer Interaction* (9.8%, 15.5%), *Web Services* (5.6%, 13.4%), *Electronic Commerce* (3.2%, 4.3%) and *Ubiquitous Computing* (3.1%, 5.3%). This seems to suggest that there is a good amount of research in the intersection of

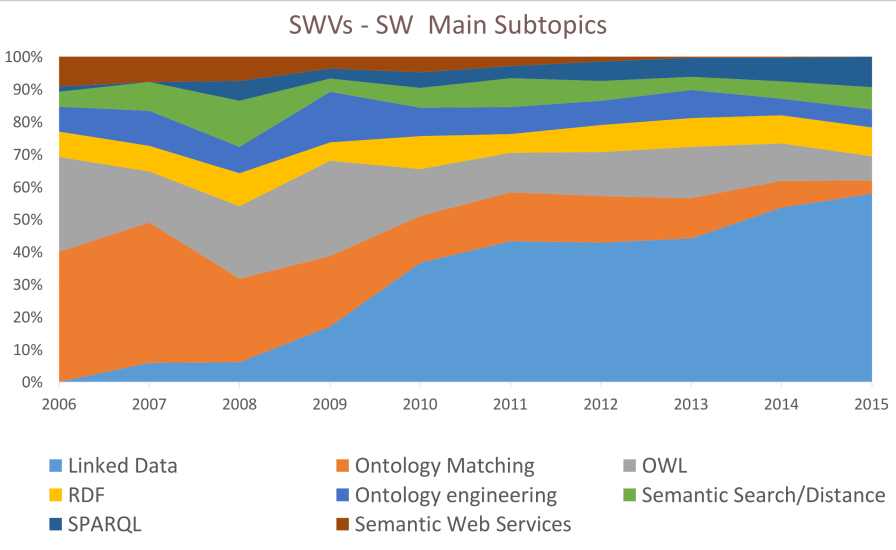



Fig. 8. Rexplore: Number of publications associated with eight Semantic Web subtopics in SWVs.

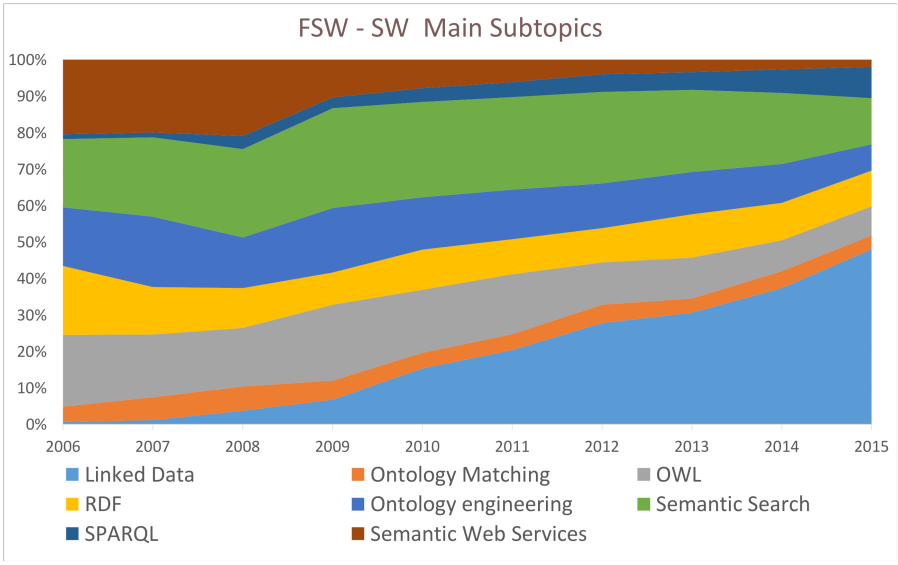



Fig. 9. Rexplore: Number of publications associated with eight Semantic Web subtopics in FSW.

these topics and Semantic Web that is not fully represented in the main venues.

The Semantic Web field subsumes several heterogeneous research areas dealing with different aspects of its vision. *Figure* 7 shows the popularity of the main Semantic Web direct subtopics in the two datasets. We include in this view also the area of *Ontology Engineering*, which is not formally a sub-topic of Semantic Web, since a very large portion of its outcomes are published in the main Semantic Web venues. It is again interesting to consider the difference between the datasets. The topics *Linked Data* (23.4, 8.8%), *Ontology Matching* (9.5%, 2.1%), *OWL* (9.4%,6.7%), and *SPARQL* (3.5%,1.9%) are more frequent in the main venues. Conversely publications addressing *Semantic Search* (4.0%, 10.6%) and *Semantic Web Services* (1.7%, 3.9%) are more popular outside these venues.

*Figure* 8 and *Figure* 9 show the popularity of the main sub-topics over the years. The two main dynamics, evident in both datasets, are the fading of Seman-

tic Web Services and the rapid growth of Linked Data and to a lesser extent of SPARQL. Indeed, *Semantic Web Services* is one of the main areas in 2004, and an integral part of the initial Semantic Web vision [1]. However, the number of papers about these topics consistently decreases and from 2013 there are almost no publications about them in the SWVs corpus and very few in FSW. The second trend is the steady growth of Linked Data from 2007. In 2015 about half of Semantic Web papers in the main venues refer to this topics. Interestingly, both trends are first anticipated by the main venues, and only later evident also in the FSW dataset. It thus seems that the tendencies of the main venues influence in time all the Semantic Web research.

### 5.3. Saffron quantitative analysis

Saffron employs a domain-independent approach to topic extraction, which is one of its biggest advantages compared to most systems in the area, in that it does not require external domain-specific classifications. Such information is often not readily available especially in niche domains, and creating a classification is very costly in terms of time, human expertise needs, and maintenance. Saffron bypasses this barrier by automatically building a domain model from the input corpus itself, and by capturing the expertise knowledge of the corpus by isolating its most generic concepts. The constructed hierarchical taxonomy can be visualized as a graph. We use Cytoscapes, an open source software tool for complex networks graph visualization[10]. It allows us to perform a network analysis on the output provided by Saffron, and a customization of the layout. In our case, the size and the color of the nodes are proportional to the number of neighbors each topic is connected to.

*Figure* 10 shows the general picture of the graph displaying the interconnected topics from the results of the analysis. The size and the colour of the nodes in the graph are related to the number of edges that are connected to them, ie. the bigger nodes with red shades are the most connected topics while the smaller and blue nodes are the leaves of the tree. The first and predominant node (i.e., the root of the taxonomy) is the *Semantic Web* topic itself. Around it, several main clusters with major keyphrases emerge, including: *RDF Data* and *Linked Data*, followed by *Natural Language*, *Data Source* and *Reasoning Task*. A strong focus is also put on *Machine Learning*, *Ontology Engineering*, *Query Execution*, and the mark-up language *OWL-S*. By concentrating on the clusters, we identify the importance of data in terms of its representation (*RDF Data*, *RDF Graph*, *Linked Data*), its accessibility (*Open Data*), and its querying (*Query Execution*, *Query Processing*). Some other main interests in the domain are visible, represented by a cluster made up of *Natural Language* and topics related to the querying of information such as *Semantic and Keyword Search*, *Keyword Query*, *Semantic Similarity* or *Information Retrieval*. *Natural Language* is also connected to another dominating topic that is *Machine Learning*, associated to *Ontology Matching and Mapping*. One of the branches originating from the *Semantic Web* topic brings together concepts related to the structure and representation of the ontology (*Knowledge Base*, *Knowledge Representation*, *Ontology Language*, *OWL Ontology*), while a sub-branch leads to logic and reasoning related topics (*Description Logic*, *Reasoning task*, *Reasoning Algorithm*). The *Ontology Engineering* node is related to topics such as *User Interface*, *Ontology Development* and *Ontology Editing*.

As demonstrated above, the main nodes are at the centre of clusters of topics that are semantically related to them. In the following analysis, we focus thus on the evolution of those major terms, which are the most prominent for a cluster. We selected the top 20 topic terms (i.e., the most connected ones) and observe the distribution of their use in the SW corpus. The two charts in *Figure* 11 show the percentage of documents containing the aforementioned topics (i.e., the number of documents where the term appear at least once), per year. We observe that *Semantic Web* as a topic is on the decline with a decrease of 20% between the beginning and the end of the studied time frame. It is still the most used topic nonetheless, reaching 91% of distribution in the documents in 2006 and lowering down to 70% in 2015. The reasons for this decline could be manifold: the term/field may be so established that it is not named explicitly in the papers anymore or the community is trying to re-brand their research with new terms such as *Linked Data*.

Indeed, the most significant progression is the use of the term *Linked Data*. While it was completely missing in 2006, it experienced a very rapid growth in particular between 2008 and 2010 where its rise was 9-fold, to eventually reach 64% of the distribution in the documents by 2015. Similarly, the *Open Data* topic increased from about 1% in 2006/2007 to 45% in 2015.

[10]Cytoscapes, http://www.cytoscape.org/

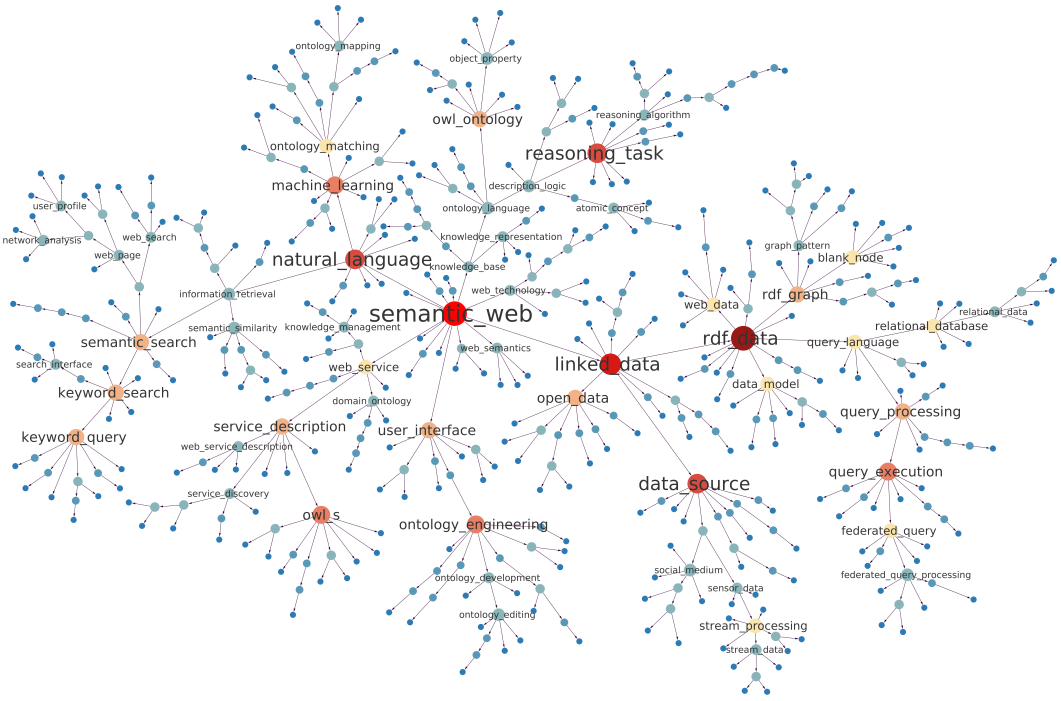
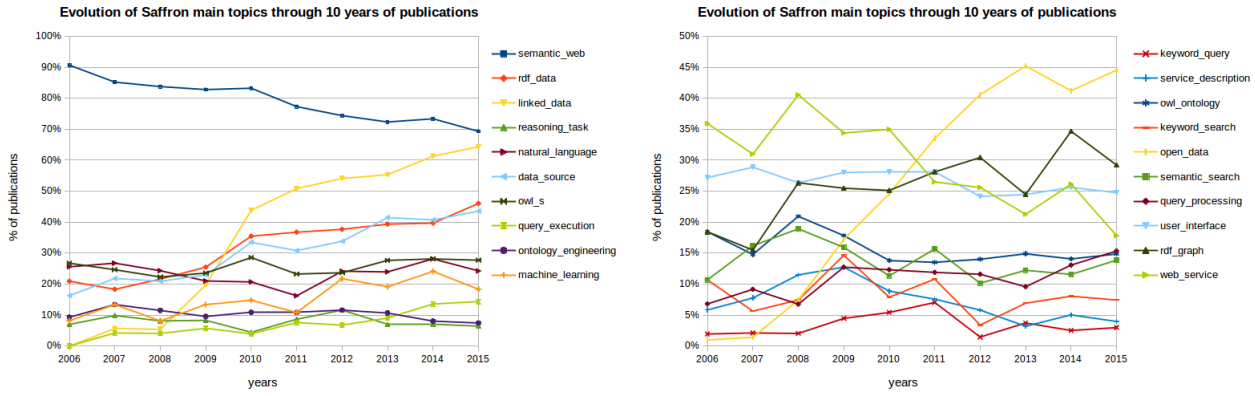



Fig. 10. Saffron: Taxonomy of Semantic Web topics.




(a) top 10 topics

(b) top 11-20 topics

Fig. 11. Saffron: Topic term occurrence evolution over time for (a) top 10 topics and (b) top 11-20 topics.

Other emerging topics include: *Query Execution*, appearing in 2015 in 15% of the publications as well as *RDF Data* and *Data Source* which doubled their presence since 2006. Topics whose popularity increased by at least twice their initial proportion include *RDF Graph* (with two peaks in 2008 and 2014), *Machine Learning* (with a peak in 2012) and *Query Processing* (with a small peak in 2009 then a quite steady line).

Among the topics experiencing strong variations through time, the term *Web Service* is a declining one. After experiencing a peak of use in 2008 with a 40% distribution in the documents, it then dropped to less than 20% in 2015. *Semantic Search* experienced two small peaks in 2008 and 2011, and slight drops in 2010 and 2012 to a more steady curve thereafter. Some topics appear to be consistent over the years, such as

*Ontology Engineering*, while some others are more volatile. The *Natural Language* topic, despite being equally cited in 2006 and in 2015, gradually dropped in the first half of the period examined, to gain in popularity again after 2011. *Keyword Search* shows quite a varied pattern, with drops in 2007, 2010 and 2012, and peaks in 2009 and 2011. As for *Service Description*, it increased slowly up to 13% by 2009, but gradually declined towards its initial value by 2015.

### 5.4. Comparison of top forty topics extracted by each tool

*Table* 8 and *Table* 9 in *Appendix* A, highlight the top 40 multi-word topics that were extracted by at least two data-driven tools and those that were only identified by a single data-driven tool, respectively, based on a simple syntactic matching of the topics. After normalizing the topic names across the sets, we found 86 unique topics. 12 of these were detected by all systems and 23 by at least two systems. We thus computed the Spearman's Rank-Order Correlation on the intersection of the three sets. We found that Rexplore and Poolparty exhibit a moderate correlation ($\rho = 0.61$) and a statistically significant association (p (2-tailed) = 0.035). Conversely, the list produced by Saffron is not correlated with the ones of Rexplore ($\rho = 0.01$, p (2-tailed) = 0.966.) or Poolparty ($\rho = 0.01$, p (2-tailed) = 0.681).

The topics uncovered by all three tools could be categorized as reflecting the core focus of the community (*knowledge base, linked data, semantic search, semantic web, web services, ontology matching, query languages*) and several well established sub-communities (*information retrieval, machine learning, natural language processing, ontology engineering*). The topics uncovered by two ore more tools further elaborate on core research topics within the community (*data integration, data source, linked open data (LOD), ontology language, open data, query processing, social networks, user interfaces, web data, web semantics, web ontology language (OWL)).* While, the topics uncovered by only one tool are a mix between supporting technology (e.g., *rdf data, rdf Graph, search engines, logic programming, SPARQL*), very specific topics (e.g., *human computer interaction, stream processing, data privacy, federated query processing*), commonly used data sources (e.g., *DBpedia, wikipedia*), and frequently used terms (e.g., *on the web, use cases, web of data*).

Although in this paper we do not go into details of the specific algorithms employed by PoolPary, Rexplore and Saffron, it is possible to speculate as to why certain topics appear in the top forty list of the various tools. For instance, considering that the PoolParty taxonomy is created from conference and journal metadata, it is not surprising that topics such as *case studies*, *use cases* and references to *on the web* or *web of data* appear, as these terms could frequently occur in calls for papers. In the case of Rexplore we see evidence of broader topics, such as *artificial intelligence* and *human computer interaction* that are reflective of the broader nature of the Rexplore taxonomy, which was generated from a more general computer science corpus. Finally, considering that Saffron not only learns the topics from the corpus, but also tries to identify distinguishing topics for papers, it is not surprising that we see evidence of specfic topics such as *federated query processing* and *stream processing*.

## 6. Topic Alignment and Findings

In this section, we compare and contrast the topics extracted by the three bottom-up data-driven approaches (Rexplore, Saffron, PoolParty) and the core and marginal topics mentioned in the seminal Semantic Web papers (discussed in *Section* 4), with primary topics identified by the data-driven approaches presented in *Section* 5. Initially we conducted the mapping exercise with the top 20 topics, however after seeing that there were no mappings for several core topics we elected to use the top 40 multi-word topics from PoolParty, Rexplore and Saffron (see *Table* 7 in *Appendix* A).

### 6.1. Core and marginal topic analysis

The analysis presented in this section is based on a comparison between the core and marginal topics mentioned in the seminal Semantic Web papers and the predominant topics uncovered by PoolParty, Rexplore and Saffron. In contrast to the aforementioned data-driven topic analysis, which was based primarily on the syntactic cross-correlation of topics extracted by PoolParty, Rexplore and Saffron, the analysis presented in this section is based on the clustering of similar topics.

Table 3
Core research topics identified in the seminal papers and their coverage by the data-driven approaches.

| | Coverage | | | Matched topics | | |
|---|---|---|---|---|---|---|
| Core topic | PoolParty | Rexplore | Saffron | PoolParty | Rexplore | Saffron |
| knowledge representation languages and standards | ✓ | ✓ | ✓ | knowledge representation, | knowledge based systems, knowledge representation, Resource Description Framework (RDF), Web Ontology Language (OWL) | rdf data, owl s, blank node, object property |
| Knowledge structures and modeling | ✓ | ✓ | ✓ | ontology/thesaurus/taxonomy management, web semantics, ontology engineering, ontology language, data models, ontology matching | ontology, ontology engineering | owl ontology, ontology engineering, rdf graph, data model, ontology language, ontology editing, web semantics, ontology development, ontology matching |
| logic and reasoning | ✓ | ✓ | ✓ | description logic, formal logic/ formal languages/description logics, logic programming | formal logic, description logic, Web Ontology Language (OWL) | reasoning task, description logic |
| search, retrieval, ranking, question answering | ✓ | ✓ | ✓ | search engines, semantic search, web search, natural language, searching/ browsing/ exploration, computer linguistics & NLP systems, information retrieval | information retrieval, semantic search/similarity, computer linguistics | keyword search, semantic search, natural language, information retrieval |
| matching and data integration | ✓ | ✓ | ✓ | ontology matching, ontology alignment, similarity measures, data integration | ontology matching, data integration | ontology matching, semantic similarity |
| privacy, trust, security, provenance | ✓ | ✓ | - | security & privacy | security of data, data privacy | - |
| semantic web databases | ✓ | - | ✓ | data sets, knowledge base, data source, knowledge management, data management | knowledge base systems | data source, relational database, knowledge base |
| distribution, decentralization, federation | - | - | ✓ | - | - | federated query, federated query processing |
| query languages and mechanisms | ✓ | ✓ | ✓ | query languages, query answering, query processing | query languages, SPARQL, SPARQL queries | query execution, keyword query, query processing, query language |
| linked data | ✓ | ✓ | ✓ | linked data, linked open data, semantic web, web of data, data integration, data creation/publishing/sharing | linked data, semantic web, linked open data, data integration | linked data, semantic web |
| knowledge extraction, discovery and acquisition | ✓ | ✓ | ✓ | information retrieval, machine learning, extraction, data mining, text mining, entity, extraction, analytics, machine learning | information retrieval, natural language processing, data mining, machine learning, natural language processing systems | machine learning, information retrieval |

*Core topic analysis:* As shown in *Table* 3 all three data-driven approaches uncovered eight out of eleven of the Research Landscape topics and all topics were uncovered by at least one data-driven approach. Notable omissions include the *distribution, decentralization and federation* topic, which was not uncovered by PoolParty and Rexplore, the *privacy, trust, security, and provenance* topic, which did not figure in the primary topics uncovered by Saffron, and the *semantic web databases* topic which was not ranked highly by Rexplore.

*Marginal topic analysis:* Comparing the output from the data-driven approaches to the marginal topics presented in *Table* 4 we observe reduced coverage, with the *multilingual intelligent agents* and *change management and propagation* topics not featuring in any of the top 40 topic lists produced by PoolParty, Rexplore and Saffron. While, the *scalability, efficiency, robust semantic approaches* topic was only identified by PoolParty and not by Rexplore and Saffron.

*Additional topics:* In order to complete the analysis in *Table* 5 we highlight the topics that were extracted by the data-driven approaches, however were not mentioned in the seminal papers. All three tools identified topics that are very general in nature and as such could not be easily mapped to the primary topics appearing

Table 4
Marginal research topics identified in the seminal papers and their coverage by the data-driven approaches.

| | Coverage | | | Matched topics | | |
|---|---|---|---|---|---|---|
| Marginal topic | PoolParty | Rexplore | Saffron | PoolParty | Rexplore | Saffron |
| multilingual intelligent agents | - | - | - | - | - | - |
| semantic web services | ✓ | ✓ | ✓ | web service, semantic web service | web services, semantic web services | web service, service description |
| visualization, user interfaces and annotation | ✓ | ✓ | ✓ | user interfaces, semantic annotation, human computer interaction & visualization, annotation, concept tagging | human computer interaction, visualization | user interface |
| (scalability, efficiency, robust semantic approaches) | ✓ | - | - | robustness, scalability, optimization and performance | - | - |
| change management and propagation | - | - | - | - | - | - |
| (social semantic web, FOAF) | ✓ | ✓ | ✓ | social network | social networks | social medium |

Table 5
Research topics covered by the data-driven approaches that were not identified by the seminal papers.

| PoolParty | Rexplore | Saffron |
|---|---|---|
| recommendations, use cases, case studies, open data, information systems, web data, semantic technology, structured data | computational linguistics, recommender systems, mobile devices, cloud computing, e-learning system, robotics, electronic commerce systems, decision support systems | open data, web data, web technology |

Table 6
Visionary research topics from the seminal papers and their coverage by the data-driven approaches.

| | Coverage | | | Matched topics | | |
|---|---|---|---|---|---|---|
| Future topic | PoolParty | Rexplore | Saffron | PoolParty | Rexplore | Saffron |
| scale changes drastically | ✓ | - | | robustness, scalability, optimization and performance | - | - |
| intelligent software agents | - | ✓ | - | - | artificial intelligence | - |
| (Internet of Things), high volume and velocity of data, e.g., streaming & sensor data | ✓ | ✓ | ✓ | dynamic data / streaming | Internet of Things | stream processing |
| data quality, e.g, representation, assessment | ✓ | - | - | quality | - | - |

in the seminal papers. For instance, *recommendations, use cases, case studies, open data, information systems, web data, semantic technology*, and *structured data* in the case of PoolParty, *computational linguistics, recommender systems, mobile devices, cloud computing, e-learning system, robotics, electronic commerce systems*, and *decision support systems* in the case of Rexplore, and *open data*, *web data*, *web technology* in the case of Saffron. Several of the topics uncovered by Rexplore stand out from the others as they are not topics per se but rather application or use case oriented keywords that were not extracted from the seminal papers.

### 6.2. Evidence of future topics

Besides using the data-driven approaches to look for evidence of the topics that the community have been actively working on, we also investigated if the data-driven approaches could also find evidence of future trends predicted in the seminal papers, in particular those mentioned by Bernstein et al. [2]. According to our mapping presented in *Table* 6, evidence with respect to each of the four main lines of future research topics was uncovered by at least one of the data-driven approaches. Interestingly, all approaches found topics relating to the *Internet of Things, streaming and sensor data*, indicating a rise in importance of this topic within the Semantic Web community. However, at the same time, the other three topics that relate to *scale*, *in-*

*telligent software agents* and *quality* were only weakly identified by the seminal papers.

### 6.3. Evidence of trends

In the following we summarize the analysis of the trends identified by PoolParty (cf. *Figure* 4- 5), Rexplore (cf. *Figure* 8- 9) and Saffron (cf. *Figure* 11). The foundational topic and trend analysis conducted via PoolParty did not yield any useful results, as generally speaking work on each of the foundational topics appear to be increasing year on year. A cross correlation of the trends highlighted by PoolParty, Rexplore and Saffron provides evidence that topics such as *linked data*, *open data* and *data sources* have an upward trend, whiles topics such as *semantic web*, *web service*, *service description* and *ontology matching* appear to be on a downward trend. When it comes to trend analysis using the data-driven approaches, it is clear that neither foundational topic analysis nor topic specific analysis, provides us with enough evidence to confirm the visions outlined in the seminal papers. For this there is a need for a more focused analysis that maps visions to relevant research topics and uses year on year aggregate counts to depict trends. Although, Fernandez Garcia et al. [16] made some initial attempts at mapping the trends identified by PoolParty to the visions from the seminal paper, unfortunately such a mapping is not very straightforward even for manual mappings and as such is left to future work.

### 6.4. Mixed methods observations

The comparative analysis of the research topics identified with the qualitative and quantitative methods, discussed in the previous sections, reveals several interesting observations on the benefits and drawbacks of these approaches, as discussed next.

**Qualitative vs. Quantitative approaches.** Comparing the quality of topic detection using data-driven methods with that of expert-driven methods (cf. *Table* 3), we observe that data-driven approaches had a high recall when it comes to detecting core topics identified by experts in the seminal papers. Data-driven methods failed however to cover multidisciplinary topics, (i.e., topics that cross boundaries between areas), such as *distribution, decentralization, federation*, or *privacy, trust, security, provenance*, or *semantic web databases*. These weakly covered topics are particularly interesting, as they indicate research areas that, although considered important by experts, have not yet attracted a critical mass of research to be reliably identified with quantitative methods.

Analyzing the coverage of *marginal topics* (cf. *Table* 4), we find an opposite phenomenon of research topics for which there is marginal agreement among experts, but strong data-driven evidence of work on those topics. Indeed, data-driven approaches confirm some of the marginal topics such as *social semantic web* and *human computer interaction*. These are topics on which a sufficient volume of work is performed to allow identification by data-driven approaches, but for which a core community has not yet been formed.

As expected, the coverage of visionary topics ( *Table* 6) was lower. Although these periphery topics are somehow addressed by the Semantic Web community, the data-driven analysis failed to represent them with the required fine-grained details. It is clear from the results of our analysis that further work on trend detection and analysis is needed in order to better detect emerging topics and to understand the research gaps with respect to the vision.

A major benefit of data-driven methods is that they are capable of providing evidence of the popularity of research areas and topics over time and consequently can be used to derive research trends (although these are somewhat sensitive to the available data and can be less accurate when data is missing, for instance towards the end of the analysis period). When it comes to topics that appear in the Research Landscape but are underrepresented according to our data-driven analysis, such information could be used to encourage publications on these topics via calls for papers of future conferences or via workshops or journal special issues.

**Comparison of Quantitative Methods.** For the quantitative analysis of our work, we employed data-driven methods that differed, among others, in the way the topic taxonomy was created. In the case of PoolParty a manually built topic taxonomy was employed which closely reflected the topics on which the community are looking for in call for papers or in conference programs. Rexplore made use of the CSO ontology, a large-scale ontology of computer science extracted from a very large corpus and covering key research areas as well as associated research topics. Finally, Saffron extracted its taxonomy of topics entirely from the corpus under analysis and used clustering to identify topics that belong to a research area (without actually deriving research area names). Obviously, these approaches of procuring the topic taxonomy are

decreasing in terms of cost as per the time of expert involvement.

In terms of overall performance, (cf. *Tables* 3, 4, 6), PoolParty identified 17/21 core, marginal and future topics (10/11 core topics; 4/6 marginal topics; 3/4 future topics). Together with Saffron, PoolParty identified the most core topics, while achieving the highest recall for the other two topic categories too (i.e., marginal and future topics). Closely after PoolParty, Rexplore identified 14 of the 21 topics of the Research Landscape (9/11 core topics; 3/6 marginal topics; 2/4 future topics), identifying in each category just one topic less than PoolParty. Finally, Saffron is overall very close in its coverage to that of the other two tools by identifying 13 out of 21 topics (10/11 core topics; 2/6 marginal topics; 1/4 future topics). While having a very good coverage of the core topics, Saffron's performance was remarkably inferior to the other tools for the other topic categories, where it primarily identified those topics which were already identified by the other tools. From the above, we conclude that the use of apriory built taxonomies of research areas, while more expensive, leads to a better coverage of research topics, especially in the analysis of marginal or emerging research topics. Moreover, we attribute the high success of PoolParty to covering research topics to the fact that it relied on a high-quality, manually built topic taxonomy that was well aligned to the domain as the topics were extracted from conference and journal metadata.

While the most cost-effective, Saffron identified a bag of topics that was less straightforward to align to research areas than the output of the other two approaches that relied on taxonomies of research areas (and associated topics). The alignment and interpretation of Saffron topics required expert knowledge and therefore Saffron should ideally also be used in settings where such expert knowledge is available.

While PoolParty had the best performance in confirming research topics from the qualitative analysis, Rexplore provided the most additional topics (cf. *Table* 5), clearly identifying research topics at the intersection of the Semantic Web and other research communities (e.g., *computational linguistics* and *cloud computing*), thus providing invaluable support in positioning the work of our community in a broader research context.

## 7. Conclusion

The analysis of research topics and trends is an important aspect of scientometrics which is expanding from qualitative expert-driven approaches to also include data-driven methods. The Semantic Web community is no different, with several seminal papers reflecting on and predicting the work of the community and data-driven methods (based on Semantic Web technologies) trying to achieve similar topic and trend detection activities (semi-)automatically.

With this study, we aimed to go beyond the various views on our community's Research Landscape scattered in several papers and obtained with different methods. To that end, we proposed the use of a *mixed methods approach* that can converge, unify but also critically compare conclusions reached with both expert or data-driven approaches. Finally, we conclude this study by revisiting the original research questions:

**Is it possible to identify the predominant Semantic Web research topics using both expert based predictions and topic and trend identification tools?** A key benefit and novelty of our work is that we identified and aligned core research topics mentioned in the seminal papers and then verified these using data-driven methods. After extracting, grouping and aligning the topics from the seminal papers, we concluded on *eleven core Semantic Web topics* (cf. *Table* 3), out of which eight were confirmed by all the data-driven approaches, while the remaining three indicate topics that are important but not sufficiently represented in papers at the key Semantic Web venues. Besides these core topics, we capture *six marginal topics* (cf. *Table* 4) out of which two are very strongly supported by evidence from data-driven methods.

From a trends perspective, it was clearly visible that topics such as *linked data*, *open data* and *data sources* have increased in importance over the years. While, at the same time, topics such as *semantic web*, *web service*, *service description* and *ontology matching* seem to appear less and less. Although we could speculate as to why this is the case (e.g., a push by the community towards using semantic technology to open up and link data may have caused a decline in work in relation to service based machine-to-machine interaction), however a more in depth analysis, involving sources other than over research papers, would be needed in order to conform our suspicions.

Looking into the future, we identify *four future topics* (cf. *Table* 6), from which the topics on *IoT, sensor and streaming data* has ample evidence in the analyzed research corpus. Finally, the Rexplore data-driven method provided insights into the interactions of our fields with other research areas, highlighting

its cross-disciplinary nature. Considering the growing interest in scientomentrics within the Semantic Web community, our findings could be used as a baseline for benchmarking other topic and trend detection methods for the same time period, or extended to cater for more recent work by the community.

**What are the strengths and weaknesses of expert-driven and data-driven topic and trend identification methods?** Qualitative, expert-driven methods benefit from insights by experts who reflect on past or present research topics and trends and predict future directions. As such, they remain valuable assets in the scientometrics tool-box. Data-driven methods challenge expert-analysis by providing a surprisingly high recall, especially for core research topics, and naturally less for marginal and emerging topics. However, a major benefit of data-driven methods is that their findings are backed-up by quantitative data which can be used to perform a range of other analytics such as research trend detection or identifying connections between research topics.

A key element of the data-driven approaches considered here is the use of a topic taxonomy which can be derived with costly, manual effort, semi-automatically or fully-automatically. Not-entirely surprising, well-curated taxonomies lead to the best performance, but these naturally age very quickly and their maintenance is not sustainable. Therefore, semi-automatic or fully-automatic taxonomy construction methods offer a cheaper and more sustainable alternative with only a slight loss of recall.

In this paper, we proposed and demonstrated the use of a mixed methods approach, which combines both qualitative and quantitative methods in an attempt to overcome their respective weaknesses. This mixed methods approach has several strengths. Firstly, it allowed us to synchronize the results of several qualitative studies and propose a unified Research Landscape of the area. Secondly, by comparing and contrasting the Research Landscape with the results of the data-driven methods, we could: (1) confirm those topics that are both seen as important by experts and for which quantitative evidence can be gathered - these are clearly core topics in the community; (2) identify topics that experts consider important but for which data-driven methods do not (unanimously) find sufficient evidence in the corpus - these are topics that the community should encourage; (3) identify topics on which not all experts agree (which is natural given some bias inadvertently brought in by experts) but which are strongly represented in the research data - these topics could benefit from community building efforts. To summarize, mixed methods allows for drawing interesting conclusions in areas where quantitative and qualitative methods agree or disagree. A weak point of the presented method is the use of manual extraction and alignment of topics which could have introduced bias. We tried to minimize this by performing each of these steps with multiple experts and then reaching agreement where their opinions differed.

In this paper we have focused on approaches to analyse and reflect about the past and to some extent the future development of our research community, using expert opinions, on the one hand, and applying our own data-driven methods, on the other. As such, the comparison and benchmarking of topic detection tools was outside the scope of the paper. Nevertheless, the collected document corpus and the results of our analysis provide the foundations for performing further analysis and benchmarking among topic detection tools in future work.

A first interesting direction would be to apply methods for citation network analysis [9,11] in order to characterize each research field with relevant clusters of papers. We could also apply techniques from the field of spatial scientometrics [18] for analyzing the geographical trends.

Additionally, we could adopt (as mentioned in the end of Section 4.2) emerging methods such as crowdsourcing for a similar reflectional exercise. That is, based on the findings and topics presented here, let the community itself on a larger scale than relying on the insights of a few of its established experts, assess the importance and future of topics for the community. Such an analysis should probably counteract biases in terms of ensuring that researchers do not assess/favor the (future) importance of their own field of research, but we would expect this to be an interesting future direction.

Other avenues for further study include: a more focused analysis that maps visions to relevant research topics and generates the corresponding trends; the deepening of the work to better understand the type of coverage offered in each of the identified research topics; and a broadening of the work to consider not only the research topics but also the application areas and domains where these technologies are routinely applied.

Also, it would be interesting to test this method in other communities (e.g., Software Engineering) and to

further improve the topic alignment methods to further reduce bias.

## Acknowledgments

This publication has emanated from research supported in part by the PROPEL research project funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) under the program "ICT of the Future", and a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 co-funded by the European Regional Development Fund. We would like to thank our colleagues from the Semantic Web Company for their support with the PoolParty analysis.

## References

[1] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific American*, 284(5):28–37, 2001.

[2] Abraham Bernstein, James A. Hendler, and Natalya Fridman Noy. A new look at the semantic web. *Commun. ACM*, 59 (9):35–37, 2016. . URL https://doi.org/10.1145/2890489.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3 (Jan):993–1022, 2003.

[4] Georgeta Bordea. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. Thesis, National University of Ireland, Galway, Ireland, 2013. uri: http://hdl.handle.net/10379/4484.

[5] Georgetas Bordea and Paul Buitelaar. Expertise mining. In *Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland*, 2010.

[6] Michel Callon, Jean-Pierre Courtial, William A Turner, and Serge Bauin. From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22(2):191–235, 1983. .

[7] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*, page 4. ACM, 2010.

[8] David Chavalarias and Jean-Philippe Cointet. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one*, 8(2), 2013. .

[9] Chaomei Chen. Citespace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Assoc. Inf. Sci. Technol.*, 57(3):359–377, 2006. . URL https://doi.org/10.1002/asi.20317.

[10] Manuel J. Cobo, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Assoc. Inf. Sci. Technol.*, 62(7):1382–1402, 2011. . URL https://doi.org/10.1002/asi.21525.

[11] Manuel J. Cobo, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. Scimat: A new science mapping analysis software tool. *J. Assoc. Inf. Sci. Technol.*, 63(8):1609–1630, 2012. . URL https://doi.org/10.1002/asi.22688.

[12] Xiang-Ying Dai, Qingcai Chen, Xiaolong Wang, and Jun Xu. Online topic detection and tracking of financial news based on hierarchical clustering. In *International Conference on Machine Learning and Cybernetics, ICMLC 2010, Qingdao, China, July 11-14, 2010, Proceedings*, pages 3341–3346. IEEE, 2010. . URL https://doi.org/10.1109/ICMLC.2010.5580677.

[13] Sheron Levar Decker. *Detection of bursty and emerging trends towards identification of researchers at the early stage of trends*. PhD thesis, uga, 2007.

[14] Jörg Diederich, Wolf-Tilo Balke, and Uwe Thaden. Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. In Edie M. Rasmussen, Ray R. Larson, Elaine G. Toms, and Shigeo Sugimoto, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007, Vancouver, BC, Canada, June 18-23, 2007, Proceedings*, page 505. ACM, 2007. . URL https://doi.org/10.1145/1255175.1255305.

[15] Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann, and Susie Stephens. The semantic web in action. *Scientific American*, 297(6):90–97, 2007.

[16] Javier David Fernandez Garcia, Elmar Kiesling, Sabrina Kirrane, Julia Neuschmid, Nika Mizerski, Axel Polleres, Marta Sabou, Thomas Thurner, and Peter Wetz. Propelling the potential of enterprise linked data in austria. roadmap and report., 2016. URL https://www.linked-data.at/wp-content/uploads/2016/12/propel_book_web.pdf.

[17] Volker Frehe, Vilius Rugaitis, and Frank Teuteberg. Scientometrics: How to perform a big data trend analysis with scienceminer. In Erhard Plödereder, Lars Grunske, Eric Schneider, and Dominik Ull, editors, *44. Jahrestagung der Gesellschaft für Informatik, Informatik 2014, Big Data - Komplexität meistern, 22.-26. September 2014 in Stuttgart, Deutschland*, volume P-232 of *LNI*, pages 1699–1710. GI, 2014. URL https://dl.gi.de/20.500.12116/2780.

[18] Koen Frenken, Sjoerd Hardeman, and Jarno Hoekman. Spatial scientometrics: Towards a cumulative research program. *J. Informetrics*, 3(3):222–232, 2009. . URL https://doi.org/10.1016/j.joi.2009.03.005.

[19] Birte Glimm and Heiner Stuckenschmidt. 15 years of semantic web: An incomplete survey. *KI*, 30(2):117–130, 2016. . URL https://doi.org/10.1007/s13218-016-0424-1.

[20] Thomas Hofmann. Probabilistic latent semantic indexing. *SIGIR Forum*, 51(2):211–218, 2017. . URL https://doi.org/10.1145/3130348.3130370.

[21] William W. Hood and Concepción S. Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001. . URL https://doi.org/10.1023/A:1017919924342.

[22] Yingjie Hu, Krzysztof Janowicz, Grant McKenzie, Kunal Sengupta, and Pascal Hitzler. A linked-data-driven and semantically-enabled journal portal for scientometrics. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul T.

Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, volume 8219 of *Lecture Notes in Computer Science*, pages 114–129. Springer, 2013. . URL https://doi.org/10.1007/978-3-642-41338-4_8.

[23] Yookyung Jo, Carl Lagoze, and C. Lee Giles. Detecting research topics via the correlation between graphs and texts. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 370–379. ACM, 2007. . URL https://doi.org/10.1145/1281192.1281234.

[24] Jon M. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373–397, 2003. . URL https://doi.org/10.1023/A:1024940629314.

[25] Nancy L Leech and Anthony J Onwuegbuzie. A typology of mixed methods research designs. *Quality & quantity*, 43(2): 265–275, 2009. .

[26] Loet Leydesdorff and Staša Milojević. Scientometrics. *arXiv preprint arXiv:1208.4566*, 2012.

[27] Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar. Exploring your research: Sprinkling some saffron on semantic web dog food. In *Semantic Web Challenge at the International Semantic Web Conference*, volume 117, pages 420–435. Citeseer, 2010.

[28] Satoshi Morinaga and Kenji Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 811–816. ACM, 2004. . URL https://doi.org/10.1145/1014052.1016919.

[29] Mizuki Oka, Hirotake Abe, and Kazuhiko Kato. Extracting topics from weblogs through frequency segments. In *Proceedings of WWW 2006 annual workshop on the weblogging ecosystem: aggregation, analysis, and dynamics*, 2006.

[30] Francesco Osborne and Enrico Motta. Klink-2: Integrating multiple web sources to generate semantic topic networks. In Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul T. Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, volume 9366 of *Lecture Notes in Computer Science*, pages 408–424. Springer, 2015. . URL https://doi.org/10.1007/978-3-319-25007-6_24.

[31] Francesco Osborne, Enrico Motta, and Paul Mulholland. Exploring scholarly data with rexplore. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 460–477. Springer, 2013. . URL https://doi.org/10.1007/978-3-642-41335-3_29.

[32] Francesco Osborne, Angelo Antonio Salatino, Aliaksandr Birukou, and Enrico Motta. Automatic classification of springer nature proceedings with smart topic miner. In Paul T. Groth, Elena Simperl, Alasdair J. G. Gray, Marta Sabou, Markus Krötzsch, Freddy Lécué, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, volume 9982 of *Lecture Notes in Computer Science*, pages 383–399, 2016. . URL https://doi.org/10.1007/978-3-319-46547-0_33.

[33] Francesco Osborne, Henry Muccini, Patricia Lago, and Enrico Motta. Reducing the effort for systematic reviews in software engineering. *CoRR*, abs/1908.06676, 2019. URL http://arxiv.org/abs/1908.06676.

[34] Sergey Parinov and Mikhail R. Kogalovsky. Semantic linkages in research information systems as a new data source for scientometric studies. *Scientometrics*, 98(2):927–943, 2014. . URL https://doi.org/10.1007/s11192-013-1108-3.

[35] Angelo Antonio Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. The computer science ontology: A large-scale taxonomy of research areas. In Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II*, volume 11137 of *Lecture Notes in Computer Science*, pages 187–205. Springer, 2018. . URL https://doi.org/10.1007/978-3-030-00668-6_12.

[36] Angelo Antonio Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. Classifying research papers with the computer science ontology. In Marieke van Erp, Medha Atre, Vanessa López, Kavitha Srinivas, and Carolina Fortuna, editors, *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*, volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2180/paper-55.pdf.

[37] Angelo Antonio Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. The CSO classifier: Ontology-driven detection of research topics in scholarly articles. In Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt, editors, *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, volume 11799 of *Lecture Notes in Computer Science*, pages 296–311. Springer, 2019. . URL https://doi.org/10.1007/978-3-030-30760-8_26.

[38] Thomas Schandl and Andreas Blumauer. Poolparty: SKOS thesaurus management utilizing linked data. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Proceedings, Part II*, volume 6089 of *Lecture Notes in Computer Science*, pages

421–425. Springer, 2010. . URL https://doi.org/10.1007/978-3-642-13489-0_36.

[39] J Michael Schultz and Mark Liberman. Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA broadcast news workshop*, pages 189–192. San Francisco: Morgan Kaufmann, 1999.

[40] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 990–998. ACM, 2008. . URL https://doi.org/10.1145/1401890.1402008.

[41] Nees Jan van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010. . URL https://doi.org/10.1007/s11192-009-0146-3.

# Appendix

## A. Additional results

Table 7
Extended topics: Top-40 multiwords in Poolparty and top-40 topics in Rexplore (MV) and Saffron

| | Poolparty | Rexplore | Saffron |
|---|---|---|---|
| 1 | semantic web | semantic web | semantic web |
| 2 | linked data | ontology | rdf data |
| 3 | knowledge base | artificial intelligence | linked data |
| 4 | web service | information retrieval | natural language |
| 5 | web semantics | query languages | data source |
| 6 | data source | linked data | reasoning task |
| 7 | data sets | knowledge based systems | machine learning |
| 8 | description logic | natural language processing systems | query execution |
| 9 | on the web | Computational Linguistics | owl S |
| 10 | natural language | formal logic | ontology engineering |
| 11 | use cases | data mining | rdf Graph |
| 12 | social network | knowledge representation | User Interface |
| 13 | query languages | human computer interaction | service description |
| 14 | search engines | ontology matching | open data |
| 15 | query answering | web ontology language (OWL) | semantic search |
| 16 | user interfaces | description logic | query processing |
| 17 | semantic annotation | linked open data (LOD) | keyword search |
| 18 | information retrieval | data integration | keyword query |
| 19 | web of data | web services | owl ontology |
| 20 | open data | resource description framework (RDF) | web service |
| 21 | data models | security of data | query language |
| 22 | semantic search | ontology engineering | data model |
| 23 | ontology matching | semantic search/similarity | ontology matching |
| 24 | information systems | social networks | web data |
| 25 | query processing | SPARQL | federated query |
| 26 | machine learning | data privacy | stream processing |
| 27 | ontology language | recommender systems | relational database |
| 28 | semantic web service | electronic commerce | blank node |
| 29 | linked open data | sensors | information retrieval |
| 30 | logic programming | ubiquitous computing | ontology language |
| 31 | knowledge management | semantic information | description logic |
| 32 | data integration | SPARQL queries | federated query processing |
| 33 | ontology engineering | pattern recognition | semantic similarity |
| 34 | semantic technology | data visualization | object property |
| 35 | ontology alignment | knowledge acquisition | ontology editing |
| 36 | web search | information technology | social medium |
| 37 | web data | mobile devices | knowledge base |
| 38 | structured data | wikipedia | web technology |
| 39 | case studies | machine learning | web semantics |
| 40 | similarity measures | DBpedia | ontology development |

Table 8
Extended topics extracted by two or more tools

| Topic | PoolParty | Rexplore | Saffron |
|---|---|---|---|
| description logic | ✓ | ✓ | ✓ |
| information retrieval | ✓ | ✓ | ✓ |
| knowledge base | ✓ | ✓ | ✓ |
| linked data | ✓ | ✓ | ✓ |
| machine learning | ✓ | ✓ | ✓ |
| natural language processing | ✓ | ✓ | ✓ |
| ontology engineering | ✓ | ✓ | ✓ |
| semantic search | ✓ | ✓ | ✓ |
| semantic web | ✓ | ✓ | ✓ |
| web services | ✓ | ✓ | ✓ |
| ontology matching | ✓ | ✓ | ✓ |
| query languages | ✓ | ✓ | ✓ |
| data integration | ✓ | ✓ | - |
| data source | ✓ | - | ✓ |
| linked open data (LOD) | ✓ | ✓ | - |
| ontology language | ✓ | - | ✓ |
| open data | ✓ | - | ✓ |
| query processing | ✓ | - | ✓ |
| social networks | ✓ | ✓ | - |
| user interfaces | ✓ | - | ✓ |
| web data | ✓ | - | ✓ |
| web semantics | ✓ | - | ✓ |
| web ontology language (OWL) | - | ✓ | ✓ |

Table 9

Extended topics extracted by only one tool

| Topic | PoolParty | Rexplore | Saffron |
|---|---|---|---|
| artificial intelligence | - | ✓ | - |
| blank node | - | - | ✓ |
| case studies | ✓ | - | - |
| Computational Linguistics | - | ✓ | - |
| data mining | - | ✓ | - |
| data models | ✓ | - | - |
| data privacy | - | ✓ | - |
| data sets | ✓ | - | - |
| data visualization | - | ✓ | - |
| DBpedia | - | ✓ | - |
| electronic commerce | - | ✓ | - |
| engineering data model | - | - | ✓ |
| federated query | - | - | ✓ |
| federated query processing | - | - | ✓ |
| formal logic | - | ✓ | - |
| human computer interaction | - | ✓ | - |
| information systems | ✓ | - | - |
| information technology | - | ✓ | - |
| keyword query | - | - | ✓ |
| keyword search | - | - | ✓ |
| knowledge acquisition | - | ✓ | - |
| knowledge management | ✓ | - | - |
| knowledge representation | - | ✓ | - |
| logic programming | ✓ | - | - |
| mobile devices | - | ✓ | - |
| object property | - | - | ✓ |
| on the web | ✓ | - | - |
| ontology | - | ✓ | - |
| ontology alignment | ✓ | - | - |
| ontology development | - | - | ✓ |
| ontology editing | - | - | ✓ |
| owl ontology | - | - | ✓ |
| owl S | - | - | ✓ |
| pattern recognition | - | ✓ | - |
| query answering | ✓ | - | - |
| rdf data | - | - | ✓ |
| rdf Graph | - | - | ✓ |
| reasoning task | - | - | ✓ |
| recommender systems | - | ✓ | - |
| relational database | - | - | ✓ |
| resource description framework (RDF) | - | ✓ | - |
| search engines | ✓ | - | - |
| security of data | - | ✓ | - |
| semantic annotation | ✓ | - | - |
| semantic information | - | ✓ | - |
| semantic similarity | - | - | ✓ |
| semantic technology | ✓ | - | - |
| semantic web service | ✓ | - | - |
| sensors | - | ✓ | - |
| service description | - | - | ✓ |
| similarity measures | ✓ | - | - |
| social medium | - | - | ✓ |
| SPARQL | - | ✓ | - |
| SPARQL queries | - | ✓ | - |
| stream processing | - | - | ✓ |
| structured data | ✓ | - | - |
| systems query execution | - | - | ✓ |
| ubiquitous computing | - | ✓ | - |
| use cases | ✓ | - | - |
| web of data | ✓ | - | - |
| web search | ✓ | - | - |
| web technology | - | - | ✓ |
| wikipedia | - | ✓ | - |