Web Science and the Two (Hundred) Cultures: Representation of Disciplines Publishing in Web Science

Clare J. Hooper
IT Innovation Centre,
University of Southampton, UK
cjh@it-innovation.soton.ac.uk

Georgeta Bordea
Digital Enterprise Research
Institute, NUI, Galway
bordea.georgeta@deri.og

Paul Buitelaar
Digital Enterprise Research
Institute, NUI, Galway
buitelaar.paul@deri.org

ABSTRACT
Web Science is an interdisciplinary field. Motivated by the unforeseen scale and impact of the web, it addresses web-related research questions in a holistic manner, incorporating epistemologies from a broad set of disciplines. There has been ongoing discussion about which disciplines are more or less present in the community, and about defining Web Science itself: there is, however, a dearth of empirical work in this area.

This paper presents an analysis of the presence of different disciplines in Web Science. We applied Natural Language Processing and topic extraction to a corpus of Web Science material, analysing it with graphing and visualisation tools, MatLab and an expert survey. We discovered four communities within Web Science, and trends in the conference series over time (a strong impact from collocation) and format (posters covering a broader range of topics than papers). The expert survey linked highly ranked terms with disciplines, yielding strong links with Communication, Computer Science, Psychology, and Sociology. Controversially, experts described highly ranked topics and suggested disciplines (extracted from WebSci CFPs) as not reflecting the nature of Web Science.

Author Keywords
Web Science discipline; community analysis; bibliometrics; disciplines; Saffron.

ACM Classification Keywords
K.4.m. Computers and Society: Miscellaneous.

General Terms
Human Factors; Measurement; Theory.

INTRODUCTION

In 1959, C. P. Snow famously gave his lecture, 'The Two Cultures' [22], in which he lamented the division of the sciences and the humanities, and the negative impact of that division upon intellectual progress across society. Web Science, like certain other disciplines, is at its heart radically interdisciplinary… or is it?

There has been ongoing discussion about the representation of various disciplines within the Web Science community. Forming a stable, diverse community is no small task: members of the Web Science Trust have worked to try and ensure that the community is balanced with a rich variety of well represented disciplines, and not dominated by one field such as Computer Science.

Defining Web Science can be difficult. Tools to describe the field include the 'Web Science butterfly' diagram, used early in the life of Web Science to convey the vision [18], but this diagram is a vision rather than an accurate depiction of the state of the field [12]. Similar, the Web Science Subject Categorisation [23] only offers a vision and structure, not information on subjects' prevalence within the community. This is problematic. Understanding the actual presence (measured by publications) of different disciplines within Web Science offers several advantages, letting us: better communicate what work is done under the WebSci flag; ground dialogue about Web Science diversity and disciplinary representation with data, identifying under- and over-represented disciplines, and absent disciplines; identify problems that need addressing, and take action by seeking collaborations and communities that would remediate current weaknesses within Web Science.

One paper at WebSci'12 began to examine this area, proposing a methodology and presenting early results that were yielded by this methodology. (The next section details differences between that work and this.) We build on that work, drawing on a corpus of papers from past Web Science conference proceedings, journal.webscience.org, and other sources. We used Natural Language Processing to extract terms from these, and conducted a network analysis of the resultant materials (which we have made available online, with links to the corpus[1]). This paper presents an analysis and discussion concerning:

1. Communities found within the corpus
2. Changes in the Web Science conference series over time
3. Changes in Web Science conference publications according to format

WebSci'13, May 2–4, 2013, Paris, France.

[1] See: clarehooper.net/WebSciCorpus

4. An expert survey regarding the mapping of terms to disciplines

We analysed communities within the corpus to gain insight into the relationship between different parts of the Web Science community. Our decision to analyse the conference series was a conscious one: the Web Science conference is in many ways the heart of Web Science, being the main annual gathering for Web Science researchers and practitioners. As such, the balance of disciplines at these conferences holds strong implications for Web Science as a whole: we therefore conducted an additional analysis examining these conferences, with two hypotheses:

The WebSci conference is often co-located with other events (WWW in 2010, Hypertext in 2011, NetSci in 2012, and CHI this year, 2013). This led to Hypothesis 1: co-location with other events influences what disciplines (as measured by terms) are present at WebSci.

We are interested in what differences, if any, can be discerned between poster and paper contributions. WebSci historically hosts extremely high quality poster sessions, but nonetheless, poster submissions are typically subject to somewhat less rigorous standards than paper submissions. This led to Hypothesis 2: the distribution of disciplines represented by posters (as measured by terms) is broader than that of disciplines represented by papers.

Finally, an expert survey was conducted. The task of relating terms to disciplines is a difficult one, and to avoid issues of subjectivity we pursue this path.

This paper is structured as follows: we open by introducing the area of bibliometrics and the use of Natural Language Processing (NLP) to analyse a corpus of data. We describe our method, from initial data gathering, through processing and visualizing the data using Saffron [15], Gephi and MatLab[2], to the expert survey to gain insight into links between terms and disciplines. We then present our results, including an analysis of communities within the resultant graph, a closer look at the WebSci conference series over time and by format, and the results of the survey. After discussing these results, we present our conclusions.

BACKGROUND AND RELATED WORK

Previous work in bibliometrics ranges from co-citation analysis [8] [24], to examination of multiple conference series [11], to geospatial visualisations of collaboration [16]. Excepting a Web Science paper last year [12], little prior work analyses the disciplinarity of conferences, although Web Science students at the University of Southampton produced an illustration of their own disciplines (based on supervisor disciplines) in March 2011.

[2] http://saffron.deri.ie, https://gephi.org, and http://www.mathworks.nl/products/matlab/

According to [5] a bibliometric map can be constructed by analysing various types of items including journals, papers, authors, and descriptive terms. The work presented in this paper is based on a basic assumption in bibliometric mapping [5], which states that a research field can be described by a list of important keywords. While previous work made use of author assigned key phrases and already built domain taxonomies [9], we applied an automatic method [6] for the extraction of domain terms as such resources are not readily available for our dataset. This method was previously applied for expert profiling [15] in Saffron, a system that provides insights in a research community or organization by analysing its main terms of investigation and the experts associated with these terms.

Implicit relations between the extracted topical descriptors can be discovered and described through word co-occurrence analysis, a content analysis technique that was effectively applied to analyse interactions in different scientific fields [7, 9]. This technique was applied to analyse the interconnections between a main field, i.e., fuzzy logic theory, and other computing techniques [14], a setting that is similar to our analysis of the Web Science field. A more recent work on co-word analysis [25] outlines several limitations related to the use of keywords and proposes a method to integrate expert knowledge into the process. A main issue with this approach is that it requires a considerable amount of human intervention for the construction of domain specific thesauri. We alleviate this challenge by completely automating the process of identifying topical descriptors and by automatically constructing a domain taxonomy.

A short WebSci'12 paper presented initial work in this area [12]. This paper covers new ground, in both breadth and depth: the earlier work analysed only 69 papers from WebSci'09, '10 and '11, compared to the 469+ articles that we examine, including the proceedings of prior Web Science conferences and other sources. Two flaws in the previous paper were that (1) it depended to a significant degree on subjective interpretations of graph structures and taxonomies and (2) attempts at subject demarcation require knowledge of the political and ideological boundaries that have developed over the years (e.g. some people see Criminology as a field that stands outside of Sociology, others see it as a sub-discipline). To mitigate these issues, we include an expert survey of terms.

METHOD

We took the overall approach of using NLP to extract terms from a large corpus of Web Science publications. We then analysed this data via graphing and visualisation. We analysed the communities present within the corpus of data and data concerning the Web Science conference series specifically. We also conducted a survey to gain insight into perceptions of the linkage between different disciplines and key terms from the corpus.

We chose not to analyse co-authorship or co-citation data. We took this approach since our focus was on disciplines, which are better identified by term and not author: many authors – particularly in the WebSci community – have written within a diversity of disciplines.

Data Gathering

The focal point of our corpus was journal.webscience.org, which aims to "collect and highlight Web Science literature and to provide a location for the research outputs of the Web Science community". Table 1 shows the data provided from this source, which spanned papers, posters, panels and keynotes (excluding 4 articles from a British Library Workshop on Ethics and the Web, as none of those articles could be processed by Saffron). Occasionally, articles on journal.webscience.org were listed without an associated PDF. When this was the case, we instead included a text file containing the paragraph abstract on the webpage.

Source	Number of items (number usable)
WebSci 2009	147 (133)
WebSci 2010	109 (109)
WebSci 2011	116 (116)
Oxford Internet Institute Symposium: Dynamics of the Internet and Society	42 (42)
Web Evolution Workshop 2008	16 (16)
Royal Society Discussion Meeting	9 (9)
PLE (Personal Learning Environment) Conference 2011	75 (43)
WWW 2001	1 (1)

Table 1. Publications (papers, posters, etc.) analysed from journal.webscience.org.

Table 2 shows the additional publications that we included:

Source	Number of items (number usable)
WebSci 2012	N/A
Foundations & Trends, from 1(1) to 3(2)	7 (7)
Other key papers on Web Science	6 (6)

Table 2. Other publications subject to analysis

The WebSci'12 proceedings were processed as 3 PDFs representing posters, papers and panels; as each article was not processed separately it is hard to count how many were usable. 366 terms were extracted in total, showing a very noisy baseline: on average in the rest of the corpus we extracted 54 terms per article. Not all items were usable; the Data Processing section gives detail about this.

We included WebSci'12 publications (sourced from the WebSci'12 webpage) and publications from Foundations & Trends in Web Science for the obvious reason: these publications are clearly a relevant part of the Web Science corpus. Note that the 7 files for Foundations & Trends constitute a large mass of data, with a total of 798 pages.

We included 6 key Web Science papers that were – surprisingly – not in our corpus, as they were not present on journal.webscience.org. Papers were drawn from the recommended reading list of a forthcoming encyclopedia article on Web Science. These are: 'Creating a Science of the Web' [2], 'Linked Data – the Story so Far' [3], 'Web Science: An Interdisciplinary Approach to Understanding the Web' [10], 'The Semantic Web Revisited' [19], 'Web Science Emerges' [20], and 'Web Science: a provocative invitation to Computer Science' [21].

Data Processing: Saffron and Gephi

The number of files processed does not equal the number of publications, since some files contained multiple articles (e.g. the 3 PDF files for WebSci'12). In total, we handled 552 files. We processed these using Saffron, an application to understand research communities [15]. Saffron uses information extracted from unstructured documents using Natural Language Processing techniques. We used the topic extraction component with the parameters: maximum topic length 5; web filter minimum 5 hits; web filter maximum 1 billion hits. We used the ACM Subject Classification to build linguistic patterns for terms in Computer Science.

Of 552 files, 491 were included in the analysis. 61 files were not processed, due to being of a format that Saffron, our processing tool, could not use. Saffron can only process plaintext and PDF files, meaning that Word documents and PowerPoint files were excluded.

The Saffron analysis yielded 5371 phrases that were identified as research term candidates, with an average of 54 candidates per document (although no term was extracted for 6 of the analysed documents). Only the top 20% of terms are considered in our analysis. This threshold was necessary because the quality of terms influences the taxonomy of concepts: it is important to choose meaningful terms before analysing the relations between them. The research terms are not manually curated, therefore they include incorrect terms such as 'future research', which is not a Web Science term. Like any other tool, term extraction and analysis has some limitations, and the appearance of 'future research' as an important term exemplifies the issue of incorrectly extracted terms.

The index used in co-word analysis to measure the strength of relationships between two research terms is defined as:

$$I_{ij} = D_{ij} / (D_i D_j)$$

where D_i is number of articles that mention the term T_i in our corpus, D_j is number of articles that mention the term T_j, and D_{ij} is the number of documents in which both terms appear.

Edges are added in the research terms graph for all the pairs that appear together in at least 3 documents. Saffron uses a

generality measure to direct the edges from generic concepts to more specific ones. This step results in a highly dense, noisy directed graph that is further trimmed using an optimal branching algorithm. An optimal branching is a rooted tree where every node but the root has in-degree 1, and that has a maximum overall weight. This algorithm was successfully applied for the construction of domain taxonomies in [17]. This yields a tree structure where the root is the most generic term and the leaves are the most specific terms.

We used a network graph tool, Gephi, to build a graph showing links between terms: nodes are extracted terms and arcs are papers that link them. This let us identify 'clusters' of closely related terms. We used the Yifan Hu algorithm [13] to layout the graph, and between-ness centrality to weight node importance. Betweenness centrality measures the fraction of shortest paths going through a node [1]: a high value indicates that nodes play an important bridging role in a network. Finally, we ran the Louvain method [4] with resolution 12 to detect communities.

Data Processing: MatLab

MatLab was used to examine the Web Science conference proceedings, by tracking four variables: keyword, year (2009, 2010, 2011), type (poster, paper), and count type (number of documents to contain keyword, overall keyword occurrence). The formatting of the WebSci'12 proceedings meant that the data was largely unsuitable for processing using our methods; this analysis concerns the proceedings of WebSci'09, WebSci'10 and WebSci'11.

Expert survey

A key issue in the original work in this area [12] was the subjectivity with which extracted terms were categorised as falling into disciplines: essentially, the three researchers reviewed the terms and came to an agreed mapping of terms to disciplines. Although it is difficult to do so, it is important to consider the relationship between terms and disciplines: for this reason, we ran an expert survey.

We chose to approach experts in the field of Web Science and ask them to map disciplines to terms. We recruited experts from our own personal networks, making a point of targeting experts from a wide range of disciplines.

We provided the experts with the top 20 extracted terms (ranked by betweenness centrality), although we removed one meaningless term ('future research', discussed above). We asked the experts to map those terms into disciplines, including a suggested discipline list but not requiring that they keep to that list. The list was made up of every discipline to have been mentioned in past 5 CFPs for the Web Science conference (2009 – 2013): Communication; Computer and Information Sciences; Criminology; Design; Digital Humanities; Economics; Geography; Language and Communication; Law; Linguistics; Management; Political Science; Sociology; Philosophy; Psychology.

13 experts responded. They came almost entirely from academia (12/13); the industrial responder is Chief Scientist at a relevant company. The academics consisted of 2 professors, 4 lecturers, 3 postdocs and 3 PhD students. 12 respondents had worked in WebSci (1 described as having done 'related work'), and 11 had published at the WebSci conference. 4 respondents described their main discipline as WebSci, with the other main disciplines described as Archaeology (2), Computer Science/Software Development (3), Digital Humanities (1), Health Sciences (1), Law (1), NLP (1). All respondents reported working in additional fields, which is unsurprising given that we specifically targeted Web Science researchers and practitioners.

RESULTS

Figure 1 shows a visualisation of the extracted terms, where larger nodes and label fonts indicate terms with a higher betweenness centrality. Table 3 lists terms with a high betweenness centrality:

Betweenness centrality value	Term
758	semantic web
590	social media
504	information retrieval
495	social networking site
456	social science
454	search engine
434	social networking
360	learning network
304	web page
297	personal learning environment
282	social interaction
270	mobile device
260	future research
258	internet user
246	uniform resource identifier
235	web science research
235	user interface
235	web community
234	web application
231	linked data principle

Table 3. The 20 terms with highest betweenness centrality

Communities

The community detection algorithm found 9 communities, shown in Figure 1. Each community had its own subset of terms, which had been ranked by the function used during topic extraction (rank range: 0 to 22). Table 4 details the 9 communities, including for each community its most highly ranked 5 terms and how many 'hot' terms (terms with a score above 10) it contains.

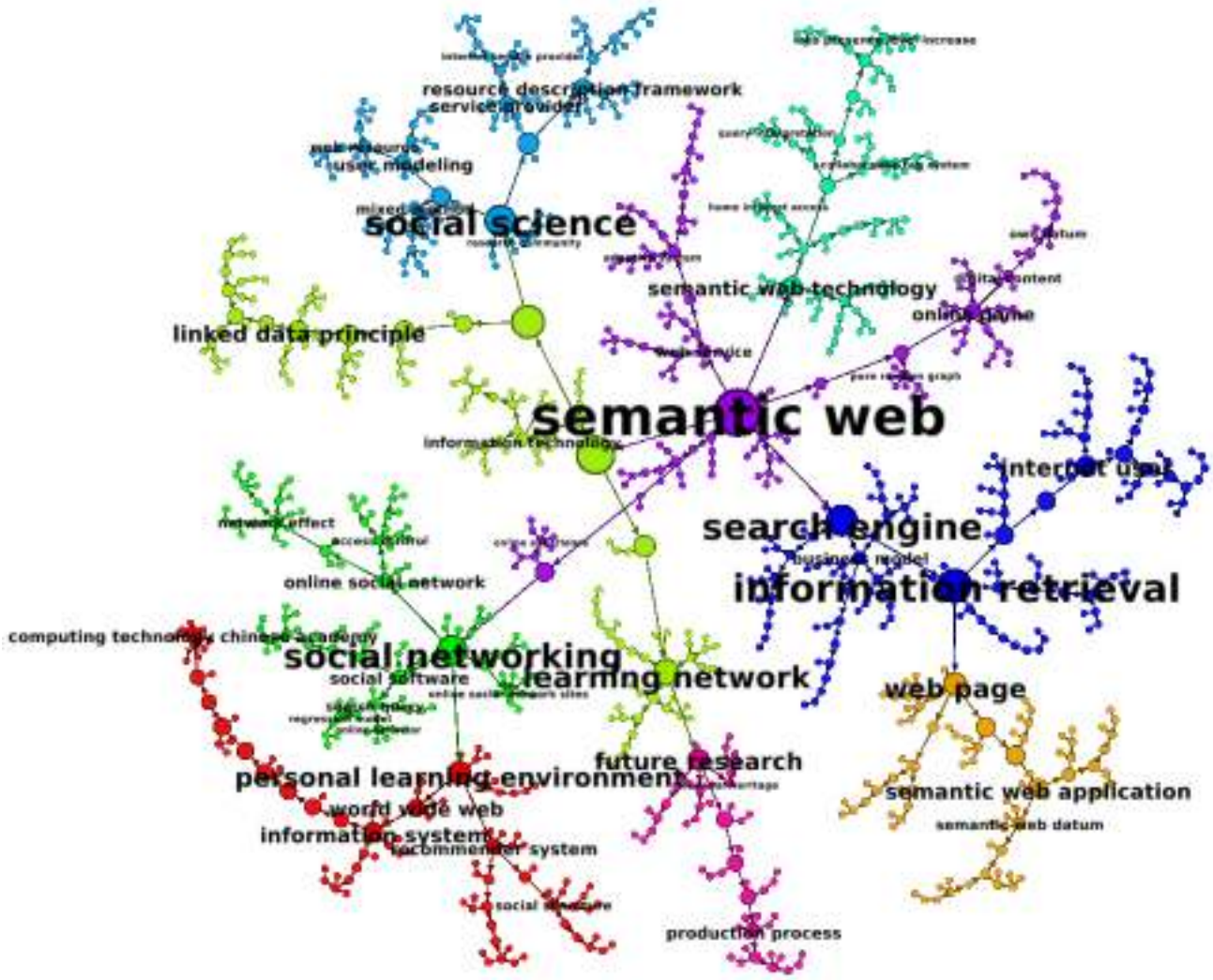

Figure 1. A visualisation of the extracted terms. Colours indicate communities.

Root Node	# hot terms	% of graph	Top 5 terms
Search Engine	22	5	search result; open data; web search; information retrieval; natural language
Semantic Web	12	10	web science; data source; random graph; graph pattern; data set
Personal Learning Environ-ment	9	10	world wide web; information system; mobile web; web archive; research information system
Social Science	9	15	web science research; p2p network; mobile device; user modeling; service provider
Social Media	8	8	social networking site; data mining; web site; information technology; social interaction Social
Web Page	6	13	web technology; user interface; web application; rdf data; semantic web application
Semantic Web Technology	3	12	social web; social bookmarking systems; linked data; information source; public sector information
Future Research	1	15	cultural heritage; learning network; production process; open source blogging

			platform; credibility evaluation
Social Networking	1	12	social software; search query; operating system; analyzing social bookmarking systems; data management

Table 4. Summary of the 9 WebSci communities

The WebSci Conference Series

Differences over time

One aspect to impact material at the conference is acceptance rate. The first three years of the WebSci conference were relatively stable, with 21%, 26% and 15% of submitted papers accepted.

To understand term diversity over time, we counted how many of the top 1000 WebSci terms were included each year. Over 2009 to 2011 the number shifted between 609 and 708 terms, showing a relatively small variation.

We sought 'peak' terms, identifying terms to occur in five or more publications. If such a term peaked in a given year in both papers and posters, it was defined as a 'peak term'. Initial results included all terms that occurred more in one year than in others, which yielded some false peaks: for example, 'public sector' occurred 3 times in 2009 and 2010 and 4 times in 2011. We discarded such results, keeping only peaks where the overall variation in frequency was greater than 5 papers in different years. We also discarded a false peak, 'commercial advantage', from 2011: this arose from a change in the wording of the copyright statement.

This yielded 2 peaks in 2009, 10 peaks in 2010 and 1 peak in 2011:

- 2009: machine learning; real world
- 2010: available online; information exchange; information retrieval; information sharing; natural language; RDF graph; real time; semantic web; share information; SPARQL query
- 2011: social media

Differences by format

We used MatLab to track term diversity across papers and posters, examining how many of the top 1000 WebSci terms they included. Papers cover 70% of top terms, while posters are more diverse, covering 83% of terms.

Peak terms' average 'height' (the difference between their minimum and maximum occurrence over time) is relevant. Average height in posters is 4.8, and in papers is 3.9.

Expert Survey

To gain insight into possible links between extracted terms and disciplines, a short survey of 13 experts was conducted. Table 5 summarises the results, showing the number of disciplines suggested in relation to each term, and enumerating disciplines that were mentioned in relation to each term by at least three experts.

Term	**Number of associated disciplines**	**Disciplines named by at least 3 experts**
linked data principle	1	CompSci
information retrieval	2	Computer and Information Sciences (CompSci)
uniform resource identifier	4	CompSci
web science research	4	Any/all; CompSci; Web Science
semantic web	7	CompSci
user interface	7	CompSci; Design
search engine	8	CompSci
web application	8	CompSci
web page	8	Any/all; CompSci
internet user	9	CompSci; Psychology; Sociology
social science	9	Sociology
personal learning environment	10	CompSci; Education
web community	10	CompSci; Psychology; Sociology
learning network	11	CompSci; Pedagogy
mobile device	11	Any/all; CompSci; (Industrial) design
social networking	11	Communication; CompSci; Sociology
social networking site	11	Communication; CompSci
social interaction	12	Sociology; Psychology
social media	12	Communication; CompSci; Network Science; Sociology

Table 5. Disciplines associated with each term.

DISCUSSION

The initial graph of the extracted terms alongside the top-rated terms by betweenness centrality present no huge surprises: terms such as semantic web and social media are central to Web Science and would be expected to be highly visible. More interesting results can be found in the deeper analyses:

Communities

9 communities were detected: in most cases, their root nodes refer to research terms. 'Future research' is an exception to this, a term used across many disparate papers, terms and disciplines: unsurprisingly, the top 5 terms of this latter community are incongruent.

The 'personal learning environment' community branches off the 'social networking' community, and has 9 hot (and related) terms. It seems likely that the majority of the material in this community comes from the Personal Learning Environment conference that was included within the corpus on journal.webscience.org.

Given the terms of the 'search engine' community (which has the highest number of hot terms: 22), it seems more accurate to describe this as the 'information retrieval' community. The 'web page' community that hangs off it and refers to such terms as web technology and user interfaces is clearly related. As this community was so large, we analysed it further, finding 3 sub-communities:

1. personal information (terms include: internet user, information network, social networking service, social sharing, law enforcement)
2. information retrieval (natural language, online community, knowledge management, sentiment analysis)
3. search engine (business model, system design, open data, web search)

It can be seen that the search engine community is divided between real world applications and domains, and core information retrieval terms. The third sub-'community' is disparate in nature.

The 'semantic web' and 'semantic web technology' communities are clearly part of the same movement within Web Science, sharing between them a total of 15 hot terms.

The 'social networking' community also only has one hot term, although its terms are more cohesive. It seems likely that the related 'social media' community (with 8 hot terms) is the reason for this community's weaker ranking.

Last but not least is the 'social science' community, branching off from the 'social networking' community and hosting 9 hot terms. This community was subject to further investigation, because unlike the others its root node refers to a (set of) discipline(s). As in shown Table 4, its 5 top terms are somewhat varied, although 3 touch on mobile networks and the internet. When analysing this community in its own right, Gephi revealed four (somewhat messy) sub-communities, shown in Figure 2:

1. RDF / knowledge representation
2. Intellectual property / machine translation / information security
3. User modeling / cognitive science
4. P2P networks

These sub-communities match various of the 9 communities of the original graph (particularly, semantic web and information retrieval), but are notably heterogeneous. It would seem that the social science community arose from Gephi picking up the 'social science' keyword across disparate research: as such, this set of data perhaps does not constitute a real 'community'.

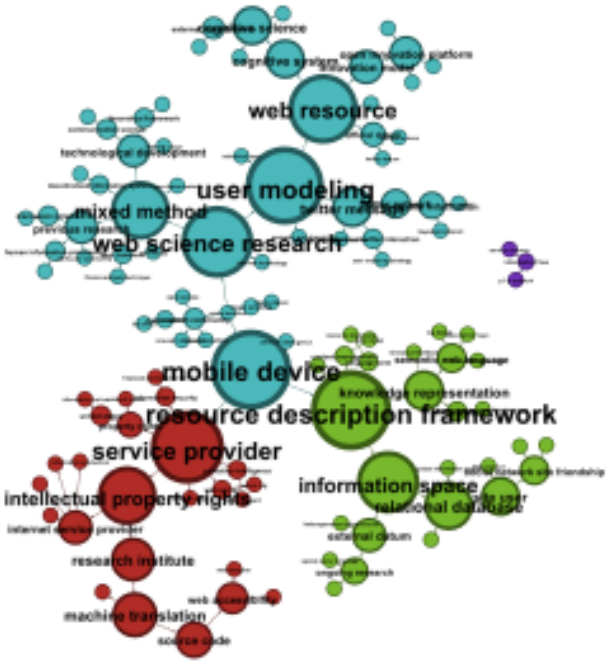

Figure 2. Visualisation of the 'social science' community.

In sum, of the 9 auto-detected communities we can clearly see representation of the following Web Science communities: information retrieval; personalised learning/elearning; semantic web; social networking.

Web Science Conference series

Differences over time

There was no significant variation in how many terms were covered during each year of the Web Science conference, nor in acceptance rate.

Figure 3 shows the number of WebSci publications to heavily use 'peak' terms: peak content and frequency are both notable.

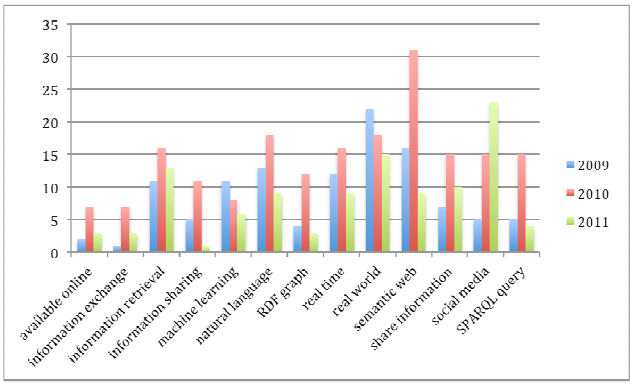

Figure 3. Number of WebSci publications to heavily use 'peak' terms

Regarding frequency, 2010 had far more peaks. This can't be related to the number of papers: we processed 133, 109,

and 116 publications respectively from 2009, 2010 and 2011. This suggests that the many peaks are due instead to a shift in focus within the accepted papers in 2010: 2009 and 2011 had a broader focus, and thus fewer peaks

Regarding content, the 2010 and 2011 peaks are noteworthy. The 2010 peaks are very strongly related to web and semantic web technologies: terms such as information retrieval, RDF graph and SPARQL query show a clear focus in this area which would seem to suggest a strong influence from colocation with WWW'10. WebSci'11 had only one peak, but this was very strongly related to Hypertext'11: 'social media' was in fact a track at the 2011 Hypertext conference.

This shows a strong influence from colocation with other conferences. This influence is, of course, precisely the point of collocating conferences: bringing together groups of people with like interests. The strength of the impact of collocation is clear: perhaps people choose to submit papers to WebSci if they know they are going to a collocated conference regardless. At the same time, some caution is advised: is the PC likely to be biased against papers that don't fit with the companion conference? We see, for example, an extreme drop in the frequency of the term 'RDF graph' moving from 2010 to 2011 (12 papers to 3), and likewise 'semantic web' (31 to 9). Does this mean that collocating with Hypertext not only benefitted researchers investigating social media, but actively disadvantaged semantic web researchers?

Differences by format

The initial results showed that Web Science papers and posters covered 70% and 83% of the 1000 identified WebSci terms. It is unsurprising to see that posters covered more terms, as one would expect greater diversity in poster contributions. We hypothesise that the percentage of Web Science terms covered by Web Science conference materials would be both higher (for both papers and posters) were the PLE conference materials (from journal.webscience.org) excluded from the corpus.

One might expect fewer (or weaker) peak terms in posters than papers, since we expect posters to cover a broader diversity of terms rather than to converge the way papers might. In fact, poster versus paper peaks are not strongly different: the average peak height (the difference between minimum and maximum instances of a term: e.g., as in Figure 3) in posters is 4.8, while papers in fact have a smaller height of 3.9. This suggests that perhaps posters and papers respond similarly to colocation.

Expert Survey

A deep analysis of the survey results is beyond the scope of this paper, but Table 5 clearly shows a high variance in the number of disciplines to be associated with a term. 'Information retrieval' and 'uniform resource locator' are prime examples of terms where the majority of respondents immediately associated the term with Computer Science and nothing else. By contrast, some terms yielded wildly diverse discipline lists: examples included all terms to do with social interaction, media and networking (each yielding over 10 disciplines), and also 'learning network' and 'mobile device'. Many of the suggested disciplines were only suggested by one or two separate experts, and so are not enumerated in Table 5: nonetheless, it can be seen that relating a discipline to these terms is controversial. While 'information retrieval' is a generic name for a set of techniques from Computer Science, 'social media' can be the object of study of multiple disciplines: it is probably this key difference that explains why some terms had more diverse connections to other disciplines.

When examining disciplines named by at least three experts, we see a preponderance of responses naming Computer Science and Sociology. Given Web Science's traditional foundation upon these two disciplines, this is perhaps no big surprise. Other strongly present disciplines were Psychology and Communication, which were both named by at least three experts in relation to three separate terms. There is no relationship between how highly ranked terms were and how controversial they were when experts related disciplines to them: the 5 top ranked terms (semantic web, social media, information retrieval, social networking site, social science) had 7, 12, 2, 11 and 9 disciplines associated to them respectively.

It is perhaps disheartening to see that at least three experts associated Computer Science with the term 'web science research', but that the same was not the case for any other discipline (except Web Science itself!) -- even Sociology, which was otherwise frequently named by the experts.

Unsolicited comments from the experts are informative. Some experts criticised the lists ("the [term] list seems to be very much slanted towards technology and away from anything like law, economics, sociology"; "you need to add all the [humanities] disciplines if you're going to add philosophy […] And what about art, design, media studies, gender studies?"; "There are some startling absences, e.g. business studies, art, culture […] and education"). This has implications for a) the meaning of the top-rated terms (do they imply that WebSci is in fact only the study of technology?) and b) the decisions made regarding what disciplines to enumerate in WebSci conference CFPs.

There is a larger debate unfolding here: can terms – whether extracted via NLP or by hand – ever reflect or represent particular disciplines? Some terms, such as information retrieval, mapped clearly to a single discipline, but many did not, occurring across many disciplines: terms such as social media, social networking, and mobile device might occur in any field of study, in very different ways. Indeed, some terms will mean wholly different things according to context (consider 'social networking' in Sociology, and in Computer Science). When a term is *in situ* in a publication, it has much contextual information (arguments made,

methods used, authors' backgrounds) that the topic extraction process strips out. Solutions include: displaying clusters of related terms; identifying related terms through co-occurrence; using related terms from the taxonomy; semantic grounding via definitions; showing the context of the term (i.e. the paragraph surrounding the term).

Although we used most of the above techniques, the experts did not have access to this information: we kept the survey short to elicit more responses. Thus, the survey showed the terms extracted but not the taxonomy: for example 'web page' is a term that was assigned to any discipline by some experts. The taxonomy reveals that it is mainly used as a subtopic of 'information retrieval', meaning we can conclude that it comes from Computer Science.

Overall Reflections

Our use of several different methods to analyse this data allows us to corroborate our results. For example, the partition algorithm identified 9 communities, which on inspection mapped to 4 key communities: information retrieval; personalised learning/elearning; semantic web; social networking. We can see these communities at the Web Science conferences: WebSci'10 clearly included the semantic web and information retrieval communities (its peak terms included those very phrases), while WebSci'11 presumably had stronger presence from the social networking community, with its peak term of 'social media'. We noticed a dearth of peak terms in the WebSci conference series related to the Personalised Learning Environment community, which further suggests that that community arose from the PLE conference included in the corpus at journal.webscience.org.

Our expert survey included terms related to the 4 communities. Of these, information retrieval was uncontroversial and mapped straight to Computer Science. Computer Science was the only discipline named by more than 3 experts in relation to the term 'semantic web', but the term did elicit a total of 7 named disciplines. The terms relating to the remaining two communities, personalised learning environment and social networking, were both controversial, eliciting 10 and 11 named disciplines apiece. We suggest that it is the controversial terms, the ones that elicit many named disciplines, which are the most important to Web Science: although technologies like linked data and information retrieval techniques are clearly necessary to Web Science, they are perhaps tools of Web Science, rather than its heart. By contrast, the controversial terms such as social networking, web community and social interaction are the terms that truly reflect the ethos of Web Science, the goals of understanding and engineering the web's impact on our society – and the impact of societies upon the web.

The WebSci'12 paper to examine this area [12] provided early results that suggested that Computer Science and Sociology were very present within Web Science publications, and that there was no real sign of Economics, Psychology, Philosophy, or Law. Although experts in our review did name all of these areas in relation to top-ranked terms, psychology was the only one of these to be named by at least three experts. Other disciplines associated with terms by at least three experts were -- in addition to the Web Science stalwarts of Computer Science and Sociology -- Communication, Design, Education, Network Science and Pedagogy. Perhaps Web Science is growing after all.

CONCLUSIONS

We have used NLP and data processing tools to make sense of a corpus of Web Science publications. We discovered four key communities (information retrieval; personalised learning; semantic web; social networking) and examined trends within the Web Science conference. We first looked at changes over time, revealing a strong difference in focus between 2010 and 2011 that reflected the collocated conferences of those years, and confirmed Hypothesis 1: co-location with other events influences what disciplines are present at WebSci. We also examined trends in the Web Science conference by format (papers versus posters), and found that although posters covered broader topics (83% of the identified 1000 topics, compared to 70% by papers), confirming Hypothesis 2: the distribution of disciplines represented by posters (as measured by terms) is broader than that of disciplines represented by papers. The formats were not, however, very different in terms of the height of peak terms (when a term was noticeably more present in one year than another).

The expert survey was conducted to gain insight into the relationship between terms and disciplines. It was clear that some terms, such as information retrieval, were closely linked with specific disciplines, but that other terms, such as social media, were not. Computer Science and Sociology were disciplines that the experts often cited, reflecting their foundational role in Web Science; other disciplines to be named frequently were Psychology and Communication. Expert responses to the lists of highly rated terms and disciplines were controversial, with claims made that these lists were biased towards technology and excluded the arts and humanities. We hope that this disparity between expert expectations of Web Science and actual Web Science materials will provide food for thought for future Web Science organising committees.

ACKNOWLEDGEMENTS

With grateful thanks to Kieron O'Hara for allowing us to draw on his currently unpublished encyclopedia article, Allison Schaap for her assistance with MatLab, and the experts who were kind enough to participate in the survey. This work has been funded in part by the European Union under Grant No. 258191 for the PROMISE project, as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

REFERENCES

1. Barthélémy M. 2004. Betweenness centrality in large complex networks, *The European Physical Journal B - Condensed Matter and Complex Systems*, 38,4, 163-168
2. Berners-Lee, T., Hall, W., Hendler, J.A., Shadbolt, N. & Weitzner, D.J. (2006). 'Creating a Science of the Web'. Science, 313/5788: 769-771
3. Bizer, C., Heath, T. & Berners-Lee, T. (2009). 'Linked Data – the Story so Far'. *International Journal on Semantic Web and Information Systems*, 5/3: 1-22.
4. Blondel V, Guillaume J, Lambiotte R, Mech E (2008) Fast unfolding of communities in large net- works. J Stat Mech: Theory Exp 2008:P10008
5. Börner, K., Chen, C. and Boyack, K. W. (2003), Visualizing knowledge domains. Ann. Rev. Info. Sci. Tech., 37: 179–255. doi: 10.1002/aris.1440370106
6. Georgeta Bordea and Paul Buitelaar. 2010. DERIUNLP: A context based approach to automatic keyphrase extraction. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 146-149.
7. Callon, M., Cortial, J.P. Turner, W.A., Bauin, S. From Translations To Problematic Networks – An Introduction To Co-Word Analysis, Social Science Information Sur Les Sciences Socials 22(2) (1983), 191–235.
8. Chen, C., Carr, L. 1999. Trailblazing the literature of hypertext: author co-citation analysis (1989–1998), in Proc. 10th ACM Conference on Hypertext and Hypermedia, 51-60.
9. Coulter, N., Monarch, I. and Konda, S. (1998), Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49: 1206–1223.
10. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. & Weitzner, D. (2008) ‚Web Science: An Interdisciplinary Approach to Understanding the Web', Communications of the ACM, 51(7), 60-69.
11. Henry, N., Goodell, H., Elmqvit, N., Fekete, J. 2007. 20 Years of 4 HCI Conferences: A Visual Exploration. *International Journal of Human Computer Interaction - Reflections on Human-Computer Interaction*, 23(3), 239-285.
12. Hooper, C. J., Marie, N., Kalampokis, E., *Dissecting the Butterfly: Representation of Disciplines Publishing at the Web Science Conference Series*, Proc. WebSci 2012, ACM Press (2012), 137-140.
13. Hu, Y. F. "Efficient, High-Quality Force-Directed Graph Drawing." The Mathematica Journal 10, no. 1 (2006): 37-71
14. Lopez-Herrera, A. G., Cobo, M. J., Herrera-Viedma, E., & Herrera, F. (2010). A bibliometric study about the research based on hybridating the fuzzy logic field and the other computational intelligent techniques: A visual approach. International Journal of Hybrid Intelligent Systems, 17, 17–32.
15. Monaghan, F., Bordea, G., Samp, K., Buitelaar, P. 2010 Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food, in *Semantic Web Challenge at the International Semantic Web Conference.*
16. Nagel, T, Duval, E., Heidmann, F. (2011). Exploring the Geospatial Network of Scientific Collaboration on a Multitouch Table, in Proc. 22nd ACM Conference on Hypertext and Hypermedia (demo).
17. Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three (IJCAI'11), Toby Walsh (Ed.), Vol. Volume Three. AAAI Press 1872-1877.
18. Shadbolt, N., *What Is Web Science?* Talk, webscience.org/webscience.html
19. Shadbolt, N., Hall, W. & Berners-Lee, T. (2006). 'The Semantic Web Revisited'. *IEEE Intelligent Systems*, 21/3: 96-101.
20. Shadbolt, N. & Berners-Lee, T. (2008). 'Web Science Emerges', Scientific American, October 2008: 32-37.
21. Shneiderman, B. 2007. Web science: a Provocative Invitation to Computer Science. Commun. ACM 50, 6 (June 2007), 25-27. DOI=10.1145/1247001.1247022
22. Snow, C. P. (1960). *The Two Cultures and the Scientific Revolution: the Rede Lecture 1959*. University Press.
23. Vafopoulos, M. 2010. Web Science Subject Categorization (WSSC), *in Proc. Of the WebSci 2010*.
24. White, H.D. 1998. Visualizing a Discipline: An Author Co- Citation Analysis of Information Science, 1972–1995, *Journal of the American Society for Information Science*, 49, 4, 327–355.
25. Zhong-Yi Wang, Gang Li, Chun-Ya Li, Ang Li (2012). Research on the semantic-based co-word analysis. Scientometrics, 90, 855-875. doi: 10.1007/s11192-011-0563-y