Semantic Web: Who is who in the field – a bibliometric analysis

Ying Ding

School of Library and Information Science, Indiana University Bloomington, IN, USA

Abstract.

The Semantic Web (SW) is one of the main efforts aiming to enhance human and machine interaction by representing data in an understandable way for machines to mediate data and services. It is a fast-moving and multidisciplinary field. This study conducts a thorough bibliometric analysis of the field by collecting data from Web of Science (WOS) and Scopus for the period of 1960–2009. It utilizes a total of 44,157 papers with 651,673 citations from Scopus, and 22,951 papers with 571,911 citations from WOS. Based on these papers and citations, it evaluates the research performance of the SW by identifying the most productive players, major scholarly communication media, highly cited authors, influential papers and emerging stars.

Keywords: citation analysis; impact analysis; research evaluation; semantic web

1. Introduction

The web is experiencing tremendous changes in its function to connect information, people and knowledge, but also facing severe challenges to integrate data and facilitate knowledge discovery. The Semantic Web (SW) is one of the main efforts aiming to enhance human and machine interaction by representing data in an understandable way for machines to mediate data and services [1]. Recently, PriceWaterhouseCoopers [2] predicted that SW technologies may revolutionize the entire enterprise of decision-making and information sharing. The profile of the SW has been further heightened by the Obama administration’s groundbreaking plan to initiate SW technologies to bring transparency to government activities [3]. Indeed, we see and hear the term ‘Semantic Web’ almost everywhere we go.

Why is the SW becoming so popular? One obvious reason is the increasing needs of individuals and society to process information with efficiency, speed and comprehensiveness. This primary need addresses the vexing issue of the web’s over-flooded information. Ten years ago the web largely contained documents, allowing users to consume these documents and mine their nuggets in a
timely and relatively straightforward fashion. But the explosively massive increase of websites has generated an information deluge, creating an often confusing and overwhelming format for gathering pertinent data within a reasonable period of time. For example, within the first few months of 2009 there was an increase of 46 million websites [4]. The human capacity to read or consume this level of data is not possible to achieve in his lifetime. There is now a serious demand to distil documents into data that extract core concepts and represent them in a concise manner, such as RDF triples. A huge amount of documents can therefore be shrunk to several data triples, which allows for easier consumption and retrieval. Yet even these improved cycles can go infinitively. As information deluge turns into data deluge, there is a need to add metadata to data, a process already in place. In turn metadata deluge will become another deluge, with no end in sight to this information abstraction. So while these abstracting processes can reduce the size of data and the burden for human use, they also create new challenges for data sharing and integration.

Another major issue for web users is the problem of data sharing and integration [5]. If data are isolated somewhere as a silo, usage and function can be significantly limited. As the world is becoming increasingly linked [6], the proper sharing of data has become essential to virtually all fields. Since data are represented in widely different syntax and semantics, the tasks of integrating data may be profoundly complex. One of the major missions for the database community is thus to find efficient ways to integrate data. But this remains as a remote goal where no ‘shortcuts’ are available. Data stored in databases are structured data, while most data on the web are unstructured [7]. Integrating and sharing data becomes more challenging, as what is called the current ‘bag-of-strings’ nature of the web does not facilitate connections that are machine readable. These problems need to be addressed and solved. But there is no clear answer. Semantic Web proposes technologies and methods that mainly address these two needs: how to add semantics to data and how to enable data integration [8].

Ontology, the backbone of the SW, is the formal representation of domain schemas. An ontology provides a shared vocabulary by modelling the semantics of data and representing them in mark-up languages proposed by the World Wide Web Consortium (W3C). W3C plays a major role in directing international efforts at specifying, developing and deploying standards for sharing information [9]. Semantically enriched data pave the crucial way to facilitate web functionality and interoperability. Semantic Web technologies thus open up new possibilities for developing applications that work across the web by modelling and linking data with best practices. They provide a fundamental infrastructure to create, represent and instantiate ontologies and metadata so as to enable intelligent retrieval and discovery. Semantic Web technologies continue to influence data sharing and management in various fields, such as digital library, knowledge management, data mining, social media, electronic commerce and web services [10].

Although SW is derived from the arena of artificial intelligence (AI), where ontological research can be traced back to the early 1980s, the groundbreaking progress in this area started from the late 1990s or the year 2000, when significant funding was secured from the European Union (EU) and the USA to support these important innovations [5]. This paper uses citations and publications to illustrate this 10-year development in the SW field, with the special focus on semantics and ontology-related research development. Following this Introduction, Section 2 gives a brief history; Section 3 provides the related work; Section 4 presents the research methods; Section 5 discusses the productivity and impact of this field, and Section 6 summarizes the results and addresses future research.

2. Brief history

Ten years ago nearly all the SW researchers could fit into one meeting room. They had to attend various conferences to explain the difference between ‘ontology’ and ‘oncology’, for the infrastructure and enabling methods/tools for SW were unclear. Researchers still struggled with the migration of existing AI methods to the web, and tried to avoid yet another ‘AI winter’ [11]. The brief history described here focuses on several ‘firsts’ in the field: the first language, first conference, first journal, and first foundation.
In Europe, we claimed that the first EU-funded SW project was OntoKnowledge (www.ontoknowledge.org/, 2000–2002) led by the Free University of Amsterdam. The major output of this project was the development of the OIL language which was not scheduled as a formal deliverable from the proposal. This project triggered the first meeting of EU and US researchers in Aachen, Germany, in August 2000. This meeting stressed the importance of the layered structure of OIL, and planned future EU funding for the community-formed Thematic Network for the Semantic Web (called the OntoWeb project, funded two years later by the EU). One month later, the second DAML and OIL meeting was held in Amsterdam. Three months later in December 2000, the DARPA Agent Markup Language Program officially announced that DAML+OIL was expected to be available that month, and in January 2001 its official version was released. DAML+OIL was later developed as OWL, which is currently the W3C standard and one of the key languages in the SW area.

The fundamental community-forming effort for the SW came from the OntoWeb project funded by the EU between 2002 and 2004. The project created several ‘firsts’ – the first conference, largely sponsored by the OntoWeb consortium, was held in Stanford in 2001, named the Semantic Web Working Symposium. The conference was subsequently renamed as the International Semantic Web Conference and has been held annually thereafter in Europe, Asia and America in alternating years. Following the same pattern, the regional conferences were created. The first European Semantic Web Symposium was held in Greece in 2004 and later became known as the European Semantic Web Conference. The first Asian Semantic Web Conference was held in Beijing in 2006. Of course, nowadays, SW-related topics are mentioned in almost all the major computer science-related conferences and broadly spread to conferences in other domains, such as, biology, chemistry, life science, medicine, library science and so on.

Creating an international journal for the field was planned in the deliverable of the OntoWeb project but was first discussed at the Dagstuhl Workshop on the Semantic Web in March 2000. The initial plan was to start the journal under the rubric of the Electronic Transactions on Artificial Intelligence (ETAI), which was published under the scientific patronage of the Royal Swedish Academy of Sciences and the European Coordinating Committee for Artificial Intelligence (ECCAI). This journal in the end found a home with Elsevier in 2003 (Journal of Web Semantics: Science, Services and Agents on the World Wide Web). The journal grew with the community and received a 2009 impact factor of 3.023. It is currently ranked as the 12th highest journal of 94 in the categories of computer science and artificial intelligence.

The first non-profit foundation was sponsored by the OntoWeb project and established in Amsterdam as ‘Stichting OntoWeb’ (Stichting is the Dutch translation for ‘foundation’) in 2001. The Foundation’s objective is the advancement of research and development in the field of ontology and SW in general, and information exchange for knowledge management and electronic commerce in particular. Later on, this Stichting was moved to Karlsruhe and renamed the Semantic Web Science Association (SWSA). Now it supervises the organization of the International Semantic Web Conference series and other related conferences, workshops and summer schools and runs the Journal of Web Semantics.

At this 10-year juncture of the SW, it is now important to identify its current status, including who the major players are, such as the most productive and highly cited authors, and the new driving forces. Since this area is moving fast and leading innovations on web engineering, data integration and service architecture, there is a pressing need to conduct research performance evaluation. This paper uses works published in this field to portray its research landscape.

3. Related work

Although various critical problems exist in bibliometric analysis as a method to evaluate research impact, such as database-related problems, inflated citation records, bias in citation rates and crediting of multi-author papers [12], it has been extensively applied over the past decades [13]. The basic approach is straightforward counting, such as how many times a particular paper has been cited [14]. Advanced techniques have been developed as well, such as author co-citation analysis [15], the $h$-index [16, 17], social network analysis [18, 19] and PageRank [20].
Recently, for example, Huang [21] collected publications associated with research on obstructive sleep apnoea (OSA) between 1991 and 2006 from the Web Of Science (WOS) to identify and predict the trends of publication output, journal patterns, country of publication, and authorship. Sorensen [17] applied citation analysis to post-1984 research on Alzheimer’s disease based on data from PubMed and WOS. Rikkonen and Vihinen [14] examined the productivity and impact of more than 700 biomedical researchers in Finland from 1966 to 2000. Thijs and Glanzel [22] used different bibliometric indicators to profile European research institutes.

But available research on using bibliometric methods to evaluate the field of SW is scarce, partially because it is still a young emerging field. Mika [23] and Mika, Elfring and Groenewegen [24] conducted social network analysis for the SW research community based on researchers who have submitted publications or held an organizing role at the first, second and third International Semantic Web Conference (ISWC2002, ISWC2003 and ISWC2004) or the first Semantic Web Working Symposium in 2001. Their dataset contains 608 researchers. They compared the in-degree and closeness centrality, structural holes, publications, and citations among these researchers and identified the core community and influential members. Zhao and Strotmann [25] used author co-citation analysis to detect school-of-thoughts for the XML field, which is broader than the SW field. As there is not a thorough citation analysis for SW research, this paper fills this gap by analysing papers and citations produced in this field.

4. Method

For citation analysis, WOS and Scopus are the two major authorized databases [26]. But since 2007, WOS has excluded all the major computer science conference proceedings and put them to the ISI proceedings which are not part of WOS anymore. Because SW is an emerging multidisciplinary field, we place our focus especially on the semantics and ontology-related research (as discussed in the Introduction), which form the core part of the SW field. In April 2009, ‘Semantic*’ or ‘Ontolog*’ were used as the search terms to retrieve related publications and their citations from titles, keywords and abstracts of papers in WOS and Scopus, with the restriction to the computer science-related areas, including library and information science. The search query in WOS was TS = (semantic* OR ontolog*) refined by subject areas related to computer science including theory and methods, artificial intelligence, information systems, software engineering, interdisciplinary, hardware and architecture, information science and library science, and cybernetics. In total, 23,670 items were identified. After excluding editorial materials, meeting abstracts and others, 22,951 articles remained. For Scopus, the search query was title-abs-key (semantic*) or title-abs-key(ontolog*) refined by subject areas in computer science, library and information science, and other related disciplines, which resulted in 46,029 items. After excluding corrections, conference review and other notes, 44,157 articles remained.

The main hypothesis for forming the search query for WOS and Scopus is that if this paper belongs to the SW area, the authors should mention either ‘ontolog*’ or ‘semantic*’ in their title, keyword or abstract. The reason why the semantic language terms are not included in the search query is that there are too many of them and they are still evolving, such as, XML, RDF, RDF-S, X-Query, SPARQL, RDFa, OWL, OIL, DAML + OIL, DAML, OWL-S, WSMO, WSML, GRIDDLE, SWRL, RIF, to name but a few. Also the OWL, OIL and DAML can lead to a large amount of noisy data, such as papers researching on OWL as an animal, or OIL as a product of oil industry. For example, in Ian Horrocks’ most cited paper on OWL, the SW and ontology are mentioned in the title and abstract. So if one paper never mentions ‘ontolog*’ or ‘semantic*’ in the title, abstract, or keywords, there is a high chance that this paper might not be directly related to the SW. So ‘ontolog*’ or ‘semantic*’ can be used as search terms to capture the majority of papers published in the SW area.

In total, there were 44,157 papers with 651,673 citations from Scopus covering 1975–2009, and 22,951 papers with 571,911 citations from WOS covering 1960–2009. We took these two datasets to analyse the research performance of the SW community. The SW is a continuous development of the World Wide Web. The major progress of this field started in early 2000 when it gradually acquired major funding from EU and USA. In order to discuss this important phase in detail, we divided the period of 2000–09 into 2000–04 and 2005–09 to better outline the dynamic changes.
5. Results

There have been consistent increases in SW publications in Scopus to date, taking into consideration that 2009 data were downloaded in April 2009 (see Figure 1). In Scopus, the 2008 publications nearly doubled the amount of total paper published during 1990–99. Since 2000, there has been an average yearly increase in publications of 31.7%. In WOS, however, these numbers significantly dropped in 2007 and 2008, due to the exclusion of conference proceedings from 2007 onwards, especially those coming from major SW events such as the International Semantic Web Conference, European Semantic Web Conference, Asian Semantic Web Conference and the World Wide Web Conference. Among the total number of SW papers in WOS and Scopus, 50% are conference papers.

5.1. Productivity

5.1.1. Journal/conference

Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence are the two major publishing channels for SW papers, all of which are conference papers (see Figure 1). This confirms that conference proceedings form the dominant publishing media reporting SW research. The top journals contributing to the publishing of SW papers are Theoretical Computer Science, Bioinformatics, Data and Knowledge Engineering, IEEE Transactions on Knowledge and Data Engineering, Information and Computation and Artificial Intelligence. Most of the journals are in English, with one journal in Chinese, the Ruan Jian Xue Bao/Journal of Software.

5.1.2. Researchers

In WOS, the number of publications produced by the authors have been counted and ranked based on the first, second, and third author respectively (see Table 2). We present them in three time periods: 1960–2009, 2000–04 and 2005–09. For example, the first authors T. Eiter (Vienna Technical University), A. Brogi (University of Pisa), and H. Zhuge (Chinese Academy of Sciences) are the top three most productive researchers between 1960 and 2009. If we look at the recent period (2000–09), H. Zhuge and T. Eiter keep their high productivity, while J.J. Jung (Yeungnam University, Korea) emerges as a new star with 13 publications in 2005–09, as does J.J. Alferes (University of Nova Lisboa) with 12 publications in 2000–04 as first author.

Scopus contains all the excluded conference proceedings of WOS, and therefore has better coverage of the field (see Table 3). Within the total period, H. Zhuge, T. Eiter and J.J. Jung are the top three
productive first authors. H. Zhuge maintains high productivity in 2000–04 and 2005–09, while J.J. Jung moves to the top position in 2005–09 with 25 publications as first author, and T. Eiter keeps his third position in 2005–09. E. Bertino (Purdue University) and M.R. Naphade (University of Illinois) are ranked as the second and third most productive first authors in 2000–04.

5.2. Impact

5.2.1. Highly cited journals/conferences
In computer science, the major scholarly communication channel is shifting from journals to conferences (http://isiwebofknowledge.com/media/pdf/cpci_faq.pdf). Statistics on the SW, as one of the fast-moving subfields, show that the major highly cited channels are various conference proceedings published as Lecture Notes in Computer Science or Lecture Notes in Artificial Intelligence. In WOS (see Table 4), Artificial Intelligence, Communication of the ACM and Theoretical Computer Science journals are ranked the top three or four during these three periods. Looking at the top 20
| Rank | First author | Second author | Third author |
|------|--------------|---------------|--------------|
| 1    | Eiter, T., 43 | Montanari, U., 32 | Eiter, T. |
| 2    | Barbuti, R., 20 | Horrocks, I., 5 | Montanari, U. |
| 3    | Lee, J., 19 | Staab, S., 5 | Zhang, Y., 8 |
| 4    | Greco, S., 19 | Lamma, E., 6 | Decker, S., 6 |
| 5    | DEBAKKER, J.W., 19 | Antoniou, G., 8 | Motta, E., 7 |
| 6    | Alferes, J.J., 18 | Antoniou, G., 8 | Li, L., 5 |
| 7    | Vogler, W., 18 | Antoniou, G., 8 | Lei, Z., 7 |
| 8    | Zhang, Y., 18 | Antoniou, G., 8 | Decker, S., 6 |
| 9    | Giacobazzi, R., 17 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 10   | Dabulis, D., 17 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 11   | Borger, E., 16 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 12   | Corradini, A., 16 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 13   | Aceto, L., 16 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 14   | Meseguer, J., 16 | Antoniou, G., 8 | de Boer, F.S., 16 |
| 15   | Boreale, M., 16 | Antoniou, G., 8 | de Boer, F.S., 16 |

Note: the numbers in this table represent the number of publications. Some of the current Chinese names, such as Liu, L, Ding, L, could be different people but it is beyond the scope of current research to differentiate author identities.
| Rank | First author | Second author | Third author |
|------|--------------|---------------|--------------|
| 1    | Di Sciascio E. | Motta E. | Baldan P. |
| 2    | Zhuge H. | Yuan L. | Naphade M. |
| 3    | Tadeusiewicz R. | Hussain F.K. | Fensel D. |
| 4    | Yang Y. | Yang Y. | Yang Y. |
| 5    | Jeong D. | Jeong D. | Jeong D. |
| 6    | Lee C.S. | Lee C.S. | Lee C.S. |
| 7    | Dong M. | Dong M. | Dong M. |
| 8    | Zhao H. | Zhao H. | Zhao H. |
| 9    | Zhang D. | Zhang D. | Zhang D. |
| 10   | Ma Z.M. | Ma Z.M. | Ma Z.M. |
| 11   | Lu L. | Lu L. | Lu L. |
| 12   | Ding Ying | Ding Ying | Ding Ying |
| 13   | Pan J.Z. | Pan J.Z. | Pan J.Z. |
| 14   | Embley D.W. | Embley D.W. | Embley D.W. |
| 15   | Mylopoulos J. | Mylopoulos J. | Mylopoulos J. |
| 16   | Bertino E. | Bertino E. | Bertino E. |
| 17   | Tao A.M. | Tao A.M. | Tao A.M. |
| 18   | Palopoli L. | Palopoli L. | Palopoli L. |
| 19   | Pimentel L. | Pimentel L. | Pimentel L. |
| 20   | Ma X. | Ma X. | Ma X. |
| 21   | Andrews R. | Andrews R. | Andrews R. |
| 22   | Llorente D. | Llorente D. | Llorente D. |
| 23   | Osiadziszewski E. | Osiadziszewski E. | Osiadziszewski E. |
| 24   | Goguen J. | Goguen J. | Goguen J. |
| 25   | De Caro L. | De Caro L. | De Caro L. |
| 26   | Gromov A. | Gromov A. | Gromov A. |
| 27   | Zeng Y. | Zeng Y. | Zeng Y. |
| 28   | Zhang R. | Zhang R. | Zhang R. |
| 29   | Zhang D. | Zhang D. | Zhang D. |
| 30   | Shi Y. | Shi Y. | Shi Y. |
| 31   | Shi Z. | Shi Z. | Shi Z. |
| 32   | Li X. | Li X. | Li X. |
| 33   | De Bruijn J. | De Bruijn J. | De Bruijn J. |
| 34   | Liu B. | Liu B. | Liu B. |
| 35   | Liu Z. | Liu Z. | Liu Z. |
| 36   | Liu S. | Liu S. | Liu S. |
| 37   | Liu M. | Liu M. | Liu M. |
| 38   | Liu H. | Liu H. | Liu H. |
| 39   | Liu J. | Liu J. | Liu J. |
| 40   | Liu D. | Liu D. | Liu D. |
| 41   | Liu W. | Liu W. | Liu W. |
| 42   | Liu G. | Liu G. | Liu G. |
| 43   | Liu K. | Liu K. | Liu K. |
| 44   | Liu Y. | Liu Y. | Liu Y. |
| 45   | Liu X. | Liu X. | Liu X. |
| 46   | Liu J. | Liu J. | Liu J. |
| 47   | Liu J. | Liu J. | Liu J. |
| 48   | Liu J. | Liu J. | Liu J. |
| 49   | Liu J. | Liu J. | Liu J. |
| 50   | Liu J. | Liu J. | Liu J. |
| 51   | Liu J. | Liu J. | Liu J. |
| 52   | Liu J. | Liu J. | Liu J. |
| 53   | Liu J. | Liu J. | Liu J. |
| 54   | Liu J. | Liu J. | Liu J. |
| 55   | Liu J. | Liu J. | Liu J. |
| 56   | Liu J. | Liu J. | Liu J. |
| 57   | Liu J. | Liu J. | Liu J. |
| 58   | Liu J. | Liu J. | Liu J. |
| 59   | Liu J. | Liu J. | Liu J. |
| 60   | Liu J. | Liu J. | Liu J. |
| 61   | Liu J. | Liu J. | Liu J. |
| 62   | Liu J. | Liu J. | Liu J. |
| 63   | Liu J. | Liu J. | Liu J. |
| 64   | Liu J. | Liu J. | Liu J. |
| 65   | Liu J. | Liu J. | Liu J. |
| 66   | Liu J. | Liu J. | Liu J. |
| 67   | Liu J. | Liu J. | Liu J. |
| 68   | Liu J. | Liu J. | Liu J. |
| 69   | Liu J. | Liu J. | Liu J. |
| 70   | Liu J. | Liu J. | Liu J. |
| 71   | Liu J. | Liu J. | Liu J. |
| 72   | Liu J. | Liu J. | Liu J. |
| 73   | Liu J. | Liu J. | Liu J. |
| 74   | Liu J. | Liu J. | Liu J. |
| 75   | Liu J. | Liu J. | Liu J. |
| 76   | Liu J. | Liu J. | Liu J. |
| 77   | Liu J. | Liu J. | Liu J. |
| 78   | Liu J. | Liu J. | Liu J. |
| 79   | Liu J. | Liu J. | Liu J. |
| 80   | Liu J. | Liu J. | Liu J. |
| 81   | Liu J. | Liu J. | Liu J. |
| 82   | Liu J. | Liu J. | Liu J. |
| 83   | Liu J. | Liu J. | Liu J. |
| 84   | Liu J. | Liu J. | Liu J. |
| 85   | Liu J. | Liu J. | Liu J. |
| 86   | Liu J. | Liu J. | Liu J. |
| 87   | Liu J. | Liu J. | Liu J. |
| 88   | Liu J. | Liu J. | Liu J. |
| 89   | Liu J. | Liu J. | Liu J. |
| 90   | Liu J. | Liu J. | Liu J. |
| 91   | Liu J. | Liu J. | Liu J. |
| 92   | Liu J. | Liu J. | Liu J. |
| 93   | Liu J. | Liu J. | Liu J. |
| 94   | Liu J. | Liu J. | Liu J. |
| 95   | Liu J. | Liu J. | Liu J. |
| 96   | Liu J. | Liu J. | Liu J. |
| 97   | Liu J. | Liu J. | Liu J. |
| 98   | Liu J. | Liu J. | Liu J. |
| 99   | Liu J. | Liu J. | Liu J. |
| 100  | Liu J. | Liu J. | Liu J. |

Note: Some popular Asian names are deleted as many researchers can have the same names. The numbers in this table represent the number of publications. Some of the current Chinese names such as Liu, L, Ding, L could be different people but it is beyond the scope of current research to differentiate author identities.
highly cited journal/conferences, one finds that SW is closely related to AI, computing theory, logic programming, database and bioinformatics.

Table 5 shows the top 20 highly cited journals or conferences from Scopus. There is no major difference between Tables 4 and 5, where between them Lecture Notes in Computer Science, Communications of the ACM, Artificial Intelligence and Theoretical Computer Science are ranked within the top three during these three periods. Nature and Science emerge within the top 20 in 2005–09.

5.2.2. Highly cited authors
The number of times authors or their work are cited can be used to measure the impact of their work on the community. Table 6 shows the top 20 highly cited authors based on 571,911 citations from WOS. In the whole period (1960–2009), R. Milner is ranked as the top author for his contribution of pi-calculus for mobile processes, M. Gelfond is second for his work on logic programming and non-monotonic reasoning, and C.A.R. Hoare is third for his Quicksort algorithm and Hoare logic, which brought him the Turing Award in 1980. Sir Tim Berners-Lee, the inventor of the World Wide Web and SW, is ranked fourth in the entire period, second for 2000–04 and first for the period 2005–09, which shows his increasing impact within the community. T. Gruber’s ontology definition and his ontology engineering work are highly cited, causing him to be ranked as third in 2000–04. I. Horrocks’s fundamental contribution to SW languages, especially OWL, move him up to second place in 2005–09.

Citations in Scopus include all authors, making it possible to rank the cited authors based on first, second, and third author. In the total period (1960–2009), R. Milner, T. Berners-Lee and I. Horrocks are ranked as the top three highly cited first authors; J. Hendler, S. Staab and H. Garcia-Molina are ranked as the top three highly cited second authors; and O. Lassila, F. van Harmelen and I. Horrocks are the top three highly cited third authors. In 2000–04, R. Milner, T. Berners-Lee and M. Abadi are

Table 4
Highly cited journals/conferences (WOS)

|     | 1960–2009 |     | 2000–04 |     | 2005–09 |
|-----|-----------|-----|---------|-----|---------|
| R   | Journal/conference | No. cited | Journal/conference | No. cited | Journal/conference | No. cited |
| 1   | LECT NOTES COMPUT SC | 34,015 | LECT NOTES COMPUT SC | 10,604 | LECT NOTES COMPUT SC | 12,706 |
| 2   | ARTIF INTELL | 6915 | ARTIF INTELL | 2164 | ARTIF INTELL | 2007 |
| 3   | THEOR COMPUT SCI | 5691 | LECT NOTES ARTIF INT | 1996 | LECT NOTES ARTIF INT | 1947 |
| 4   | COMMUN ACM | 5669 | THEOR COMPUT SCI | 1856 | COMMUN ACM | 1816 |
| 5   | LECT NOTES ARTIF INT | 4494 | COMMUN ACM | 1714 | THEOR COMPUT SCI | 1799 |
| 6   | J LOGIC PROGRAM | 3238 | INFORM COMPUT | 1206 | BIOINFORMATICS | 1181 |
| 7   | INFORM COMPUT | 3216 | J LOGIC PROGRAM | 1119 | INFORM COMPUT | 1048 |
| 8   | IEEE T SOFTWARE ENG | 3090 | IEEE T SOFTWARE ENG | 857 | IEEE T PATTERN ANAL | 1027 |
| 9   | ACM T DATABASE SYST | 2891 | IEEE T KNOWL DATA EN | 800 | NUCLEIC ACIDS RES | 1000 |
| 10  | J ASSOC COMPUT MACH | 2710 | ACM T PROGR LANG SYS | 778 | IEEE T KNOWL DATA EN | 987 |
| 11  | ACM T PROGR LANG SYS | 2567 | J ASSOC COMPUT MACH | 723 | IEEE T SOFTWARE ENG | 863 |
| 12  | IEEE T KNOWL DATA EN | 2289 | ACM T DATABASE SYST | 722 | ACM T PROGR LANG SYS | 705 |
| 13  | IEEE T PATTERN ANAL | 1842 | IEEE INTELL SYST APP | 653 | IEEE INTELL SYST APP | 662 |
| 14  | ACTA INFORM | 1617 | IEEE T PATTERN ANAL | 569 | DATA KNOWL ENG | 643 |
| 15  | ACM COMPUT SURV | 1478 | P ACM SIGMOD INT C M | 564 | J LOGIC PROGRAM | 595 |
| 16  | J ACM | 1431 | J LOGIC COMPUT | 462 | ACM T DATABASE SYST | 572 |
| 17  | P ACM SIGMOD INT C M | 1414 | J AM SOC INFORM SCI | 443 | J ASSOC COMPUT MACH | 542 |
| 18  | BIOINFORMATICS | 1400 | ACTA INFORM | 440 | J AM MED INFORM ASSN | 516 |
| 19  | J COMPUT SYST SCI | 1398 | DATA KNOWL ENG | 436 | VLDB J | 512 |
| 20  | NUCLEIC ACIDS RES | 1392 | ACM COMPUT SURV | 436 | INT J HUM-COMPUT ST | 511 |
Table 5
Highly cited journal/conference (Scopus)

| 1960–2009 | 2000–04 | 2005–09 |
|------------|----------|----------|
| R          | Journal/conference | No. cited | Journal/conference | No. cited | Journal/conference | No. cited |
| 1          | Lecture Notes in Computer Science | 22,923 | Lecture Notes in Computer Science | 6721 | Lecture Notes in Computer Science | 15,176 |
| 2          | Communications of the ACM | 5913 | Theoretical Computer Science | 2511 | Communications of the ACM | 2983 |
| 3          | Theoretical Computer Science | 5564 | Communications of the ACM | 2228 | Artificial Intelligence | 2429 |
| 4          | Artificial Intelligence | 5069 | Artificial Intelligence | 2056 | IEEE Intelligent Systems | 2061 |
| 5          | IEEE Intelligent Systems | 2844 | Information and Computation | 1230 | Theoretical Computer Science | 2060 |
| 6          | Information and Computation | 2722 | IEEE Transactions on Software Engineering | 1018 | Bioinformatics | 1943 |
| 7          | Journal of the ACM | 2472 | Journal of the ACM | 995 | Lecture Notes in Artificial Intelligence | 1483 |
| 8          | Lecture Notes in Artificial Intelligence | 2371 | Journal of Logic Programming | 821 | Journal of the ACM | 1260 |
| 9          | Bioinformatics | 2203 | Lecture Notes in Artificial Intelligence | 800 | Computational Linguistics | 1188 |
| 10         | IEEE Computer | 1861 | IEEE Intelligent Systems | 771 | IEEE Transactions on Knowledge and Data Engineering | 1160 |
| 11         | IEEE Transactions on Software Engineering | 1842 | IEEE Computer | 677 | Information and Computation | 1142 |
| 12         | Computational Linguistics | 1679 | ACM Transactions on Programming Languages and Systems | 629 | Scientific American | 1113 |
| 13         | IEEE Transactions on Knowledge and Data Engineering | 1670 | ACM Computing Surveys | 574 | IEEE Computer | 1067 |
| 14         | ACM Computing Surveys | 1554 | Fuzzy Sets and Systems | 546 | IEEE Internet Computing | 988 |
| 15         | Scientific American | 1440 | IEEE Transactions on Knowledge and Data Engineering | 442 | IEEE Transactions on Software Engineering | 973 |
| 16         | Fuzzy Sets and Systems | 1389 | Electronic Notes in Theoretical Computer Science | 434 | Nucleic Acids Res | 908 |
| 17         | Data and Knowledge Engineering | 1353 | Science of Computer Programming | 432 | Data and Knowledge Engineering | 874 |
| 18         | IEEE Internet Computing | 1305 | Computational Linguistics | 432 | ACM Computing Surveys | 864 |
| 19         | ACM Transactions on Programming Languages and Systems | 1254 | ACM Transactions on Database Systems | 408 | Science | 858 |
| 20         | SIGMOD Record | 1212 | Acta Informatica | 390 | Nature | 833 |

the top three highly cited first authors; J. Hendler, V. Lifschitz and H. Prade are the top three highly cited second authors; and O. Lassila, F. van Harmelen and H. Prade are the top three highly cited third authors. In 2005–09, T. Berners-Lee, I. Horrocks and R. Milner are the top three highly cited
first authors; J. Hendler, S. Staab and I. Horrocks are the top three highly cited second authors; and O. Lassila, F. van Harmelen and A. Joshi are the top three highly cited third authors.

### 5.2.3. Highly cited papers

Table 6 shows the ranks of highly cited papers in three different periods from WOS. T. Gruber’s ontology paper has been consistently highly cited and ranked as the lead paper for all periods. M. Gelfond’s stable model semantics for logic programming is ranked as the second highly cited paper in 1960–2009 and third in 2000–04. A. van Gelder’s well-founded semantics for general logic programs is ranked as the third in 1969–2009 and second in 2000–04. T. Berners-Lee, J. Hendler and O. Lassila’s famous article about the vision of SW published in *Scientific American* is ranked as the second highly cited paper in 2005–09. M. Ashburner’s gene ontology article is ranked as the third highly cited paper in the same period. Through examining the highly cited papers in this field, one sees a clear shift from its beginning as being AI-dominated with a focus on knowledge representation, logic programming and theory proving, to more data-driven practical approaches designed to realize the SW vision by converting the current document web into a data web. During 2005–09, more papers from data mining, natural language processing and database are highly cited. Ontology forms the heart of the SW vision and approaches, and the community has accepted ontology definitions from T. Gruber. Ontology engineering is also moving from creating a theoretical foundation for ontology to the mapping of different ontologies. Ontology languages have slowly evolved from various logic languages derived from the core AI. Semantic web services emerged in 2005–09, mainly represented by OWL-S initiatives (e.g. that of J. Hendler and S. Mcilarith).

Table 9 shows the highly cited papers from Scopus. As per Table 8, T. Gruber’s ontology paper published in *Knowledge Acquisition* in 1993 is again ranked as the most highly cited paper during all three periods. T. Berners-Lee, J. Hendler and O. Lassila’s *Scientific American* article is ranked as the second most highly cited paper in 2005–09 while R. Reiter’s logic for default reasoning is ranked as the second most highly cited article in 2000–04. S. Deerwester’s latent semantic analysis from
Table 7
Highly cited first, second and third authors (Scopus)

|       | 1960–2009 |            | 2000–04 |            | 2005–09 |            |
|-------|-----------|------------|---------|------------|---------|------------|
|       | First author | Second author | Third author | First author | Second author | Third author |
| 1     | Milner R. | Hendler J. | Lassila O. | Milner R. | Hendler J. | Lassila O. |
|       | 2182      | 1937       | 1538     | 916        | 503      | 364        |
| 2     | Berners-Lee T. | Staub S. | Van Harmelen F. | Berners-Lee T. | Lifschitz V. | Van Harmelen F. |
|       | 2033      | 944        | 614      | 511        | 310      | 184        |
| 3     | Horrocks I. | Garcia-Molina H. | Horrocks I. | Abadi M. | Prade H. | Prade H. |
|       | 1376      | 807        | 437      | 438        | 292      | 181        |
| 4     | Salton G. | Lifschitz V. | Walker D. | Dubois D. | Cousot R. | Walker D. |
|       | 1161      | 781        | 433      | 399        | 267      | 164        |
| 5     | Gruber T.R. | Horrocks I. | Johnson R. | Abiteboul S. | Garcia-Molina H. | Johnson R. |
|       | 1121      | 741        | 393      | 388        | 262      | 154        |
| 6     | Fensel D. | Van Harmelen F. | Lenzerini M. | Fensel D. | Huang T.S. | Montanari U. |
|       | 1047      | 710        | 371      | 384        | 261      | 143        |
| 7     | Guarino N. | Huang T.S. | Joshi A. | Salton G. | Staab S. | Horrocks I. |
|       | 1022      | 659        | 359      | 379        | 239      | 137        |
| 8     | Baader F. | Cousot R. | Sheth A. | Zadeh L.A. | Meseguer J. | Harper R. |
|       | 1019      | 641        | 350      | 376        | 236      | 127        |
| 9     | Zadeh L.A. | Prade H. | Decker S. | Hoare C.A.R. | Montanari U. | Vianu V. |
|       | 1001      | 634        | 341      | 371        | 225      | 126        |
| 10    | Abadi M. | Dumais S.T. | Rahm E. | Abramsky S. | Pnueli A. | Lenzerini M. |
|       | 923       | 633        | 340      | 337        | 224      | 120        |
| 11    | Hoare C.A.R. | Finin T. | Montanari U. | Cardelli L. | Parrow J. | Decker S. |
|       | 899       | 524        | 338      | 336        | 193      | 116        |
| 12    | Dubois D. | Bernstein P.A. | Prade H. | Cousot P. | Cardelli L. | Siccu D. |
|       | 522       | 828        | 334      | 332        | 180      | 526        |
| 13    | Abramsky S. | Worring M. | Hendler J. | Guarino N. | Gorrieri R. | Rice J. |
|       | 828       | 517        | 333      | 330        | 176      | 107        |
| 14    | Cousot P. | Montanari U. | Staub S. | Harel D. | Lenzerini M. | Wu J. |
|       | 822       | 513        | 326      | 319        | 168      | 102        |
| 15    | Abiteboul S. | Sattler U. | Santini S. | Alur R. | Dumais S.T. | Ullman J.D. |
|       | 813       | 512        | 311      | 309        | 166      | 100        |

© The Author(s), DOI: 10.1177/0165551510365295
Table 7  
(Continued)

| R | First author | Second author | Third author | First author | Second author | Third author | First author | Second author | Third author |
|---|--------------|---------------|--------------|--------------|---------------|--------------|--------------|---------------|--------------|
| 16 | Noy N.F., 798 | Parsia B., 507 | Domingos P., 307 | Gelfond M., 301 | Fikes R., 162 | Widom J., 99 | Alur R., 441 | Musen M.A., 336 | Finin T., 211 |
| 17 | Gelfond M., 795 | Meseguer J., 507 | Fensel D., 294 | Gruber T.R., 256 | Fensel D., 160 | Jacobson I., 98 | Calvanese D., 406 | Kawamura T., 320 | Fensel D., 206 |
| 18 | Alur R., 784 | Lenzerini M., 506 | Vianu V., 290 | Meseguer J., 249 | Grumberg O., 159 | Steele G., 97 | Abadi M., 406 | Paolucci M., 301 | Volz R., 202 |
| 19 | Harel D., 773 | Patel-Schneider P.F., 505 | Harper R., 288 | Horrocks I., 240 | Helm R., 156 | Booch G., 90 | Gelfond M., 405 | Cousot R., 292 | Sattler U., 195 |
| 20 | Eiter T., 759 | Pnueli A., 485 | Finin T., 280 | Rui Y., 237 | Bernstein P.A., 146 | Eker S., 88 | Foster I., 388 | Lenzerini M., 280 | Boley H., 194 |
| No. | Paper                                                                 | 1960–2009 | 2000–04 | 2005–09 |
|-----|-----------------------------------------------------------------------|-----------|---------|---------|
| 1   | Gruber TR (1993), A translation approach to portable ontology specifications, *Knowl Acquis*, 5, 199 | 513       | 175     | 313     |
| 2   | Gelfond M (1988), The stable model semantics for logic programming, *Int J Log Progr*, 1070 | 393       | 108     | 238     |
| 3   | Vangelder A (1991), The well-founded semantics for general logic programs, *J Assoc Comput Mach*, 38, 620 | 311       | 106     |         |
| 4   | Deerwester S (1990), Indexing by latent semantic analysis, *J Am Soc Inform Sci*, 41, 391 | 265       | 100     |         |
| 5   | Bernerslee T (2001), The semantic web – a new form of web content that is meaningful to computers will unleash a revolution of new possibilities, *Sci Am*, 284, 34 | 264       | 95      |         |
| 6   | Reiter R (1980), A logic for default reasoning, *Artif Intell*, 13, 81 | 255       | 94      |         |
| 7   | Milner R (1992), A calculus of mobile processes, *Inform Comput*, 100, 1 | 245       | 89      |         |
| 8   | Gelfond M (1991), Classical negation in logic programs and disjunctive databases, *New Generat Comput*, 9, 365 | 240       | 80      |         |
| 9   | Chen PPS (1976), The entity-relational model – toward a unified view of data, *ACM T Database Syst*, 1, 9 | 239       | 80      |         |
| 10  | Vanemenden MH (1976), Semantics of predicate logic as a programming language, *J Assoc Comput Mach*, 23, 733 | 228       | 71      |         |
Table 8
(Continued)

| Year      | Paper                                                                 | No. cited | Year      | Paper                                                                 | No. cited |
|-----------|------------------------------------------------------------------------|-----------|-----------|------------------------------------------------------------------------|-----------|
| 1960–2009 | 11  Harel D (1987), Statecharts – a visual formalism for complex-systems, Sci Comput Program, 8, 231 | 210       | 2000–04   | 12  Ashburner M (2000), Gene ontology: tool for the unification of biology, Nat Genet, 25, 25 | 191       |
|           |                                                                                                                                   |           |           | 13  Gruber TR (1995), Toward principles for the design of ontologies used for knowledge sharing, Int J Hum-Comput St, 43, 907 | 191       |
|           | 14  Miller GA (1995), WORDNET – a lexical database for English, Commun ACM, 38, 39                                              | 187       | 2005–09   | 15  Uschold M (1996), Ontologies: principles, methods and applications, Knowl Eng Rev, 11, 93 | 171       |
|           |                                                                                                                                   |           |           | 16  Zadeh LA (1965), Fuzzy sets, Inform Contr, 8, 338                    | 155       |
|           | 17  Girard JY (1987), Linear logic, Theor Comput Sci, 50, 1                                                                    |           |           | 18  McIlraith SA (2001), Semantic web services, IEEE Intell Syst App, 16, 46 | 148       |
|           |                                                                                                                                   |           |           | 19  McIlraith SA (2001), Semantic web services, IEEE Intell Syst App, 16, 46 | 142       |
|           | 20  Smeulders AWM (2000), Content-based image retrieval at the end of the early years, IEEE T Pattern Anal, 22, 1349             | 140       |           | 21  Smeulders AWM (2000), Content-based image retrieval at the end of the early years, IEEE T Pattern Anal, 22, 1349 | 140       |
|           |                                                                                                                                   |           |           | 22  Miller GA (1990), Introduction to WORDNET: an on-line lexical database, Int J Lexicogr, 3, 235 | 49        |
|           | 23  Reiter R (1980), A logic for default reasoning, Artif Intell, 13, 81                                                        |           |           | 24  Gelfond M (1988), The stable model semantics for logic programming, Int C Log Progr, 1070 | 70        |
|           |                                                                                                                                   | 70        |           | 25  Horrocks I (2003), From SHIQ and RDF to OWL: the making of a web ontology language, J Web Semant, 1, 7 | 67        |
|           | 26  Ashburner M (2000), Gene ontology: tool for the unification of biology, Nat Genet, 25, 25                                   | 67        |           | 27  Vangelder A (1991), The well-founded semantics for general logic programs, J Assoc Comput Mach, 38, 620 | 66        |
|           | 28  Girard JY (1987), Linear logic, Theor Comput Sci, 50, 1                                                                    | 51        |           | 29  Porter M.F. (1980), An algorithm for suffix stripping, Program, 14, 130-137 | 55        |
|           |                                                                                                                                   | 55        |           | 30  Kalfoglou Y (2003), Ontology mapping: the state of the art, Knowl Eng Rev, 18, 1 | 58        |
|           | 31  Hoare CAR (1969), An axiomatic basis for computer programming, Commun ACM, 12, 576                                          | 53        |           | 32  Wiederhold G (1992), Mediators in the architecture of future information-systems, IEEE Comput, 25, 38 | 53        |
|           | 33  McCarthy J (1980), Circumscription – a form of non-monotonic reasoning, Artif Intell, 13, 27                               | 50        |           | 34  Landauer TK (1998), An introduction to latent semantic analysis, Discourse Process, 25, 259 | 50        |
|           | 35  McIlraith SA (2001), Semantic web services, IEEE Intell Syst App, 16, 46                                                   |           |           | 36  Sebastiani F (2002), Machine learning in automated text categorization, ACM Comput Surv, 34, 1 | 51        |
|           | 37  Hendlr J (2001), Agents and the semantic web, IEEE Intell Syst App, 16, 46                                                 |           |           | 38  McIlraith SA (2001), Semantic web services, IEEE Intell Syst App, 16, 46 | 50        |
|           | 39  Miller GA (1990), Introduction to WORDNET: an on-line lexical database, Int J Lexicogr, 3, 235                            |           |           | 40  Miller GA (1990), Introduction to WORDNET: an on-line lexical database, Int J Lexicogr, 3, 235 | 49        |
|           | 41  Smeulders AWM (2000), Content-based image retrieval at the end of the early years, IEEE T Pattern Anal, 22, 1349         |           |           | 42  Dempster AP (1977), Maximum likelihood from incomplete data via EM algorithm, J Roy Stat Soc B, 39, 1 | 49        |
Journal of the American Society for Information Science is ranked as the third most highly cited paper in 1960–2009 and 2005–09. There is no major difference between Tables 8 and 9, even though WOS and Scopus have a significant different number of SW articles.

The most highly cited papers from WOS and Scopus can be grouped into different schools of thought:

- **Vision**: T. Berners-Lee’s ‘The semantic web’.

- **Ontology engineering**: T. Gruber’s ‘A translation approach to portable ontology specifications’; T. Gruber’s ‘Toward principles for the design of ontologies used for knowledge sharing’; M. Uschold’s ‘Ontologies: principles, methods and applications’; Y. Kalfoglou’s ‘Ontology mapping: the state of the art’ and R. Studer’s ‘Knowledge engineering: Principles and methods’.

- **Ontological languages**: I. Horrocks’s ‘From SHIQ and RDF to OWL’.

- **Semantic Web services**: S. Mcilraith’s ‘Semantic Web services’; J. Hendler’s ‘Agents and the Semantic Web’ and H. Zhuge’s ‘China’s E-science knowledge grid environment’.

- **Core AI**: M. Gelfond’s ‘The stable model semantics for logic programming’; A. van Gelder’s ‘The well-founded semantics for general logic programs’; R. Reiter’s ‘A logic for default reasoning’; R. Milner’s ‘A calculus of mobile processes’; M. Gelfond’s ‘Classical negation in logic programs and disjunctive databases’; M. Van Emde Boas’s ‘Semantics of predicate logic as a programming language’; K. Clark’s ‘Negation as failure’; L. Zadeh’s ‘Fuzzy sets’; J. Girard’s ‘Linear logic’; C. Hoare’s ‘An axiomatic basis for computer programming’; J. McCarthy’s ‘Circumscription – a form of non-monotonic reasoning’; R. Kifer’s ‘Logical foundations of object-oriented and frame-based languages’; M. Gelfond’s ‘Classical negation in logic programs and disjunctive databases’; M. Van Emde Boas’s ‘Semantics of predicate logic as a programming language’; R. Alur’s ‘A theory of timed automata’; S. Kraus’s ‘Non-monotonic reasoning, preferential models and cumulative logic’; J. Meseguer’s ‘Conditional rewriting logic as a unified model of concurrency’; B. Jacobs’s ‘A tutorial on (co)algebras and (co)induction’; P. Cohen’s ‘Intention is choice with commitment’; E. Moggi’s ‘Notions of computation and monads’ and D. Harel’s ‘Statecharts – a visual formalism for complex systems’.

- **Information retrieval**: S. Deerwester’s ‘Indexing by latent semantic analysis’; A. Smelkel’s ‘Content-based image retrieval at the end of the early years’; T. Landauer’s ‘An introduction to latent semantic analysis’ and A. Tversky’s ‘Features of similarity’.

- **Database**: P. Chen’s ‘The entity-relational model’; G. Wiederhold’s ‘Mediators in the architecture of future information-systems’ and E. Rahme’s ‘A survey of approaches to automatic schema matching’.

- **Bioinformatics**: M. Ashburner’s ‘Gene ontology’.

- **Natural language processing**: G. Miller’s ‘Wordnet – a lexical database for English’; M. Porter’s ‘An algorithm for suffix stripping’ and P. Resnik’s ‘Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language’.

- **Data/text mining**: F. Sebastiani’s ‘Machine learning in automated text categorization’ and A. Dempster’s ‘Maximum likelihood from incomplete data via EM algorithm’.

These highly cited papers in the related fields do not belong to the SW area, but they are highly cited articles by SW researchers. For example, S. Deerwester’s ‘Indexing by latent semantic analysis’ is one of the best algorithms to derive topics therefore forms the fundamental methods for ontology learning. Similar for highly cited papers in database and mediator (as RDF triple stores are related to database), text mining and natural language processing (as they are the major building blocks for ontology learning and mapping), and bioinformatics (as it is one of the leading areas which applies SW technologies and achieves appealing results).

### 5.3. New stars in the Semantic Web

Table 10 shows the top 20 authors with the highest increase of their citations from 2000–04 to 2005–09. In WOS, M.A. Harris (gene ontology-related research), T. Harris (design and implementation
| R  | Paper                                                                 | 1960–2009 | 2000–04 | 2005–09 |
|----|-----------------------------------------------------------------------|-----------|---------|---------|
| 1  | Gruber TR (1993), A translation approach to portable ontology specifications, *Knowl Acquis*, 5, 199 | 598       |         | 121     |
| 2  | Bernerslee T (2001), The semantic web – a new form of web content that is meaningful to computers will unleash a revolution of new possibilities, *Sci Am*, 284, 34 | 416       |         | 54      |
| 3  | Deerwester S (1990), Indexing by latent semantic analysis, *J Am Soc Inform Sci*, 41, 391 | 132       |         | 52      |
| 4  | Gelfond M (1991), Classical negation in logic programs and disjunctive databases, *New Generat Comput*, 9, 365 | 110       |         | 47      |
| 5  | Reiter R (1980), A logic for default reasoning, *Artif Intell*, 13, 81 | 104       |         | 46      |
| 6  | Landauer TK (1998), An introduction to latent semantic analysis, *Discourse Process*, 25, 259 | 89        |         | 89      |
| 7  | Harel D (1987), Statecharts – a visual formalism for complex-systems, *Sci Comput Program*, 8, 231 | 88        |         | 41      |
| 8  | Rahm E (2001), A survey of approaches to automatic schema matching, *VLDB J*, 10, 334 | 87        |         | 39      |
| 9  | Sebastiani F (2002), Machine learning in automated text categorization, *ACM Comput Surv*, 34, 1 | 86        |         | 35      |
| 10 | Porter M.F. (1980), An algorithm for suffix stripping, *Program*, 14, 130–137 | 84        |         | 32      |

(Continued)
| R | Paper                                                                 | No. Cited | Paper                                                                 | No. Cited |
|---|----------------------------------------------------------------------|-----------|----------------------------------------------------------------------|-----------|
| 11 | Kalfoglou Y (2003), Ontology mapping: the state of the art, *Knowl Eng Rev*, 18, 1 | 79        | Meseguer J. (1992), Conditional rewriting logic as a unified model of concurrency, *Theoretical Computer Science*, 96, 73–155 | 29        |
| 12 | Tversky A. (1977), Features of similarity, *Psychological Review*, 84, 327-352 | 79        | McCarthy J (1980), Circumscription – a form of non-monotonic reasoning, *Artif Intell*, 13, 27 | 29        |
| 13 | Alur R. (1994), A theory of timed automata, *Theoretical Computer Science*, 126, 183-235 | 79        | Porter M.F. (1980), An algorithm for suffix stripping, *Program*, 14, 130–137 | 28        |
| 14 | Zadeh LA (1965), Fuzzy sets, *Inform Contr*, 8, 338 | 78        | Jacobs B. (1997), A tutorial on (co)algebras and (co)induction | 25        |
| 15 | Uschold M (1996), Ontologies: principles, methods and applications, *Knowl Eng Rev*, 11, 93 | 75        | Rahm E (2001), A survey of approaches to automatic schema matching, *VLDB J*, 10, 334 | 25        |
| 16 | Studer R. (1998), Knowledge engineering: principles and methods, *Data and Knowledge Engineering*, 25, 161–197 | 74        | Cohen P.R. (1990), Intention is choice with commitment, *Artificial Intelligence*, 42, 213–261 | 23        |
| 17 | Vangelder A (1991), The well-founded semantics for general logic programs, J Assoc Comput Mach, V38, P620 | 73        | Hendler J (2001), Agents and the semantic web, *IEEE Intell Syst App*, 16, 30 | 23        |
| 18 | Hendler J (2001), Agents and the semantic web, *IEEE Intell Syst App*, 16, 30 | 72        | Milner R. (1992), A calculus of mobile processes, *Information and Computation*, V100, PP. 1–77 | 44        |
| 19 | Landauer TK. (1997), A solution to plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review*, 104, 211–240 | 71        | Miller GA (1995), WORDNET – a lexical database for English, *Commun ACM*, 38, 39 | 23        |
| 20 | Horrocks I (2003), From SHIQ and RDF to OWL: the making of a web ontology language, *J Web Semant*, 1, 7–26 | 70        | Moggi E. (1991), Notions of computation and monads, *Information and Computation*, 93, 55–92 | 23        |

(Continued)
of programming languages) and L. Ding (Swoogle – Semantic Web search engine) are ranked as the top three authors with the highest increase of citations. Coming from Scopus, D. Roman (Semantic Web services), J. De Bruijn (logic programming) and L. Ding (Swoogle) are ranked as top three for the significant increase in number of citations.

6. Conclusion

This paper conducted citation analysis for the SW field covering 1960–2009. Papers and citations were collected from two major databases, WOS and Scopus. The productivity and impact of the SW community have been analysed, notably within the last decade of development for the periods 2000–04 and 2005–09. The major publication channels in the SW field are conference proceedings, especially those published by Springer as the series Lecture Notes in Computer Sciences. Major journals that publish SW papers are Theoretical Computer Science, Bioinformatics, Data and Knowledge Engineering and IEEE Transactions on Knowledge and Data Engineering. The most productive authors are T. Eiter, A. Brogi and H. Zhuge. J.J. Jung is the newly emerging, very productive author in this field.

The research impact has been analysed based on citation counting. In the whole period (1960–2009), R. Milner, M. Gelfond and C.A.R. Hoare are ranked as the top three authors. T. Berners-Lee is ranked fourth throughout the period. Scopus citation data allow the ranking of cited second or third authors. J. Hendler, S. Staab and H. Garcia-Molina are ranked as the top three highly cited second authors, while O. Lassila, F. van Harmelen and I. Horrocks are the top three highly cited third

| R | Name               | Web of Science Times of increase | Scopus Name          | Scopus Times of increase |
|---|--------------------|----------------------------------|----------------------|--------------------------|
| 1 | HARRIS MA          | 30.5                             | Roman, D.            | 72.5                     |
| 2 | HARRIS T           | 21.5                             | De Bruijn, J.        | 70                       |
| 3 | DING L             | 20.7                             | Ding, L.             | 43                       |
| 4 | MARCUS A           | 20.5                             | Harris, T.           | 37.5                     |
| 5 | ROMAN D            | 19                                | Rao, J.              | 36                       |
| 6 | CHEN YX            | 18.5                             | Carroll, J.J.        | 35                       |
| 7 | ANTONIOL G         | 17.5                             | Hollink, L.          | 34                       |
| 8 | HAASE P            | 16.8                             | Monay, F.            | 32                       |
| 9 | KNUBLAUCH H        | 16                                | Lara, R.             | 30.5                     |
| 10| ALSHAHROUR F       | 15.3                             | Tang, J.             | 29                       |
| 11| JEON J             | 15                                | Gu, T.               | 28.5                     |
| 12| LIERLER Y          | 14                                | Haase, P.            | 27.6                     |
| 13| LARA R             | 14                                | Bowers, S.           | 27.5                     |
| 14| DONNELLY M         | 14                                | Gauch, S.            | 27.5                     |
| 15| PATWARDHAN S       | 14                                | Snoek, C.G.M.        | 27.25                    |
| 16| PRUDHOMMEAUX E     | 13.6                             | Rosati, R.           | 26.7                     |
| 17| MA YF              | 13                                | Pang, B.             | 26.5                     |
| 18| PANTEL P           | 12.7                             | Prud’hommeaux, E.    | 26                       |
| 19| WANG P             | 12.3                             | Ding, Z.             | 25.5                     |
| 20| FU X, MAXIMILIEN EM, VENNEKENS J | 12 | Akkiraju, R. | 25 |

Notes: times of increase = [(no. of being cited in 2005–09) – (no. of being cited in 2000–04)]/(no. of being cited in 2000–04)
authors. In WOS, T. Gruber’s ontology paper has been consistently highly cited and ranked top for all sub-periods. A. Van Gelder’s theory proving paper is ranked second, and S. Deerwester’s latent semantic analysis paper is ranked third. T. Berners-Lee, J. Hendler and O. Lassila’s article about the vision of the SW, published in *Scientific American*, is ranked as the top second highly cited paper in 2005–09, while in Scopus, Gruber’s ontology paper and Berners-Lee’s *Scientific American* papers are the top two highly cited papers in 1960–2009 and 2005–09. In both WOS and Scopus, the highly cited journals and conferences are *Lecture Notes in Computer Science* or *Lecture Notes in Artificial Intelligence*, *Artificial Intelligence*, *Communication of the ACM* and *Theoretical Computer Science*. In WOS, M.A. Harris, T. Harris and L. Ding are ranked as the top three authors with the highest increase of citations, while from Scopus, D. Roman, J. De Bruijn and L. Ding are ranked as the top three for the significant increase in number of citations.

By comparing highly cited articles in 2000–04 and 2005–09, one can see the research shifting from core AI-related logic programming, logic reasoning and theory proving, to ontological languages (e.g. RDF, OWL), semantic data conversion and ontology mapping. One may therefore predict that, within the next 10 years, the following topics may become mainstreams in this field:

- **Creating, converting and enriching semantic data**: this mainstream effort is led by the Linked Open Data (LOD) Initiatives created by C. Bizer (Free University of Berlin, Germany). LOD bubbles will grow to an amazing degree, becoming the major showcase of SW technologies. LOD creates the test bed for semantic query, reasoning and service/data mash-ups. It demonstrates a powerful, simple, flexible and efficient approach to integrating heterogeneous datasets and triggers the industrial, governmental and academic adoption. Between 2010 and 2020, the efforts might focus on the quality issue of the LOD data, scalability of managing and querying LOD data, and security on data and SPARQL query.

- **Mining semantic RDF/OWL graphs**: SW creates better technologies to represent and integrate data, while all these efforts should lead to the final goal – providing better search technologies. Since RDF data form graphs, the searching and retrieving of RDF data utilizes the current Google approach: PageRank or HITS, wherein the topologies of graphs play the major role in ranking nodes in the networks. RDF graphs contain more semantics than normal graphs in Google, as the links and nodes are instances of the ontologies. Various weighted, topic-sensitive or semantic-sensitive PageRank may therefore become a new research topic in the ranking of semantic nodes. Provenance data once again becomes meaningful, wherein datasets need to be integrated. This development traces different steps of data integration and enables provenance-based layered data analysis, query and visualization.

- **Simple reasoning**: revolutionary breakthroughs should happen during the next few years as complex reasoning fails to scale up. Reasoning should be kept as simple as possible, scalable and error-tolerant. Relaxed or simplified logic may thus be invented to make this fly.

- **Benchmarking and evaluating ontologies**: nowadays ontologies have been created nearly everywhere, as noted in the introduction – a necessary step for solving the information-deluge problem. There is a pressing need to create a benchmark or widely adopted framework to evaluate and test these ontologies. Notably during the process of generating ontologies, domain experts may have a handbook to ensure right decisions on the modelling of their classes, properties or instances. Examples may be found from other communities, such as TREC in information retrieval.

- **Interfacing SW**: the next 10 years should see the creation of an innovative user-friendly interface to showcase the SW. Actually achieving goals of the SW is still currently impossible, as the search interface or SPARQL Endpoints for LOD datasets are not really targeted for normal users, and are instead accessible to SW gurus or hackers. To bring the SW out of the research lab and make its debut for normal users, a simple interface design is essential.

- **Utilizing social web (Web 2.0)**: the current social network fever in Web2.0 facilitates the generation of social semantic data, such as social tagging, commenting, voting and recommending. These data identify existing relationships and create new ones, forming a ‘social power’ that helps the LOD community
snowball their datasets and introduce mash-up powers of SW technologies. In the next 10 years, we may predict that Web 2.0 and the SW will be merged or interwoven in the manner that motivates normal Web 2.0 users to contribute more social metadata, while SW should provide better technologies to mash-up these data and further stimulate data generation. The difference between Web 2.0 and the SW will become blurred, as they finally merge to become the next generation web – Web 3.0 – which extends current Web 2.0 applications using SW technologies and graph-based open data [27].

- **Embracing eScience and eGovernment**: in the next 10 years, eScience and eGovernment will be the major adopters of SW technologies. The current trend toward data integration, interlinking and analysis within health sciences, biology, medicine, pharmaceuticals and chemistry will lead to new technologies such as bio2rdf, Linked Open Drug Data and YeastHub. Semantic publishing will create new norms for the next generation of publishing. Journal or conference papers are no long just pure ‘static strings’. They contain important RDF triples which are interlinked in the paper, with other related papers (e.g. citations), and outside related semantic datasets (e.g. LOD bubbles). The substantial funding secured from National Institutes of Health for CTSA and research networking for life science indicates the confidence and uptake of SW technologies from other major funding agents in the USA, including the National Science Foundation. The recent groundbreaking news from the USA and UK is that their governments are ready to use the potential of SW technologies to build their transparent eGovernement platform [5]. These will create the tremendous momentum and broad social and societal impact of the SW. This momentum will radiate through other fields, such as environmental science, to integrate data from hydrology, climatology, ecology, and oceanography [28].

The challenges to the SW may be as significant as its promises. As I. Horrocks mentioned in his recent article, ‘The vision of a Semantic Web is extremely ambitious and would require solving many long-standing research problems in knowledge representation and reasoning, databases, computational linguistics, computer vision, and agent system.’ [9]. To carry on and further realize this vision, the SW community needs to work with researchers from related fields to establish the SW as the emerging interdisciplinary field – called ‘web science’ – to view the World Wide Web as an important entity to be studied in its own right, and to understand its future as a computational structure and an interacting platform for people and machines [10].

Although there are twice as many SW papers in Scopus as in WOS, the citation analysis does not show a significant difference between the two. For future research, we plan to use social network analysis to detect research groups or communities in this field. The use of self-citation also poses a new area of research that can be further extended to group self-citation or project self-citation in papers citing or cited by authors from the same research group or related projects. This may help identify the knowledge diffusion and transfer patterns in this field, as new and existing thinkers within this closely knit community become necessarily self-referential.

Endnotes

1 http://isiwebofknowledge.com/media/pdf/cpci_faq.pdf
2 Of course, there can be many other terms to retrieve related data in the SW field due to its multidisciplinary feature. But in this paper, we set our focus on research related to semantics and ontology (as addressed in the Introduction), which are the crucial parts of the field. Other potential terms (e.g. RDF, XML, OWL, Linked Open Data, LOD, SPARQL, et al.) are therefore not included to retrieve data.
3 TS in WOS include Title, Abstract, Author Keywords and Keywords Plus
4 Just for testing purposes, in February 2010, there are around 7600 articles in WOS having OWL* in the title, keyword or abstract. Less than 10% of them are related to SW. Among them, more than 95% have ontology* or semantic* in the title, keyword or abstract. By using RDF* as a search term for WOS, less than 30% of articles with RDF in the title, abstract or keyword are related to the SW. Among them, more than 90% have ontology* or semantic* in the title, keyword or abstract.
References

[1] T. Berners-Lee, J. Hendler and O. Lassila, The semantic web, *Scientific American* 284(5) (2001) 34–43.
[2] PriceWaterhouseCoopers, *Report on Semantic Web* (2009). Available at: http://ivan-herman.name/2009/06/01/pwc-report-on-semantic-web/ (accessed 19 September 2009).
[3] D. Peterson, *Obama’s groundbreaking use of the Semantic Web* (2009). Available at www.sitepoint.com/blogs/2009/03/19/obama-groundbreaking-use-semantic-web/ (accessed 19 September 2009).
[4] Pingdom, *Web growth peaked in 2007 but might be back with a vengeance in 2009* (2009). Available at: http://royal.pingdom.com/2009/05/07/web-growth-peaked-in-2007-but-might-be-back-with-a-vengeance-in-2009/ (accessed 19 September 2009).
[5] N. Shadbolt, W. Hall and T. Berners-Lee, The semantic web revisited, *IEEE Intelligent Systems* 21(3) (2006) 96–101.
[6] A.L. Barabasi, *Linked: The New Science of Networks* (Perseus Publishing, Cambridge, MA, USA, 2002)
[7] www.oracle.com/newsletters/information-insight/content-management/feb-07/forrester-future.html
[8] C. Bizer, T. Heath and T. Berners-Lee, Linked data – the story so far, *International Journal on Semantic Web & Information Systems* 5(3) (2009) 1–22.
[9] I. Horrocks, Ontologies and the Semantic Web, *Communications of the ACM* 51(12) (2008) 58–67.
[10] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee and T. Weitzner, Web Science: an interdisciplinary approach to understanding the web, *Communication of the ACM* 51(7) (2008) 60–69.
[11] J. Hendler, Avoiding another AI winter, *IEEE Intelligent Systems* 23(2) (2008) 2–4.
[12] V. Retzer and G. Jurasinski, Towards objectivity in research evaluation using bibliometric indicators: a protocol for incorporating complexity, *Basic And Applied Ecology* 10(5) (2009) 393–400.
[13] H.F. Moed, *Citation Analysis in Research Evaluation* (Kluwer Academic Publishers, Dordrecht, 2005)
[14] P. Riikonen and M. Vihinen, National research contributions: a case study on Finnish biomedical research, *Scientometrics* 77(2) (2009) 207–222.
[15] H.D. White and K.W. McCain, Visualizing a discipline: an author co-citation analysis of information science, 1972–1995, *Journal of the American Society for Information Science* 49(4) (1998) 327–355.
[16] J.E. Hirsch, An index to quantify an individual’s scientific research output, *Proceedings of the National Academy of Sciences* 102(46) (2005) 16569–16572.
[17] A.A. Sorensen, Alzheimer’s disease research: scientific productivity and impact of the top 100 investigators in the field, *Journal of Alzheimer’s Disease* 16(3) (2009) 451–465.
[18] M.E.J. Newman, The structure of scientific collaboration networks, *Proceedings of the National Academy of Science of the United States of America* 98(2) (2001) 404–409.
[19] E. Yan and Y. Ding, Applying centrality measures to impact analysis: a coauthorship network analysis, *Journal of the American Society for Information Science and Technology* 60(16) (2009) 2107–2118.
[20] Y. Ding, E. Yan, A. Frazho and J. Caverlee, PageRank for ranking authors in co-citation networks, *Journal of the American Society for Information Science and Technology* 60(11) (2009) 2229–2243.
[21] C.P. Huang, Bibliometric analysis of obstructive sleep apnea research trends, *Journal of the Chinese Medical Association* 72(3) (2009) 117–123.
[22] B. Thijs and W. Glanzel, A structural analysis of benchmarks on different bibliometrical indicators for European research institutes based on their research profile, *Scientometrics* 79(2) (2009) 377–388.
[23] P. Mika, Flink: semantic web technology for the extraction and analysis of social networks, *Journal of Web Semantics* 3(2005) 211–223.
[24] P. Mika, T. Elfring and P. Groenewegen, Application of semantic technology for social network analysis in the sciences, *Scientometrics* 68(1) (2006) 3–27.
[25] D. Zhao and A. Strotmann, Information Science during the first decade of the web: an enriched author co-citation analysis, *Journal of the American Society for Information Science and Technology* 59(6) (2008) 916–937.
[26] L.I. Meho and K. Yang, Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar, *Journal of the American Society for Information Science and Technology* 58(13) (2007) 2105–2125.
[27] J. Hendler, Web 3.0 engineering, *IEEE Computer* 42(1) (2009) 111–113.
[28] E. Ostrom, A general framework for analyzing sustainability of social–ecological systems, *Science* 325(5939) (2009) 419–422.