Undefined 0 (2018) 1–0
IOS Press

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

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

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. 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 which can offer a quantitative analysis of the research field's evolution. 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 a combined use of both qualitative and quantitative approaches. Concretely, we build on a qualitative analysis of the main seminal papers, which have adopted 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 in the last decade. In this process, we both use the latter for "fact-checking" on the former and also derive key findings in relation to the strengths and weaknesses of top-down and bottom-up approaches to research topic identification. Overall, we provide a reflectional study on the past decade of Semantic Web research, however 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 [18].

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 the research interests evolve and allow the research community to better plan its future work. While, visions of future topics can inspire and channel the work of a research community. As with other research domains, topics and trends analysis is vital for the Semantic Web research community as it helps to identify both under-represented and emerging research topics.

Semantic Web technologies have been an area of intense research in the academic community for almost two decades. During this timeframe, several papers have been published by researchers from the Semantic Web community that endeavor to predict Semantic Web research topics and trends [2,3], or as the research advanced over the years, to analyze these topics and trends [16,19]. In parallel, several researchers from the Semantic Web community [8,22,26,29,30,33] have been 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 [2,3,16,19] 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 trends alone while overlooking under-represented or emerging trends. Also, we suspect that a detailed analysis of the scientific publications in the field over time might reveal the degree to which the expert predictions have come true, or even falsify their predictions, and at the least allow us to quantifiably assess them. These shortcomings could be well-addressed by (semi-) automatic, data-driven approaches, which identify research trends in a *bottom-up* fashion from large corpora.

The primary goal of this paper is to provide a broader and more complete picture of Semantic Web topics and trends in the last decade by relying on both top-down and bottom-up approaches. A secondary goal is to better understand the strengths and weaknesses of these two families of approaches in terms of topic identification (i.e., expert-based versus data-driven). To this end, we adopt a *mixed methods* research methodology [25], which involves the combination of quantitative and qualitative research methods, in order to gain better insights with respect to research conducted within the Semantic Web community.

Concretely, our approach has three main components. Firstly, in a qualitative study we converge the findings of four top-down style seminal papers [2,3, 16,19] at different points in time, into a unified Research Landscape. Secondly, we employ three diverse data-driven quantitative approaches to uncover Semantic Web topics and trends from a corpus of research papers in a bottom-up fashion. The corpus analyzed with these tools covers the academic literature that emerged from five popular international publishing venues for Semantic Web researchers: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), the SEMANTiCS conferences, the Semantic Web Journal (SWJ) and the Journal of Web Semantics (JWS), over a 10 year period from 2006 to 2015 inclusive. Each of the the data-driven approaches employ different techniques for topic extraction and analysis, for instance a handcrafted taxonomy in PoolParty[1] and an automated taxonomy in Rexplore[2] and Saffron[3]. Thirdly, we compare and contrast the topics derived from both the quantitative and qualitative approaches, in order to provide: (i) a broader picture of Semantic Web topics and, in that process (ii) a better understanding of strengths and weaknesses of the various approaches involved.

Based on our analysis we were able to classify research topics into three different groups that indicate: (i) topics that are deemed important by experts and frequently occur in papers published in popular Semantic Web conferences and journals; (ii) topics that experts consider important, however there was not sufficient evidence in the papers to confirm this; and (iii) topics that only some experts highlight as important, however they were strongly represented in the research papers. However, although it was relatively easy to align the topics in the seminar papers to topics identified using the data drive approaches, the trend analysis was not so straightforward. In essence, we discovered that neither trends derived from broad foundational topics nor specific multi-word topics provide enough evidence to confirm expert visions outlined in the seminal papers.

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 overarching methodology that guided our analysis. *Section 4* provides a snapshot of the Semantic Web research community based on the observations of several Semantic Web researches [2,3,16,19]. This is followed by the presentation of the topic analysis of papers published in the core Semantic Web venues over a 10 year period from 2006 to 2015 with PoolParty, Rexplore, and Saffron in *Sections 5-7*. A discussion on the findings of our analysis is presented in *Section 8*. Finally, *Section 9* concludes the paper and presents directions for future work.

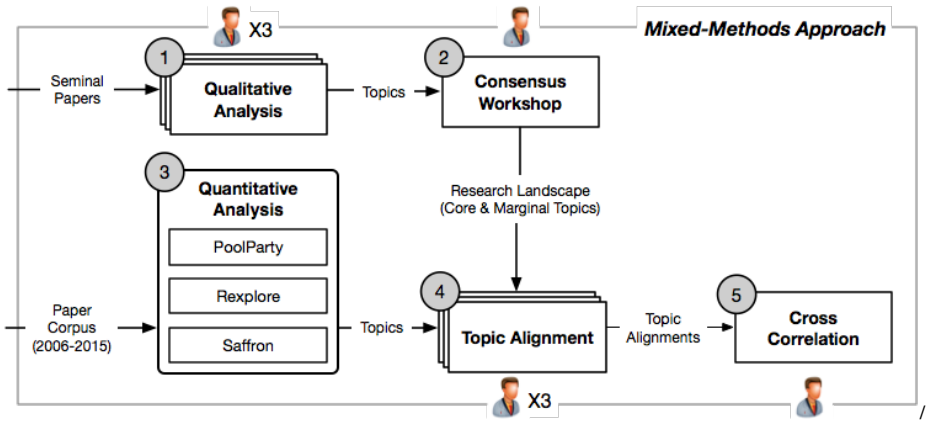


Fig. 1. Overview of the mixed methods-based methodology.

2. Related Work

Detecting topics from 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 [13], social networks [10], blogs [28], emails [27], to name but a few. In this section, we discuss approaches for the more focused task of detecting research topics from scientific literature, and in particular methods employed by the Semantic Web community.

A classical way to model the topics of a document is to extract a list of significant terms [9] (e.g., using tf-idf) and cluster them [34]. Another common solution is the adoption of probabilistic topic models, such as Latent Dirichlet Analysis (LDA) [4] 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 [27] present a method that exploits a Finite Mixture Model to detect research topics and to track the emergence of new topics. Derek et al [14] developed an approach that matches scientific articles with a manually curated taxonomy of topics that is used to analyze topics across different timescales. Chavalarias et al [11] 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. Other approaches exploit the citation graph. For example, Chen et al [12] designed a tool called CiteSpace which combines cocitation analysis and burst detection [24] to identify new emerging trends. While, Jo et al [23] detect topics by combining distributions of terms with the citation graph related to publications containing these terms.

Public tools for the exploration of research data usually identify research areas by using keywords as proxies (e.g., DBLP++ [15], Scival[4]), adopting probabilistic topic models (e.g., aMiner [35]) or exploiting handcrafted classifications (ACM[5], Microsoft Academic Search[6]). However, 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

[1] PoolParty, https://www.poolparty.biz/system-architecture/
[2] Rexplore, https://technologies.kmi.open.ac.uk/Rexplore/
[3] Saffron, http://saffron.insight-centre.org/

[4] http://www.elsevier.com/solutions/scival
[5] https://www.acm.org/publications/class-2012
[6] http://academic.research.microsoft.com/

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 [8] focus on topic extraction and modeling, demonstrating how domain terminology and document keywords can be clustered in order to form semantic concepts that can be used for expert finding. In a related work, Monaghan et al. [26] present their expertise finding platform Saffron 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 [30], an environment for exploring and making sense of scholarly data that integrates statistical analysis, semantic technologies, and visual analytics. Rexplore builds on Klink-2 [29], 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 access 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. While, Parinov and Kogalovsky [33] 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.

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 tend analysis, by comparing and contrasting the results of three data-driven and four expert-based topic-detection approaches in the context of Semantic Web research.

3. Methodology

The primary objective of this paper is to gain a better understanding of the topics and the trends in the Semantic Web community over a ten year period from 2006-2015. Towards this end, we use a *mixed methods* approach to topic analysis. The thesis being that a combination of different research methods will allow us to gain a more comprehensive view of the topics and trends within the community. In this section, we describe our overall mixed methods methodology (Section 3.1) and then explain its individual quantitative and qualitative stages in the rest of the subsections.

3.1. A mixed methods approach to topic analysis

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. In our work we adopted the mixed methods approach that is illustrated in *Figure* 1 and described next.

Qualitative research was employed to manually extract key research topics from four seminal papers that discuss the past, present and future of the Semantic Web technology [2,3,16,19] (Section 3.2 and Section 4). The extraction of keywords was performed individually by three authors of this paper (Step 1) and their findings were converged during a consensus workshop (Step 2). The result of this workshop is a high level overview of the Semantic Web research up to 2016, which we call the *Research Landscape* (cf. *Table* 2).

Quantitative research consisted of the use of three different topic analysis tools to automatically extract and rank topics in order of importance, from our Semantic Web papers corpus from 2006 to 2015 inclusive (Section 3.3 and Sections 5-7). The quantitative analysis was performed with three different tools (i.e., PoolParty, Rexplore, and Saffron) that enable users to gain insights on the various research topics that appear in research papers published in popular Semantic Web publishing outlets.

Finally, we *combine the results obtained both with qualitative and quantitative methods* in order to bet-

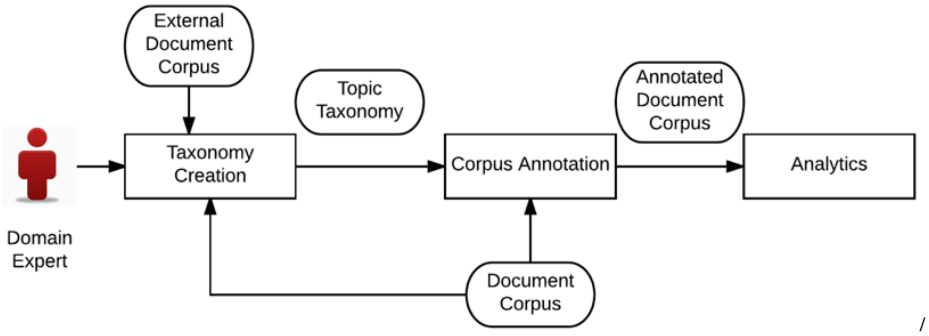


Fig. 2. Conceptual overview of topics detection approaches: main steps and data sources

ter understand the topics that are frequently discussed within the community (Section 3.4). The alignment of the Research Landscape topics to those extracted by quantitative tools was performed individually by three authors of this paper (Step 4) and then their findings were cross-correlated (cf. *Tables* 3- 6) to derive the main findings of this paper (Step 5, Sections 8 and 9).

3.2. Qualitative study of seminal papers

In terms of qualitative analysis, the following approach was used to identify the topics mentioned in the seminal papers [2,3,16,19]. First (Step 1 in Figure 1), each paper was read by three of this paper's authors who identified the main technical keywords and topic descriptions mentioned in the paper (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. Second, the extracted keywords were grouped into broader topic areas by each author (e.g., knowledge structures and modeling). Third, the results of the separate analysis were discussed and aligned during a consensus workshop (Step 2 in Figure 1). The primary outcome of the qualitative analysis was the development of a unified *Research Landscape* (shown in *Table* 2) based on the alignment of the topics mentioned in the four seminal papers.

3.3. Quantitative analysis of research papers

For the quantitative analysis, rather than using a single tool we leveraged three different tools (i.e., PoolParty, Rexplore, and Saffron) that employ different approaches to topic extraction (Step 3 in Figure 1). Before describing the different quantitative approaches employed by the aforementioned tools, we first describe the conceptual steps that are typically followed when it comes to data-driven topic analysis. The workflow for topic extraction and analysis (as depicted in *Figure* 2) typically follows three sequential steps, namely `taxonomy creation`, `corpus annotation`, and `analytics`.

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, trend detection analytics, expert profiling and recommendations.

A tabular overview of the approaches adopted by PoolParty, Rexplore, and Saffron with respect to these main steps is presented in *Table* 1, while a highlevel overview of each of the tools is presented below. The corpus analyzed with these tools covers the academic literature that emerged from five popular international publishing venues for Semantic Web researchers: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), the SEMANTiCS conferences, the Semantic Web Journal (SWJ) and the Journal of Web Semantics (JWS), over

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	SW Venues 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	(MV) SW Venues 2006-2015 (FSW) Scopus 2006-2015	Automatic (abstracts, titles, keywords)	Number of papers and citations associated with a topic	Taxonomy learning, expert profiling
Saffron	Automatic from document corpus through clustering of co-occurrences	Flat list of terms	SW Venues 2006-2015	Automatic (full-text)	Topic frequency in text	Taxonomy learning, expert profiling

a 10 year period from 2006 to 2015 inclusive. In the case of ISWC, ESWC and the JWS the papers spanned the entire timeframe, however it is worth noting that the SEMANTICs and SWJ papers were only available from 2009 and 2010 respectively.

PoolParty is a semantic technology suite that supports the analysis of documents guided by a taxonomy of the domain of interest. In the case of the analysis described in this paper the taxonomy was manually created from conference and journal metadata (i.e., call for papers, sessions, tracks, special issues etc.). Our hypothesis was that a hand crafted topic hierarchy based on the information extracted from said sources should broadly speaking reflect the topics that the community are interested in and also what existing researchers are actively working on. The PoolParty analysis was conducted over the full text of the research articles, from ISWC, ESWC, Semantics, the JWS and the SWJ. Although the PoolParty suite is capable of extracting topics automatically, following our purely manual approach we simply used the tool to annotate the documents based on the manually generated taxonomy.

Rexplore is an interactive environment for exploring scholarly data that leverages data mining, semantic technologies and visual analytics techniques [30]. In the context of this paper, we used Rexplore technologies for tagging research papers in two datasets with relevant research topics from the Computer Science Ontology (CSO), an automatically generated ontology of research areas, and produce a number of analytics. The approach for tagging the publications took into consideration their title, keywords and abstract and is a slight variation of the method adopted by Springer Nature for characterizing semi-automatically their Computer Science proceedings [31].

Saffron is a term and taxonomy extraction tool, which is primarily used for expert finding [5]. Saffron facilitates the extraction of knowledge from text in a fully automatic manner, by leveraging both text mining and Linked Data principles. Its algorithms include: key-phrase extraction, entity linking, taxonomy extraction, expertise mining, and data visualization. The Saffron analysis, which was conducted over the full text of the papers (as per the PoolParty analysis) is based on a taxonomy that is automatically generated during the analysis of the Semantic Web corpora (composed of papers from ISWC, ESWC, Semantics, the JWS and the SWJ) that was specifically constructed for our analysis.

3.4. 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. The output of each of the three data-driven approaches was mapped by one of the authors to the topics of the Research Landscape (Step 4 in Figure 1). 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.

Individual topic alignments were cross-checked by the two additional authors to reduce any bias and further discussed during an analysis and cross-correlation workshop (Step 5 in *Figure* 1). The results of this workshop are depicted in *Tables* 3- 6.

4. Qualitative Study: Research Topic Identification from Seminal Papers

In the Semantic Web area, a handful of well-known papers identify research topics and discuss trends within the community [2,3,16,19]. Some of these papers predict future topics [2,3], while others reflect on research topics in the past years or in the present [3,16,19]. In this section, we briefly introduce the seminal papers before presenting the topics mentioned in these papers in a format that allows them to be used as a basis for comparison with topics identified via the data-driven methods.

4.1. The seminal papers

At the turn of the millennium (2001), Berners-Lee et al. [2] coined the term "Semantic Web" and set a research agenda for a young and multi-disciplinary research field around a handful of topics. Six years later, Feigenbaum et al. [16] 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. Later in 2016, two important papers were published by [3,19], coinciding with the 15-year anniversary of the Semantic Web community. Bernstein et al. [3] provide their vision of Semantic Web 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. Glimm and Stuckenschmidt [19] in turn, look 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-centric analysis of research topics and trends, which was performed completely manually.

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 quantitative, 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 three or four of the seminal papers (cf. *Table* 2). All four papers agree on the following eight core research topics:

Knowledge representation languages and standards, such as XML, RDF and a so-called Semantic Web language, were considered crucial to enabling the vision of intelligent software agents by Berners-Lee et al. [2]. Work on the development of web-based knowledge representation languages (now also including OWL) continued over the next 7 years [16]. By 2016 this was seen as a core line of research extending also to the standardization of representation languages for services [3,19]. As for the future, Bernstein et al. [3] 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. [2] consider knowledge structures such as ontologies, taxonomies and vocabularies as essential compo-

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

	Berners-Lee et al. [2] Future	Feigenbaum et al. [16] Past (2000-2007)	Glimm and Stuckenschmidt [19] Past (2000-2016)	Bernstein et al. [3] Past (2000-2016)	Bernstein et al. [3] Future from 2016
Core topics	knowledge representation languages and standards	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, knowledge graphs	ontologies and modeling, (PR) knowledge graphs	-
	logic and reasonings	logic and reasonings	logic and reasonings	logic and reasoning	-
	search and question answering	(ranking)	search, retrieval and ranking	(PR) question answering systems	-
	(data integration)	(ontology matching)	data integration & matching and integration	(PR) needs-based, lightweight data integration	integration of heterogeneous data
	proof & trust	privacy, trust, access control	security, trust, provenance	personal information, privacy	trust & data provenance (representation, assessment)
	databases	semantic web databases	-	database management systems	-
	decentralization	(decentralization)	distributed data storage and federated query processing	vastly distributed heterogeneous data	(decentralization)
	-	query language (SPARQL)	query processing (SPARQL, federated query processing)	developing efficient query mechanisms	-
	-	(linked data, DBpedia)	linked data	(PR) linked data (open government data), (social data)	-
	(machine learning, prediction, analysis, automatic report)	knowledge extraction and discovery	automatic knowledge acquisition	latent semantics, knowledge acquisition, ontology learning	-
Marginal topics	intelligent software agents	-	-	multilingual intelligent agents	-
	(semantic web services)	-	semantic web services	-	-
	(Internet of Things)	-	-	-	high volume and velocity of data, e.g., streaming & sensor data
	-	visualization	user interfaces and annotation	-	-
	-	(scalability, efficiency, robust semantic approaches)	-	-	scale changes drastically
	-	change management and propagation	-	-	-
	-	(social semantic web, FOAF)	-	-	
	-	-	-	-	data quality, e.g., representation, assessment

nents of the Semantic Web. Follow up papers confirm active research on the creation of ontologies [3,16] entailing research topics such as modeling patterns, large-scale modeling efforts, and knowledge acquisition [19]. Bernstein et al. [3] introduce knowledge graphs as novel knowledge representation structures. Based on their trend analysis in terms of number of publications in each topic published at ISWC conferences, Glimm and Stuckenschmidt [19] conclude that research on both language standards and ontologies has diminished in impact over time.

Logic and Reasoning. Berners-Lee et al. [2] conjured that inference rules and expressive rule languages will enable logic-based automated reasoning on the Semantic Web. Their prediction was abundantly confirmed in follow-up papers: Feigenbaum et al. [16] reporting work on the development of inference engines for reasoning by 2007; Glimm and Stuckenschmidt [19] discussing several sub-topics in this area such as scalable and efficient reasoning, non-standard reasoning or the combination of logics with non-logical reasoning paradigms; and Bernstein et al. [3] confirming work on developing tractable and efficient reasoning mechanisms. The quantitative analysis of ISWC papers reported in [19] suggests that the number of papers focusing on reasoning was relatively stable from 2002 to 2012, however experienced a slight decline between 2012 and 2014.

Search, retrieval, ranking, question answering. Besides intelligent agents, Berners-Lee et al. [2] predicted that search and question answering programs would also benefit from the Semantic Web. In 2007 Feigenbaum et al. [16] indirectly refer to this topic in the context of ranking, however this research topic becomes increasingly important according to papers published in 2016: Glimm and Stuckenschmidt

[19] identify sub-topics such as search algorithms, domain specific search engines and natural language access to linked data. While, Bernstein et al. [3] 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. [2]. 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 [16]. By 2016, the community explored subtopics such as data and knowledge integration, ontology matching (schema matching) and entity matching (record linkage) [19] with a new trend towards needsbased, lightweight data integration [3]. Work on this topic regularly appeared at ISWC conferences during 2002-2014 [19]. For the future, Bernstein et al. [3] discuss the need to integrate heterogeneous data as part of the broader topic of data management.

Privacy, Trust, Security, Provenance. Berners-Lee et al. [2] envision proofs and digital signatures as key aspects of the Semantic Web in order to enable more trustworthy data exchange. While the topic of privacy was only vaguely mentioned in 2007 [16], it became well established by 2016 and covered topics such as security, trust and provenance [19]. According to Bernstein et al. [3] 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. [2], Feigenbaum et al. [16] 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. [3] many of these tools evolved into commercial tools by 2016.

Distribution, decentralization, federation. Berners-Lee et al. [2] envisioned that the Semantic Web would be as decentralized as possible, bringing new interesting possibilities at the cost of losing consistency. Feigenbaum et al. [16] exemplified one of these novel scenarios by mentioning FOAF as an example of a decentralized social-networking system. In turn, Glimm and Stuckenschmidt [19] identified distributed data storage and federated query processing as a primary goal of Linked Data. Finally, Bernstein et al. [3] 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. [2], but were mentioned by the other three papers. These are:

Query Languages and Mechanisms. By 2007, research also focused on the development of query languages, most notably SPARQL [16] and its evolution towards topics such as federated query processing techniques [19] and developing efficient query mechanisms [3]. Work on query processing has been increasingly published at ISWC over time [19].

Linked Data. By mentioning DBpedia, Feigenbaum et al. [16] intuitively pointed to the future research topic of Linked Data. This topic became well established by 2016, with key sub-topics including publishing Linked Data (tools and guidelines) as well as accounts of concrete data publication projects [19]. 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 [3].

Knowledge extraction, discovery and acquisition. In 2007, Feigenbaum et al. [16] hint at this topic with terms such as machine learning, prediction and analysis. By 2016, knowledge extraction and discovery emerged as a field of its own focusing on knowledge acquisition, information extraction from text and general purpose knowledge bases and exploring techniques from data mining and machine learning among others [19]. Automatic knowledge acquisition was boosted by more powerful statistical and machine learning approaches as well as improved computational resources [3]. For the future, Bernstein et al. [3] 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.

4.3. Marginal topics from the seminal papers

Our analysis also identified several marginal topics, mentioned by individual papers, or by a maximum of two of the seminal papers. These topics, which are also presented in *Table* 2 are discussed here:

Intelligent software agents. The underpinning theme of Berners-Lee et al. [2]'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. [3] discuss work on training conversational intelligent agents based on multilingual textual data on the web.

Semantic Web Services. Berners-Lee et al. [2] also envisioned the applicability of Semantic Web technologies for advertising and discovering web-services. This intuition was the precursor of the *Semantic Web Services* research field established a few years later, which focused on semantic service description, search, matchmaking, automatic composition and execution [19].

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. [2]. 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. [3] 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.

Human-Computer Interaction. Feigenbaum et al. [16] mention visualization as features of user-centric applications, while Glimm and Stuckenschmidt [19] identify an emerging *user interfaces and annotation* topic, which investigates topics such as humans in the loop, making use of human input, user interfaces for the Semantic Web, and involving users in annotation tasks.

Scalability, efficiency and robustness. Feigenbaum et al. [16] position *scalability, efficiency and robust semantic approaches* as key factors needed to address semantic web challenges, in particular integration, knowledge management and decision support. In turn, Bernstein et al. [3] recognize that new research is needed given that the *scale changes drastically*.

Change management and propagation. Feigenbaum et al. [16] 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. [16] intuitively pointed to the future research topic on the *Social Semantic Web*.

Data quality Under the heading of data management, Bernstein et al. [3] 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 paper focus primarily of research topic identification, they also offer some hints on the way these topics evolve over time (i.e., trends). We discuss this aspect of the seminal papers in this section.

In 2001, Berners-Lee et al. [2], used a fictitious scenario to describe the authors' 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. [16] reflected on the ideas presented in [2] and highlighted that although the original autonomous agent vision was far from being realised, the technologies themselves were proving to be highly effective in terms of tacking *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 brought about by the ability to link disparate data sources.

In 2016, Glimm and Stuckenschmidt [19] observed a gradual decline in research on *knowledge representation standards and languages*, *knowledge structures and modeling*, and *Semantic Web services*. Newer topics, such as *search, retrieval and ranking*, *privacy, trust, security, provenance* or *user interfaces and annotation* are identified as less prominent but nevertheless maintain a constant presence over the years. Whereas, they highlight that *query languages and mechanisms*, *Linked Data*, and *knowledge extraction, discovery and acquisition* have gained in importance over the years.

Also in 2016, discussing present research topics, Bernstein et al. [3] 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.

In the next sections we continue with describing the results of topic and trend analysis by employing data-driven tools.

5. PoolParty Quantitative Analysis

Although PoolParty is a semantic technology suite primarily targeted toward knowledge organization and management in a business context, in this paper we use PoolParty to manually construct a taxonomy of topics that are of interest to the Semantic Web community, and subsequently use this taxonomy to analyze the topics that are mentioned in papers that are published in prominent Semantic Web venues.

Taxonomy Creation. Even though there are several Natural Language Processing (NLP) tools that can be used to automatically extract topics from text, considering that accuracy was a key factor for this study, we elected to manually generate and curate a list of topics, based on the venue metadata that are relevant for the community. First, we manually extracted the metadata (i.e., call for paper line items and the track, session and invited talk titles, etc..) from the ISWC, ESWC and the SEMANTiCS conference series and the SWJ and the JWS journal websites, from 2006 to 2015 inclusive. Where the metadata was not available we consulted the Internet Archive Wayback Machine. Once all of the metadata was gathered we created a dataset composed of the aforementioned venue metadata. We started by merging the text from the titles of the invited talks, the line items from the call for papers, the various track names and the session titles, per year per venue. We subsequently went through each line item and noted all of the topics (words and phrases that are relevant for the community) that appeared. It is worth noting that more often than not there were multiple topics per line item. Following on from this we merged all of the topics into a single topic list and removed the duplicates, resulting in a dictionary containing 3,421 unique topics. In the spirit of open science all of the metadata gathered was fed into the Scholarly Data [7] initiative, and was used to strengthen the existing ontology [8].

In parallel, we derived a set of foundational technologies (based on existing taxonomies and well known Semantic Web research sub-communities) that reflect key research areas within the Semantic Web community. The complete list of foundational technologies can be seen in *Figure* 3. Finally, we manually created a coarse grained taxonomy from the 3,421 unique dictionary topics, by assigning each topic to one or more foundational technologies worked on by the community. It is worth noting that although many of the dictionary topics could be associated with foundational topics, we noticed that there were a significant number of topics that did not fit into the technological foundations that we had identified. Generally speaking, topics could be grouped into the following seven high level metadata categories: applied (e.g. applications, tools, terms relating to usage), business (e.g. profit, business value, impact), domain (e.g. healthcare, media, education), evaluation/-metrics/studies (e.g. measure, ranking, negative results), foundational technologies (e.g. knowledge representation, data management, security and privacy), methods (e.g. finding, planning, transformation), and standards (e.g. SPARQL, standardization, W3C). It is worth noting that the results presented in this paper are based solely on the topics that could be aligned with the foundational technologies. A wider study on applications and domains is left to future work.

Corpus Annotation. The full text of the papers was extracted from conference proceedings and journals and uploaded to the Semantic Web Company's PoolParty platform. We subsequently used PoolParty to annotate the documents based on the topics appearing in the taxonomy and to count the occurrences of each topic. It is worth noting that, although the PoolParty suite also includes automatic entity extraction techniques, these were not considered during our analysis as the objective was to use a purely hand crafted taxonomy in order to extract key research topics. Additionally, in order to get an indication of the coverage of the technical foundations, across the five venues under analysis, we aggregated the number of occurrences for each of the topics within a given foundation.

Analysis. The chart presented in *Figure* 3 provides details on the % coverage for each of the eighteen

[7] http://www.scholarlydata.org/#resources
[8] http://www.scholarlydata.org/ontology/doc/

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

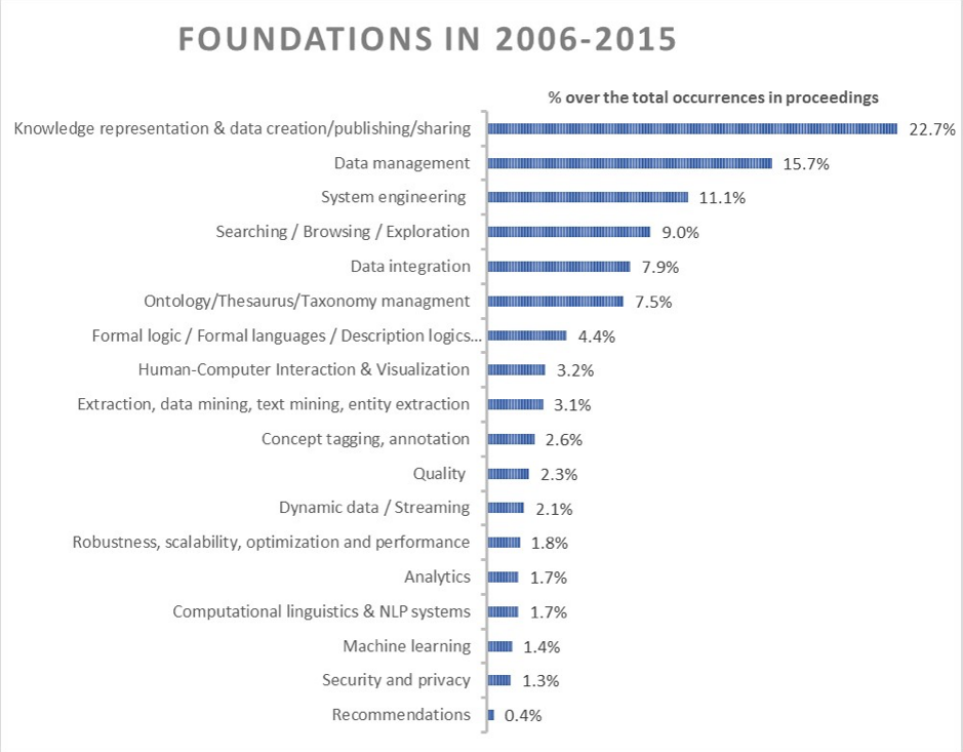
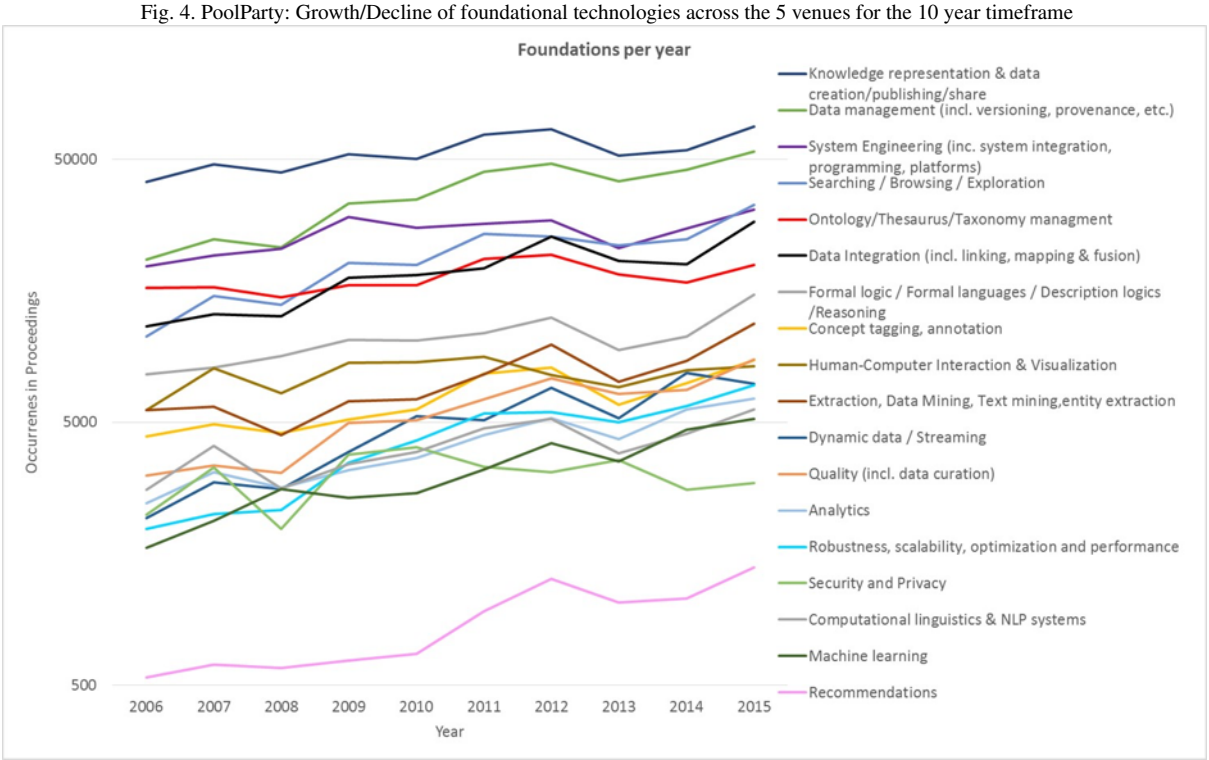
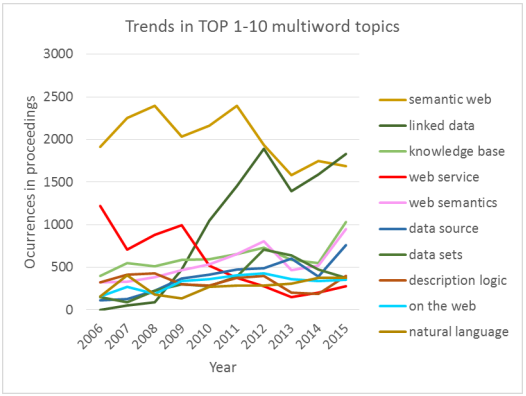
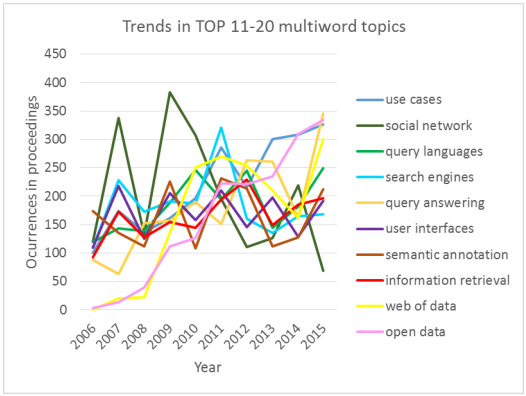


Fig. 4. PoolParty: Growth/Decline of foundational technologies across the 5 venues for the 10 year timeframe





(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.

foundations, across the five venues for the 10-year timeframe under examination. As expected, knowledge representation and data creation/publishing/sharing is the top foundation, with almost 23% of the total occurrences in all documents. Note that 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 management 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 logic and reasoning and concept tagging and annotation represent a modest 4.4% and 2.6% respectively, and cross-topics, such as human computer interaction, machine learning, computational linguistics and NLP, security and privacy, recommendations and analytics are only marginally represented. It is also worth noting that topics that relate to quality, dynamic data and scalability are also under-represented (at around 2%).

In order to gain some insights into the research trends within the Semantic Web community 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, it is worth mentioning that robustness, scalability, optimization and performance, dynamic data/streaming, searching/browsing/exploration and machine learning have increased by more than 200% since 2005. While in contrast, security and privacy and ontology/thesaurus/taxonomy management have had marginal growth of only 30% for the same period.

While, *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.

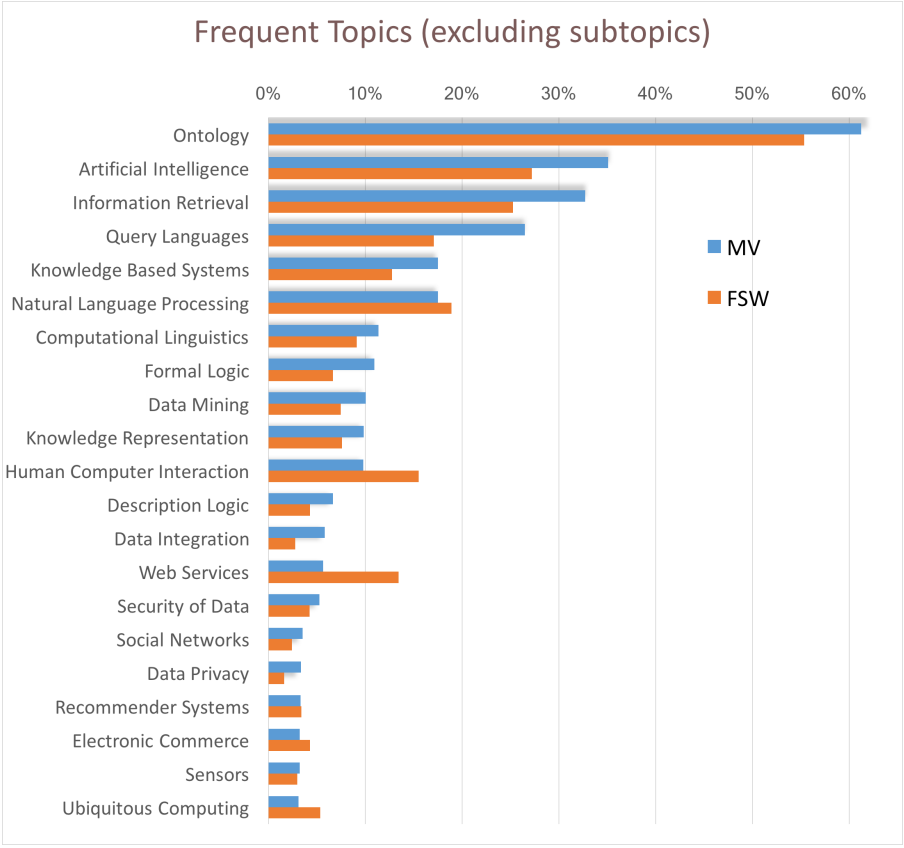
6. Rexplore Quantitative Analysis

Rexplore [30] is a system that implements novel solutions in large-scale data mining, semantic technologies and visual analytics, to provide an innovative environment for exploring and making sense of scholarly data.

Taxonomy Creation. Rexplore uses the Klink-2 algorithm [29] for automatically generating and regularly updating the Computer Science Ontology (CSO)[9], a very large ontology of research areas. Klink-2 processes networks of research entities, such as publications, authors, venues, and technologies, filters out terms that are not research topics, and automatically infers semantic relationships between topics. Some branches of the ontology, regarding in particular Semantic Web and Software Engineering, were also re-

[9]http://cso.kmi.open.ac.uk

Fig. 6. Rexplore: Frequent topics (excluding subtopic) in MV (blue) and FSW (red).



fined by domain experts as a result of previous evaluations [29,32]. The CSO model[10] is an extension of the BIBO ontology[11] which in turn builds on SKOS[12]. It includes three semantic relations: *relatedEquivalent*, which indicates that two topics can be treated as equivalent for the purpose of exploring research data (e.g., Ontology Matching, Ontology Mapping), *skos:broaderGeneric*, which indicates that a topic is a sub-area of another one (e.g., Linked Data, Semantic Web), and *contributesTo*, which indicates that the research outputs of one topic contributes to another (e.g., Ontology, Semantic Web). The last version of the CSO[13] was generated from 16 million publications in the Rexplore dataset and includes about 17k topics linked by 70k semantic relationships. The main root is Computer Science, however the ontology also includes a few secondary roots, such as Linguistics, Geometry, Semantics, and so on.

CSO presents two main advantages over manually crafted categorizations used in Computer Science (e.g., 2012 ACM Classification, Microsoft Academic Search Classification). First, it can characterize higher-level research areas by means of hundreds of subtopics and related terms, which allows Rexplore to map very specific terms to higher-level research areas. Secondly, it can be easily updated by running Klink-2 on a set of new publications. A comprehensive discussion of the advantages of adopting an automatically generated ontology in the scholarly domain can be found in [29].

Since we intend to investigate the trend of Semantic Web topics both within high tier domain conferences and the literature, we generated two different datasets: one focusing on the main Semantic Web venues (MV) and the second on all literature containing reference to the Semantic Web (Full Semantic Web, FSW). *MV* includes 4,734 publications and it was generated by retrieving the set of articles associated with several core

[10] http://technologies.kmi.open.ac.uk/rexplore/ontologies/BiboExtension.owl
[11] http://purl.org/ontology/bibo/
[12] https://www.w3.org/2004/02/skos/
[13] http://cso.kmi.open.ac.uk/download/CSO.nt

Fig. 7. Rexplore: Number of publications associated with eight Semantic Web subtopics in MV.

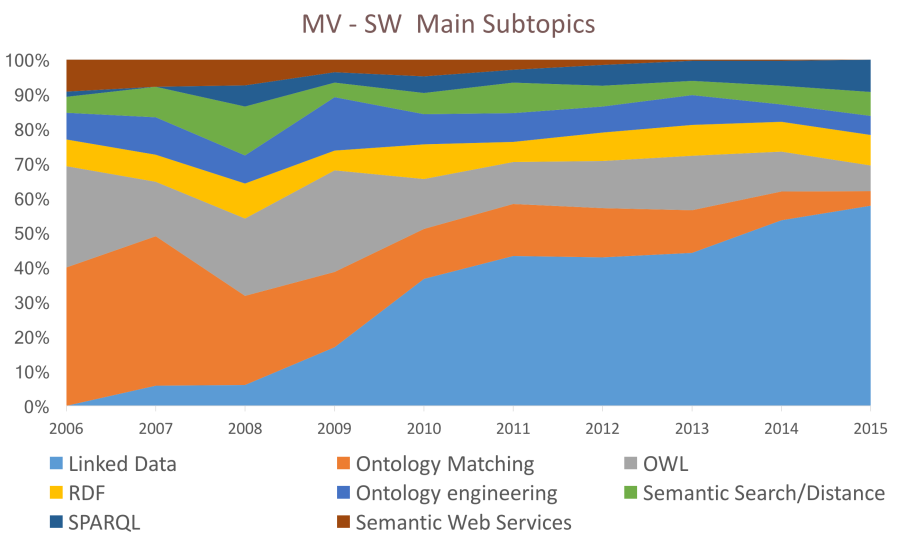
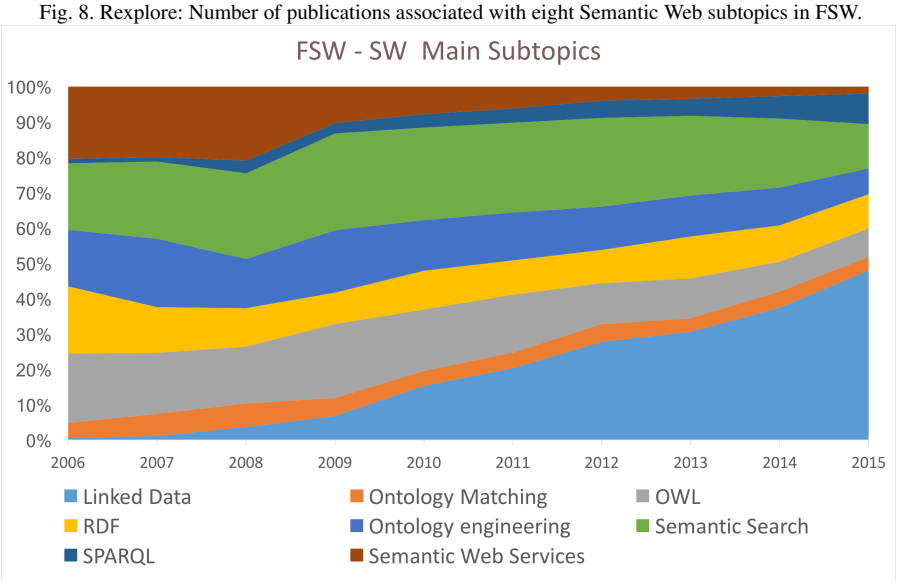


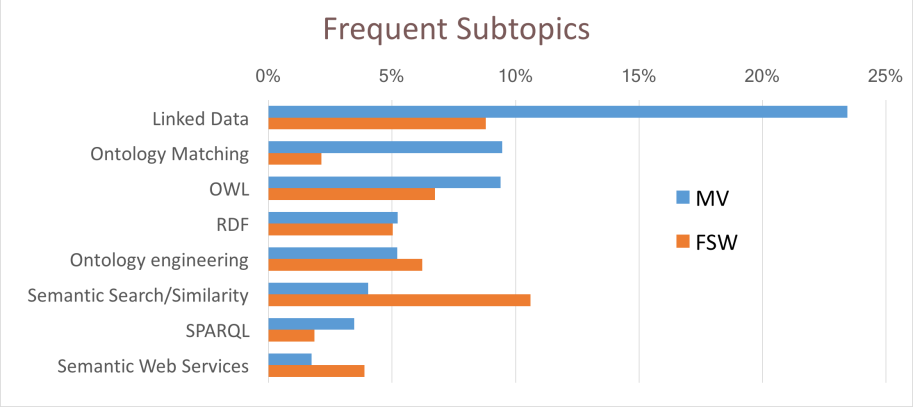
Fig. 8. Rexplore: Number of publications associated with eight Semantic Web subtopics in FSW.



Semantic Web venues: ESWC, ISWC, SEMANTICS, Semantic Web Journal (SWJ), and Journal of Web Semantics (JWS) from the Scopus dataset. *FSW* includes 32,431 publications and it was produced by selecting all articles associated with the topic Semantic Web or with its 96 associated subtopics in CSO (e.g., Linked Data, RDF, Semantic Web Services) in both MV and in the dump containing all Scopus papers in Computer Science for the interval 2000-2015. We label this dataset the Full Semantic Web (FSW). We analyzed both datasets in the 2006-2015 period. In MV the number of publications grows steadily, from less than 50 publications a year to more than 650. While, the number of publications in FSW decreases in the last four years of the period under analysis, because the Scopus Dump adopted in this analysis was complete only until 2013.

Corpus Annotation. The analytics presented in this section were produced by associating to each topic in CSO (e.g., Human Computer Interaction) all publications that contain in the title, in the abstract, or in the keyword field either 1) the topic label, 2) the

Fig. 9. Rexplore: Frequent Semantic Web subtopics in MV (blue) and FSW (red).



alternative labels (e.g., HCI, Human-Computer Interaction), and 3) its subtopics (e.g., Graphical User Interfaces, Tangible User Interfaces, User Centered Design). For example, we will associate with the topic Human Computer Interaction all publications tagged with one of the 293 terms associated with its labels and its subtopics in CSO. This same strategy was implemented in Springer Nature semantic pipeline to annotate proceeding books [31] and was previously evaluated on the software engineering domain, proving that its performance in classifying papers was not statistically significantly different from the ones of six senior researchers [32]. We then counted the number of papers associated with each topics in each year and produced relevant analytics.

Analysis. In this section, we will first analyse the main topics appearing in each dataset and then zoom on Semantic Web subtopics.

Figure 6 shows the main research fields addressed by the Semantic Web papers in both MV 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 MV, 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 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* 9 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 7 and *Figure* 8 show the popularity of the main sub-topics over the years. The two main dynamics, evident in both datasets, are the fading of Semantic Web Services and the rapid grow 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 [2]. However, the number of papers about these topics consistently decreases and from 2013 there are almost no publications about them in the MV 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.

7. Saffron Analysis

Saffron is a tool for extracting knowledge structure from text, by gathering and summarizing expertise information. Although its principal aim is expertise topic extraction, its functions are multiple and its applications various. In terms of functions, it uses NLP and Linked Data techniques to construct domain-specific topical hierarchies, catering for: topic extraction, domain-independent term extraction, topic linking, expert finding and linking. Possible applications include expert search in a research domain or an enterprise environment [6], exploring the triggers of a financial crisis [7], forecasting emerging trends from a scientific area [1], and expert profiling.

Taxonomy Creation. Saffron follows a domain-independent method, 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. The idea behind the domain model is to capture the expertise knowledge of the corpus by isolating the most generic concepts (i.e., the basic level categories of the specific domain). The domain model is represented as a vector of words, collected based on linguistic features, distribution properties, and coherence evaluation. In a second stage, a multi-word term extraction is used to collect intermediate level keyphrases (known as candidate terms). The domain model is subsequently used as a basis to calculate semantic relatedness between the candidate terms and the domain of expertise. We then connect and organize the resulting topics in an understandable and usable manner by means of a hierarchical taxonomy. This is achieved by connecting related concepts and directing the edges from broader to more specific relations using a global generality measure. This eventually yields a taxonomy of topics, where the root is the most generic term of the domain, and the leaves the most specific terms. The construction of topic hierarchies from corpora is discussed in detail in [5].

Corpus Annotation. In order to prepare the data to be ingested by Saffron, all collected proceedings were split into single paper documents. The papers, all available as pdfs, were subsequently converted to raw text and cleansed of any special or bad characters as well as badly converted images and figures, which could result in noise for the analysis or lead to a failure of the system. Furthermore, the reference section of every paper was removed in order to reduce potential noise caused by the list of authors' names, conferences, dates, etc... After term extraction and ranking, the resulting topics are connected together by directed edges to form a hierarchical taxonomy, represented as a graph. We visualize the results of the Saffron analysis with Cytoscapes, an open source software tool for complex networks graph visualization[14]. 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.

Analysis. The size of the nodes in the graph is related to the number of edges connected to them. We can see in Figure 10 topics that stand out from the analysis. Not surprisingly, the first and predominant one (i.e., the root of the taxonomy) over the ten years period is the Semantic Web topic itself. Around it, we observe several main clusters with major keyphrases. The main (i.e., the most connected) keywords are 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 query (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 query of information such as Semantic and Keyword Search, Keyword Query, Semantic Similarity or Information Retrieval. Natural Language is also connected to another dominating subject that is the one of Machine

[14] see http://www.cytoscape.org/

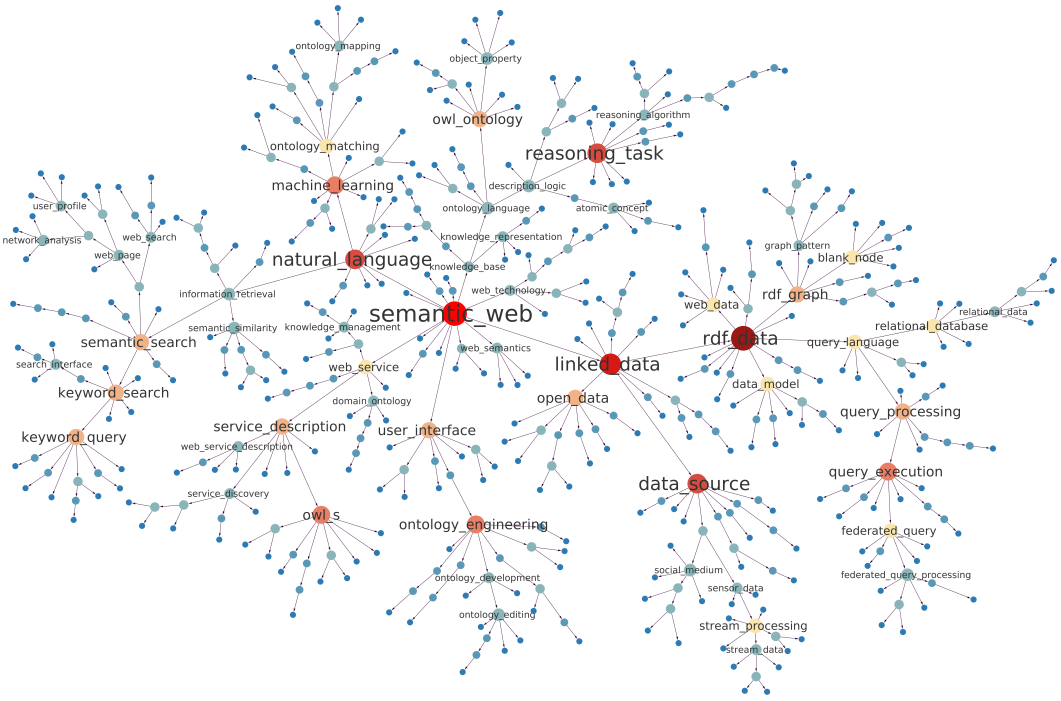
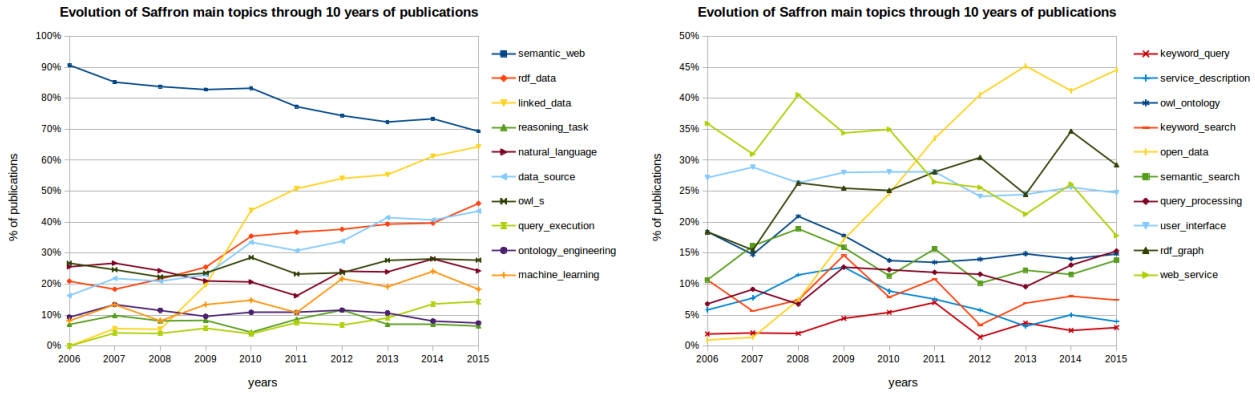


Fig. 10. Saffron: Taxonomy of Semantic Web topics.



(a) top 10 topics (b) top 11-20 topics

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

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). Around the Ontology Engineering node, User Interface seems to be an important related matter, as well as tasks like Ontology Development and Ontology Editing.

Given that Saffron establishes and calculates relations between the topics, it thus puts the focus on the terms that share the most connections with other domain specific words. The most connected terms are

considered the most influential ones as they are used in the context of many other domain-related topics.

Following the processing of the corpus of ten years of Semantic Web publications, we selected the 20 first main topics extracted by Saffron (i.e., the most connected ones), with the aim of observing the distribution of their use in the documents across the years. The two charts in Figure 11 show the percentage of documents mentioning the aforementioned topics (i.e., the number of documents where they appear at least once), per year.

The first thing that catches the eye when looking at the figure is the fact that *Semantic Web* as a topic is on the decline. When speculating as to what this means, on the surface clearly Semantic Web as a topic is mentioned less in more recent papers. However the reasons could be manifold, for example the term/field may be so established that the community has established confidence to "not have to name it" in their papers, or that the community is trying to re-brand their research in terms of new terms such as *Linked Data*. Both would be possible explanations, a more in depth understanding of which would certainly require more analysis, but the speculation about it alone might already bring about a debate concerning our self-understanding to the community. Emerging or new terms and topics, such as said "Linked Data" are easier to assess and judge.

On the graph showing the first 10 topics (ordered by the number of connections) extracted from Saffron (Figure 11), the biggest and the most significant progression is striking in 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 multiplied by 9, to eventually reach 64% of the distribution in the documents by 2015. Another dramatic increase among the top 20 terms is the Open Data topic, remarkably developing from about 1% in 2006/2007 to 45% in 2015. This is particularly meaningful, and demonstrates the emergence and widening use of Linked Data and Open Data through the years. We can notice from the graphs another emerging topic which is slowly but gradually gaining importance, namely Query Execution, appearing in 2015 in 15% of the publications. RDF Data and Data Source are topics whose popularity have also increased quite significantly, multiplying their presence by more than twice in the documents since 2006. This shows that the Semantic Web community has placed a stronger focus on those concepts through time. Other topics whose popularity increased with time 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).

On the contrary, the very hot topic of Semantic Web itself is less cited in publications, with a decrease of 20% between the beginning and the end of the studied time frame. Its high degree of genericity may account for why it is progressively less used in the publications: specific facets of the domain can be mentioned without pointing to the higher level topics. It is still the most used keyphrase nonetheless, reaching 91% of distribution in the documents in 2006 and lowering down to 70% in 2015.

Among the topics experiencing strong variations through time, 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. Other main terms decreased with time, although in a less pronounced way, like Semantic Search which experiences 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.

8. Topic Alignment and Findings

In this section, we compare and contrast 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 Sections 5-7). For this analysis, in order to cover a wider spectrum of topics, we consider the top 40 multiword topics from PoolParty, Rexplore and Saffron (see Table 7 in Appendix A). 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 40 topics instead. The analysis described herein is summarized in *Table* 3, which depicts the core research topics identified in the seminal papers and their coverage by the data-driven ap-

proaches, and *Table* 4, which presents the marginal research topics. Finally, the research topics covered by the data-driven approaches that were not identified by the seminal papers are presented in *Table* 5, while the visionary research topics from the seminal papers and their coverage by the data-driven approaches are detailed in *Table* 6.

8.1. Core and marginal topic analysis

The results presented below are based on a comparison between the core and marginal topics mentioned in the seminal Semantic Web papers and the major topics uncovered by PoolParty, Rexplore and Saffron.

PoolParty Analysis: It is not surprising that it was easy to align the PoolParty foundational topics with the *core topics* identified by the seminal papers (cf. *Table* 3), as said topics represent eighteen high level foundational topics from the Semantic Web community identified by analyzing existing taxonomies and well known Semantic Web research sub-communities. One notable omission is *distribution, decentralization, and federation*, which is an emerging topic in the Semantic Web community. Comparing the PoolParty output to the marginal topics presented in *Table* 4, we again observe good coverage with the exception of the topics relating to *multilingual intelligent agents* and *change management and propagation*. However, several topics that didn't figure in the seminal papers were uncovered by the PoolParty analysis, such as *recommendations, use cases, case studies, open data, information systems, web data, semantic technology*, and *structured data* (cf. *Table* 5). Given that these topics are very general in nature they could not be easily mapped to the primary topics appearing in the seminal papers.

Rexplore Analysis: As shown in *Table* 3, Rexplore extracted topics that can be mapped to the majority of *core topics* with the exception of two topics: *semantic web databases* and *distribution, decentralization, federation*. Like PoolParty, when it came to the *marginal topics* (cf. *Table* 4), *multilingual intelligent agents* and *change management and propagation* were not among the core topics extracted by Rexplore, additionally evidence of research on *scalability, efficiency, robust semantic approaches* was not present in the top 40 topics. However, Rexplore also identified several application or use case oriented keywords that were not mentioned in the seminal papers (cf. *Table* 5), such as *computational linguistics, recommender systems, mobile devices, cloud computing, e-learning system, robotics, electronic commerce systems*, and *decision support systems*.

Saffron Analysis: The mapping of the seminal paper topics to the results of the Saffron topic analysis affirms that all core topics that appear in the seminal papers appear in the Semantic Web corpus. Interestingly, Saffron is the only approach that uncovered evidence of *distribution, decentralization and federation*, however in contrast *privacy, trust, security, and provenance* did not figure in the top 40 topics uncovered by Saffron. Although Saffron has good coverage of the core topics, it was less successful at identifying keywords that could be aligned to the *marginal topics* (cf. *Table* 4). Like PoolParty and Rexplore, topics relating to *multilingual intelligent agents* and *change management and propagation* were not present in the top 40, however nor were *scalability, efficiency, robust semantic approaches* and *semantic web services*. Like PoolParty, Saffron also identified general topics that could not be aligned with topics from the seminal papers, namely, *open data*, *web data*, *web technology*.

8.2. 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. [3]. According to our mapping presented in *Table* 6, evidence of 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 which indicate that research relating to the *Internet of Things, streaming and sensor data* is becoming sufficiently wide-spread to indicate its rise in importance within the Semantic Web community. However, at the same time, the other three topics that relate to *scale*, *intelligent software agents* and *quality* were only weakly identified by the seminal papers.

8.3. Evidence of Trends

In the following we summarize the analysis of the trends identified by PoolParty (cf. *Figure* 4- 5), Rexplore (cf. *Figure* 7- 8) and Saffron (cf. *Figure* 11). The foundational topic and trend analysis conducted via PoolParty did not yield any useful results, as gener-

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

ally 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. [17] 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.

8.4. Mixed Method Observations

The comparative analysis of the research topics identified with the qualitative and quantitative methods, discussed in the previous sections, reveals several

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	-	-

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. While, data-driven methods failed 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. These topics, could for example be especially encouraged in calls for papers of future conferences or via workshops or journal special issues.

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. These are obviously interesting topics for the community and work on them should be fos-

tered with community building efforts such as organizing dedicated workshops or benchmarking activities.

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. As per the other categories, work on these topics should be encouraged via workshops, call for papers, and special issues. However, 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).

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 corpora and 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 a-priory built taxonomies of research areas, while more expensive, it 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.

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 keywords required expert knowledge and therefore Saffron should mostly 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.

9. 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.

In this paper, 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. The main conclusions

of our work refer not only to the important topics for Semantic Web research but also refer to strengths and weaknesses of the various topic extraction and analysis methods.

An overarching Semantic Web Research Landscape. 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 year. 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 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.

Strengths and weaknesses of 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 surprisingly, 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 paper 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 mainly focused on two approaches to analyze and reflect about the past and to some extent the future development of our research community, using expert opinions and applying our own data-driven methods. However, we could go a

third way, adopting either (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.

Additional 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.

References

[1] Kartik Asooja, Georgeta Bordea, Gabriela Vulcu, and Paul Buitelaar. Forecasting emerging trends from scientific literature. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

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

[3] Abraham Bernstein, James Hendler, and Natalya Noy. A new look at the semantic web. *Communications of the ACM*, 59(9): 35–37, 2016.

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

[5] 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.

[6] Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, and Bianca O Pereira. Expertise mining for enterprise content management. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3495–3498, 2012.

[7] Georgeta Bordea, Kartik Asooja, Paul Buitelaar, and Leona O'Brien. Gaining insights into the global financial crisis using saffron. *NLP Unshared Task in PoliInformatics*, 2014.

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

[9] 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.

[10] 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.

[11] David Chavalarias and Jean-Philippe Cointet. Phylomemetic patterns in science evolutionâĂŤthe rise and fall of scientific fields. *PloS one*, 8(2):e54847, 2013.

[12] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology*, 57 (3):359–377, 2006.

[13] Xiang-Ying Dai, Qing-Cai Chen, Xiao-Long Wang, and Jun Xu. Online topic detection and tracking of financial news based on hierarchical clustering. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 6, pages 3341–3346. IEEE, 2010.

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

[15] Jörg Diederich, Wolf-Tilo Balke, and Uwe Thaden. Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 505–505. ACM, 2007.

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

[17] 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`.

[18] Volker Frehe, Vilius Rugaitis, and Frank Teuteberg. Scientometrics: How to perform a big data trend analysis with sciencemineer. In *GI-Jahrestagung*, 2014.

[19] Birte Glimm and Heiner Stuckenschmidt. 15 years of semantic web: An incomplete survey. *KI-Künstliche Intelligenz*, 30(2): 117–130, 2016.

[20] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[21] William Hood and Concepción Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001.

[22] Yingjie Hu, Krzysztof Janowicz, Grant McKenzie, Kunal Sengupta, and Pascal Hitzler. A linked-data-driven and semantically-enabled journal portal for scientometrics. In *International Semantic Web Conference*, pages 114–129.

Springer, 2013.

[23] Yookyung Jo, Carl Lagoze, and C Lee Giles. Detecting research topics via the correlation between graphs and texts. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 370–379. ACM, 2007.

[24] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

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

[26] 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, 2010.

[27] Satoshi Morinaga and Kenji Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 811–816. ACM, 2004.

[28] Mizuki Oka, Hirotake Abe, and Kazuhiko Kato. Extracting topics from weblogs through frequency segments. In *Proc. of the Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[29] Francesco Osborne and Enrico Motta. Klink-2: integrating multiple web sources to generate semantic topic networks. In *International Semantic Web Conference*, pages 408–424. Springer, 2015.

[30] Francesco Osborne, Enrico Motta, and Paul Mulholland. Exploring scholarly data with rexplore. In *International semantic web conference*, pages 460–477. Springer, 2013.

[31] Francesco Osborne, Angelo Salatino, Aliaksandr Birukou, and Enrico Motta. Automatic classification of springer nature proceedings with smart topic miner. In *International Semantic Web Conference*, pages 383–399. Springer, 2016.

[32] Francesco Osborne, Patricia Lago, Muccini Henry, and Motta Enrico. Reducing the effort for systematic reviews in software engineering. In preparation, 2018.

[33] Sergey Parinov and Mikhail Kogalovsky. Semantic linkages in research information systems as a new data source for scientometric studies. *Scientometrics*, 98(2):927–943, 2014.

[34] 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.

[35] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

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