

Report on the  
“Future of Research Communications”  
Workshop  
Dagstuhl, August 15-18, 2012

Ivan Herman  
and the other workshop co-chairs:  
Tim Clark, Eduard Hovy, and Anita de Waard

Let me tell you a story...

I heard about an interesting paper  
on a social site

# So I looked up the journal on the Web

Journal of Web Semantics - Elsevier

http://www.journals.elsevier.com/journal-of-web-semantics/

2012 My Mercurial My Feedly TR LocalData Private Mailing lists Social SW Python RDFa it! Bookmarklets

**ELSEVIER** Type here to search on Elsevier.com **Advanced Product Search**

**Books & journals** **Online tools** **Authors, editors & reviewers** **About Elsevier** **Help**

**Journal of Web Semantics**  
**Science, Services and Agents on the World Wide Web**

The Journal of Web Semantics is an interdisciplinary journal based on research and applications of various subject areas that contribute to the development of a knowledge-intensive and intelligent service...

[View full aims and scope](#)

**Editors-in-Chief:** T. Finin, S. Staab  
[View full editorial board](#)

[Guide for Authors](#)  
[Submit Your Paper](#)  
[Track Your Paper](#)  
[Order Journal](#)  
[Access Full Text](#)

**Impact Factor:** 2.789  
5-Year Impact Factor: 3.593  
Imprint: ELSEVIER  
ISSN: 1570-8268

**Stay up-to-date**  
Register your interests and receive email alerts tailored to your needs  
[Click here to sign up](#)

**Most Downloaded Articles** **ScienceDirect**

- Relevance feedback between hypertext and Semantic Web search: Frameworks and evaluation**  
Harry Halpin | Victor Lavrenko
- Semantic Web MiningState of the art and future directions**  
Gerd Stumme | Andreas Hotho | ...
- An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size**  
Md. Hanif Seddiqui | Masaki Aono

[VIEW ALL](#)

**Most Cited Articles** **Scopus**

- Pellet: A practical OWL-DL reasoner**  
Sirin, E. | Parsia, B. | ...
- Ontologies are us: A unified model of social networks and semantics**  
Mika, P.
- A survey of trust in computer science and the Semantic Web**  
Artz, D. | Gil, Y.

[VIEW ALL](#)

# Found the paper

The screenshot shows a web browser window displaying a ScienceDirect article. The browser's address bar shows the URL: <http://www.sciencedirect.com/science/article/pii/S1570826811000473>. The page header includes the SciVerse and ScienceDirect logos, navigation links (Home, Browse, Search, My settings, My alerts, Shopping cart), and user options (Register, Login, Go to). A search bar is visible with the text "Search ScienceDirect".

The article title is "Web Semantics: Science, Services and Agents on the World Wide Web", Volume 9, Issue 4, December 2011, Pages 365–401. It is a JWS special issue on Semantic Search. The Elsevier logo is present on the left.

The main article title is "Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine". The authors listed are Aidan Hogan<sup>a</sup>, Andreas Harth<sup>b</sup>, Jürgen Umbrich<sup>a</sup>, Sheila Kinsella<sup>a</sup>, Axel Polleres<sup>a</sup>, and Stefan Decker<sup>a</sup>. The affiliations are:  
<sup>a</sup> Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland  
<sup>b</sup> AIFB, Karlsruhe Institute of Technology, Germany

The article is available online as of 22 June 2011. The DOI is <http://dx.doi.org/10.1016/j.websem.2011.06.004>. There are links for "How to Cite or Link Using DOI" and "Permissions & Reprints". A "View full text" button is visible.

On the right side, there are two sections: "Related articles" and "Related reference work articles".

**Related articles:**

- Relevance feedback between hypertext Sem...  
*Web Semantics: Science, Services an*
- Sig.ma: Live views on the Web of Data  
*Web Semantics: Science, Services an*
- Exploring the Geospatial Semantic Wet  
*Web Semantics: Science, Services an*
- Chapter 3 - RDF—The basis of the Ser  
*Semantic Web for the Working Ontolog*
- Scalable and distributed methods for er  
*Web Semantics: Science, Services an*

**Related reference work articles (e.g. encyclopedias):**

- Web Searching  
*Encyclopedia of Language & Linguistic*
- Internet  
*Encyclopedia of Ecology*
- Library Applications  
*Encyclopedia of Information Systems*

# However, I was not at my institute...

Searching and browsing Linked Data with SWSE: The Semantic ...vices and Agents on the World Wide Web | ScienceDirect.com

http://www.sciencedirect.com/science/article/pii/S1570826811000473

Back to results Export citation Purchase More options...

Stefan Decker<sup>a, \*</sup>

<sup>a</sup> Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

<sup>b</sup> AIFB, Karlsruhe Institute of Technology, Germany

Available online 22 June 2011.

<http://dx.doi.org/10.1016/j.websem.2011.06.004>, How to Cite or Link Using DOI

Permissions & Reprints

**View full text**

**Purchase**

**Related reference work articles**  
e.g. encyclopedias

- Web Searching  
*Encyclopedia of Language & Linguistics*
- Internet  
*Encyclopedia of Ecology*
- Library Applications  
*Encyclopedia of Information Systems*

More related reference work articles

**Abstract**

In this paper, we discuss the architecture and implementation of the Semantic Web Search Engine (SWSE). Following traditional search engine architecture, SWSE consists of crawling, data enhancing, indexing and a user interface for search, browsing and retrieval of information; unlike traditional search engines, SWSE operates over RDF Web data – loosely also known as Linked Data – which implies unique challenges for the system design, architecture, algorithms, implementation and user interface. In particular, many challenges exist in adopting Semantic Web technologies for Web data: the unique challenges of the Web – in terms of scale, unreliability, inconsistency and noise – are largely overlooked by the current Semantic Web standards. Herein, we describe the current SWSE system, initially detailing the architecture and later elaborating upon the function, design, implementation and performance of each individual component. In so doing, we also give an insight into how current Semantic Web standards can be tailored, in a best-effort manner, for use on Web data. Throughout, we offer evaluation and complementary argumentation to support our design choices, and also offer discussion on future directions and open research questions. Later, we

# But... I knew the secret!

The screenshot shows a web browser window titled "About the Preprint Server" with the URL <http://www.websemanticsjournal.org/index.php/ps/about>. The browser's address bar includes a "Reader" button and navigation icons. The website header features the logo for the Journal of Web Semantics and the text "Journal of Web Semantics: PREPRINT SERVER". Below the header, a navigation menu includes links for HOME, ABOUT, LOG IN, REGISTER, SEARCH, CURRENT, ARCHIVES, and ANNOUNCEMENTS. The main content area is titled "Home > About the Preprint Server" and contains a paragraph describing the server's purpose, followed by a list of links: Aims & Scope, People, Privacy Statement, Submissions, Sponsorship, Site Map, and About this Publishing System. A section titled "AIMS & SCOPE" provides a detailed description of the journal's interdisciplinary focus. The right sidebar contains a "JOURNAL CONTENT" section with a search box and a "Browse" section with links for "By Issue", "By Author", and "By Title". Below this is a "JWS BLOG" section with several article links. At the bottom of the sidebar, there is an "AUTHOR" section with links for "Aims & Scope", "Guide for Authors", and "Submit an Article", and an "INFORMATION" section.

Go to "<http://www.websemanticsjournal.org/index.php/ps/about>"

# But... I knew the secret!

Searching and Browsing Linked Data with SWSE: the Semantic ...antics: Science, Services and Agents on the World Wide Web

http://websemanticsjournal.org/index.php/ps/article/view/240

2012 My Mercurial My Feedly TR LocalData Private Mailing lists Social SW Python RDFa it! Bookmarks

## Journal of Web Semantics: PREPRINT SERVER

The Preprint Server provides readers with free electronic access to article preprints of the Journal of Web Semantics: Science, Services and Agents on the World Wide Web at Elsevier.

USER

Username

Password

Remember me

Log In

HOME ABOUT LOG IN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS

Home > Vol 9, No 4 (2011) > Hogan

### SEARCHING AND BROWSING LINKED DATA WITH SWSE: THE SEMANTIC WEB SEARCH ENGINE

Aidan Hogan, Andreas Harth, Juergen Umrich, Sheila Kinsella, Axel Polleres, Stefan Decker

#### ABSTRACT

Abstract: In this paper, we discuss the architecture and implementation of the Semantic Web Search Engine (SWSE). Following traditional search engine architecture, SWSE consists of crawling, data enhancing, indexing and a user interface for search, browsing and retrieval of information; unlike traditional search engines, SWSE operates over RDF Web data -- loosely also known as Linked Data -- which implies unique challenges for the system design, architecture, algorithms, implementation and user interface. In particular, many challenges exist in adopting Semantic Web technologies for Web data: the unique challenges of the Web -- in terms of scale, unreliability, inconsistency and noise -- are largely overlooked by the current Semantic Web standards. Herein, we describe the current SWSE system, initially detailing the architecture and later elaborating upon the function, design, implementation and performance of each individual component. In so doing, we also give an insight into how current Semantic Web standards can be tailored, in a besteffort manner, for use on Web data. Throughout, we offer evaluation and complementary argumentation to support our design choices, and also offer discussion on future directions and open research questions. Later, we also provide candid discussion relating to the difficulties currently faced in bringing such a search engine into the mainstream, and lessons learnt from roughly six years working on the Semantic Web Search Engine project.

Full Text: [PDF](#)

ARTICLE TOOLS

- Print version
- Indexing metadata
- How to cite item
- Email this article

(Login required)

JOURNAL CONTENT

Search

All

Search

Browse

- By Issue
- By Author
- By Title

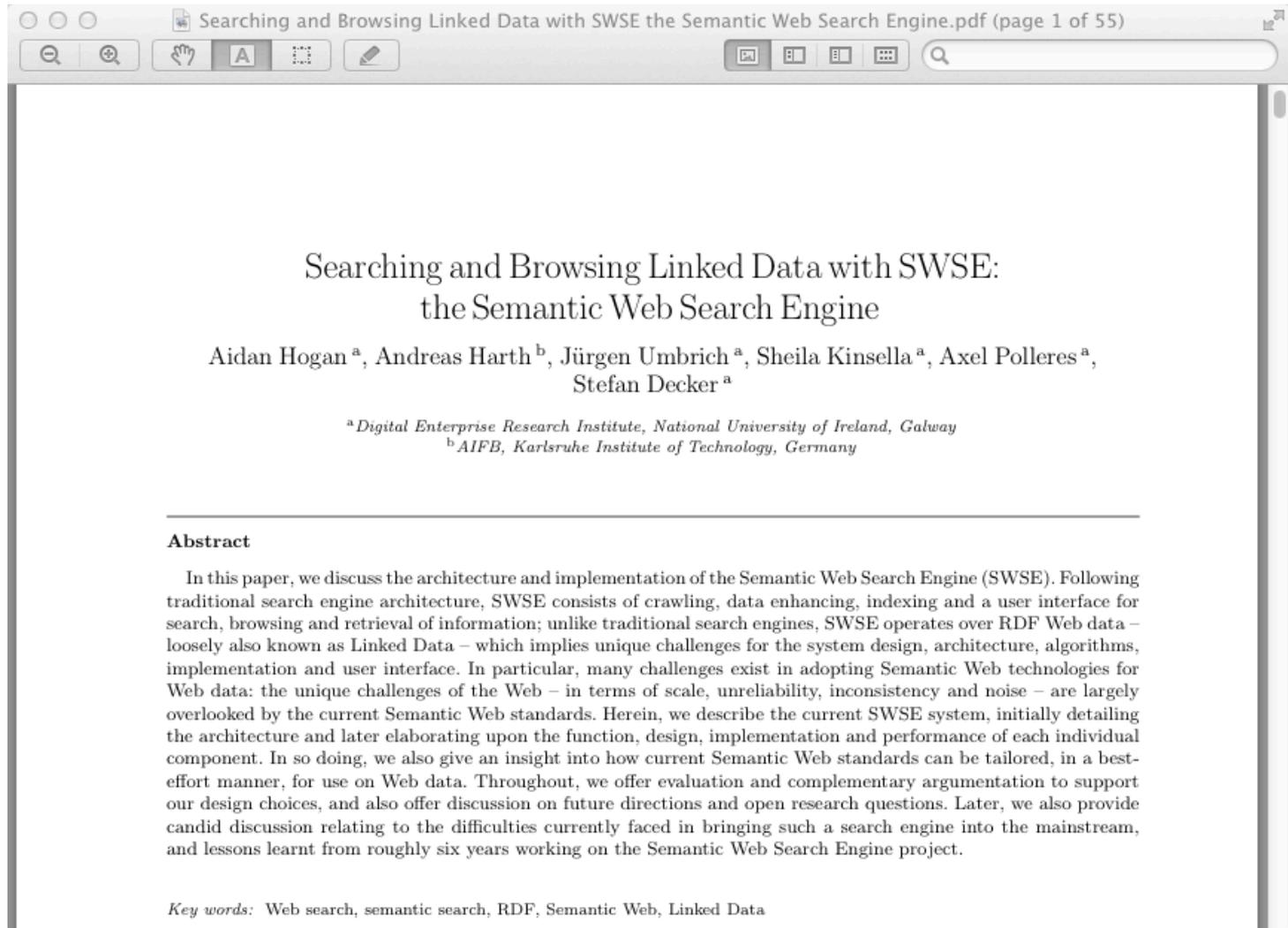
Search all

- Research Papers
- Survey Papers
- Ontology Papers
- System Papers

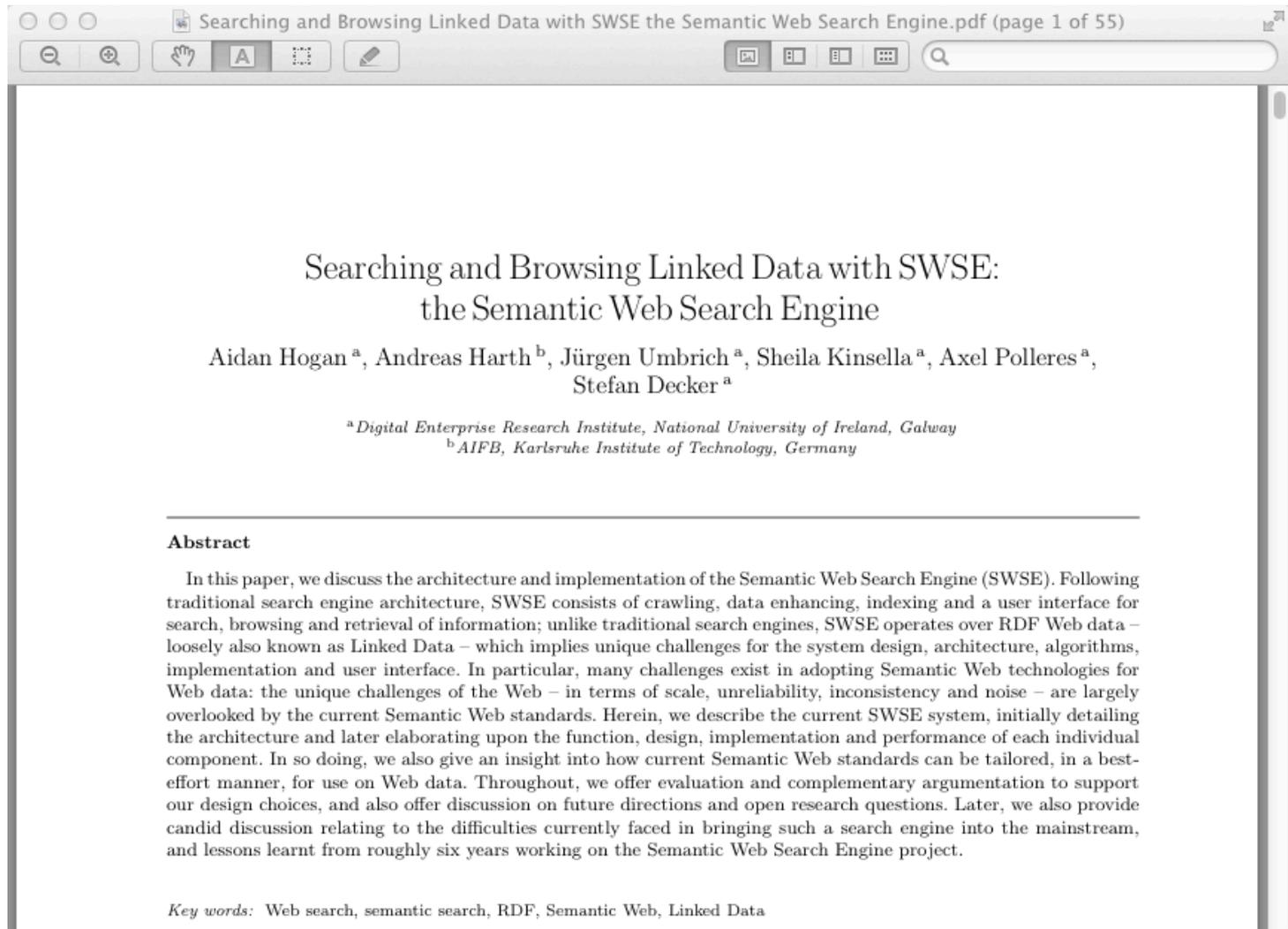
JWS BLOG

- JWS special issue: reasoning with...
- Journal of Web Semantics, volume 11...
- JWS special issue on scalability. v10...

# So I could read the paper.



# The paper also had...



# ... (low resolution) diagrams, and ...

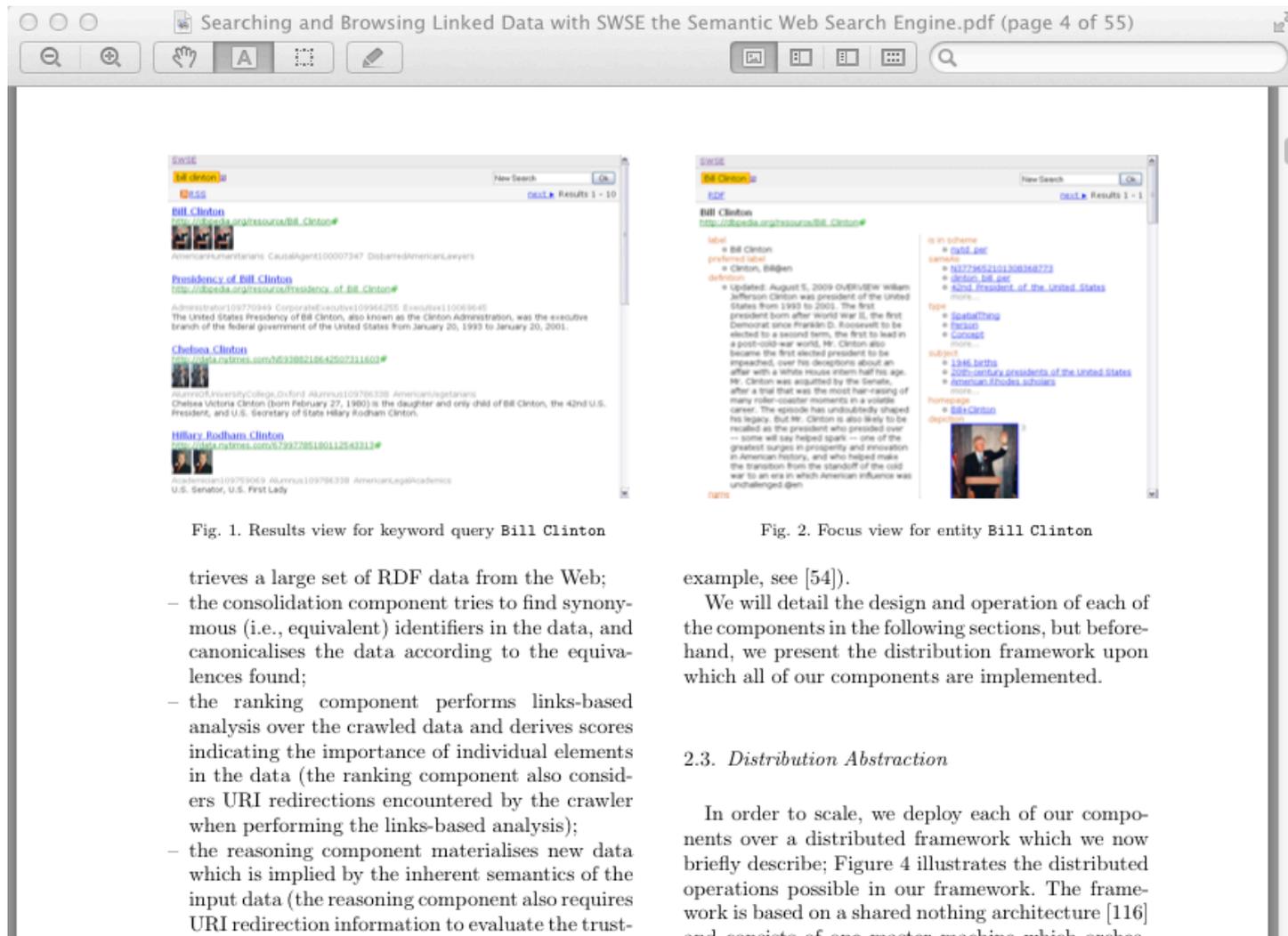


Fig. 1. Results view for keyword query Bill Clinton

Fig. 2. Focus view for entity Bill Clinton

trieves a large set of RDF data from the Web;

- the consolidation component tries to find synonymous (i.e., equivalent) identifiers in the data, and canonicalises the data according to the equivalences found;
- the ranking component performs links-based analysis over the crawled data and derives scores indicating the importance of individual elements in the data (the ranking component also considers URI redirections encountered by the crawler when performing the links-based analysis);
- the reasoning component materialises new data which is implied by the inherent semantics of the input data (the reasoning component also requires URI redirection information to evaluate the trust-

example, see [54]).

We will detail the design and operation of each of the components in the following sections, but beforehand, we present the distribution framework upon which all of our components are implemented.

### 2.3. Distribution Abstraction

In order to scale, we deploy each of our components over a distributed framework which we now briefly describe; Figure 4 illustrates the distributed operations possible in our framework. The framework is based on a shared nothing architecture [116] and consists of one master machine which orches-

# ... algorithms to read and to understand ...

– **Scale:** The crawler should employ scalable techniques, and on-disk indexing as required.

– **Quality:** The crawler should prioritise crawling URIs it considers to be “high quality”.

Thus, the design of our crawler is inspired by related work from traditional HTML crawlers. Additionally – and specific to crawling structured data – we identify the following requirement:

– **Structured Data:** The crawler should retrieve a high percentage of RDF/XML documents and avoid wasted lookups on unwanted formats: e.g., HTML documents.

Currently, we crawl for RDF/XML syntax documents – RDF/XML is still the most commonly used syntax for publishing RDF on the Web, and we plan in future to extend the crawler to support other formats such as RDFa, N-Triples and Turtle.

The following algorithm details the operation of the crawler, and will be explained in detail throughout this section.

### 5.1. High-level Approach

Our high-level approach is to perform breath-first crawling, following precedent set by traditional Web crawlers (cf. [15] [69]): the crawl is conducted in rounds, with each round crawling a *frontier*. On a high-level, Algorithm 1 represents this round-based approach applying ROUNDS number of rounds. The frontier comprises of seed URIs for round 0 (Algorithm 1, Line 1), and thereafter with novel URIs extracted from documents crawled in the previous

---

**Algorithm 1** Algorithm for crawling

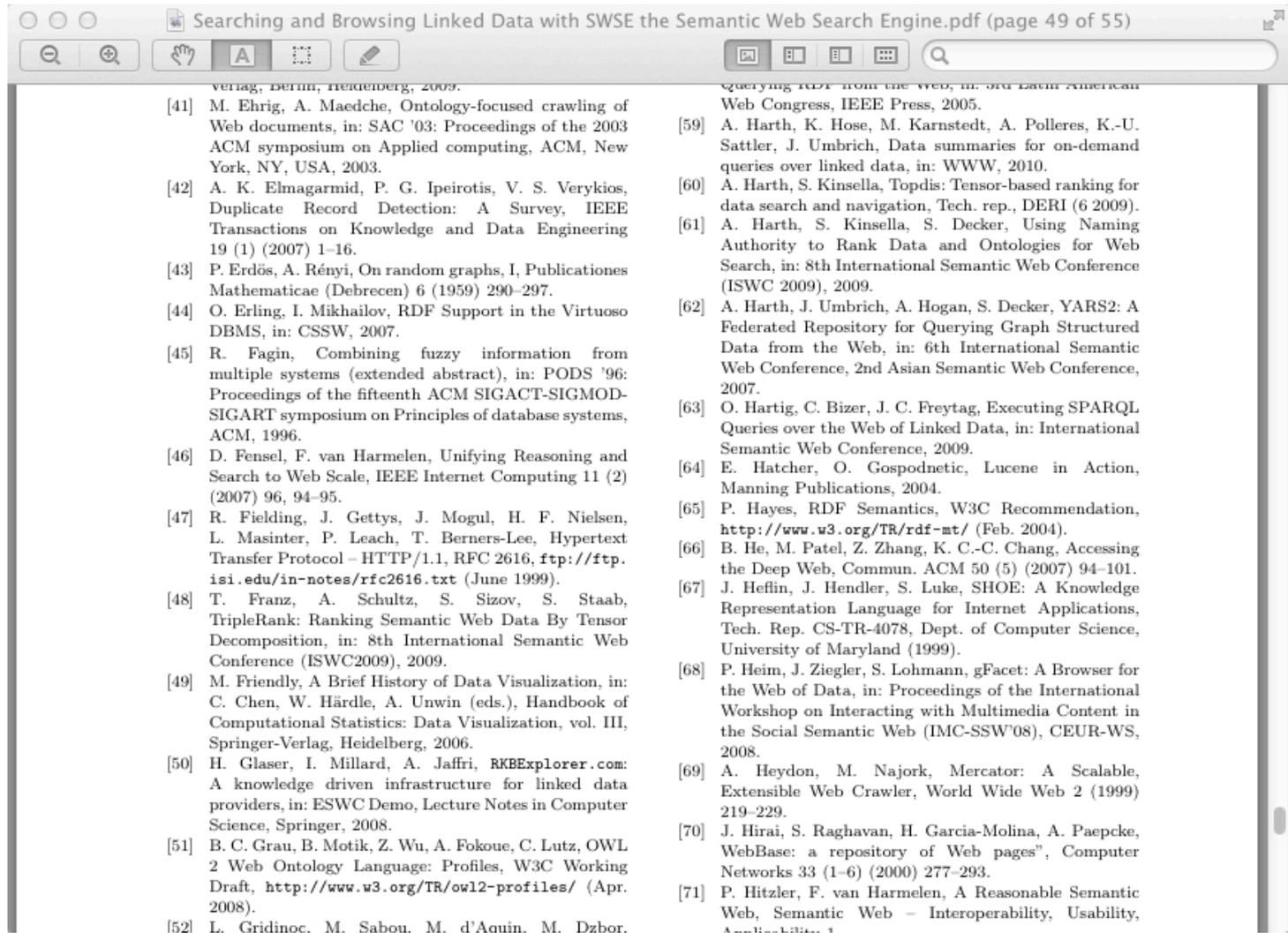
---

**Require:** SEEDS, ROUNDS, PLD-LIMIT, MIN-DELAY

```
1: frontier ← SEEDS
2: pld0...n ← new queue
3: stats ← new stats
4: while rounds + 1 < ROUNDS do
5:   put frontier into pld0...n
6:   while depth + 1 < PLD-LIMIT do
7:     for i = 0 to n do
8:       prioritise(pldi, stats)
9:     end for
10:    start ← current_time()
11:    for i = 0 to n do
12:      curi = calculate_cur(pldi, stats)
13:      if curi > random([0,1]) then
14:        get uri from pldi
15:        urideref = deref(uri)
16:        if urideref = uri then
17:           $\mathcal{G}$  = get(uri)
18:          output  $\mathcal{G}$ 
19:           $\overline{U}_{\mathcal{G}}$  ← URIs in  $\mathcal{G}$ 
20:           $\overline{U}_{\mathcal{G}}$  ← prune blacklisted from  $U_{\mathcal{G}}$ 
21:          add unseen URIs in  $\overline{U}_{\mathcal{G}}$  to frontier
22:          update stats wrt.  $\overline{U}_{\mathcal{G}}$ 
23:        else
24:          if urideref is unseen then
25:            add urideref to frontier
26:            update stats for urideref
27:          end if
28:        end if
29:      end if
30:    end for
```



# ... lots of references.



venag, Berni, Heidelberg, 2009.

[41] M. Ehrig, A. Maedche, Ontology-focused crawling of Web documents, in: SAC '03: Proceedings of the 2003 ACM symposium on Applied computing, ACM, New York, NY, USA, 2003.

[42] A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering 19 (1) (2007) 1–16.

[43] P. Erdős, A. Rényi, On random graphs, I, Publicationes Mathematicae (Debrecen) 6 (1959) 290–297.

[44] O. Erling, I. Mikhailov, RDF Support in the Virtuoso DBMS, in: CSSW, 2007.

[45] R. Fagin, Combining fuzzy information from multiple systems (extended abstract), in: PODS '96: Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, ACM, 1996.

[46] D. Fensel, F. van Harmelen, Unifying Reasoning and Search to Web Scale, IEEE Internet Computing 11 (2) (2007) 96, 94–95.

[47] R. Fielding, J. Gettys, J. Mogul, H. F. Nielsen, L. Masinter, P. Leach, T. Berners-Lee, Hypertext Transfer Protocol – HTTP/1.1, RFC 2616, <ftp://ftp.isi.edu/in-notes/rfc2616.txt> (June 1999).

[48] T. Franz, A. Schultz, S. Sizov, S. Staab, TripleRank: Ranking Semantic Web Data By Tensor Decomposition, in: 8th International Semantic Web Conference (ISWC2009), 2009.

[49] M. Friendly, A Brief History of Data Visualization, in: C. Chen, W. Härdle, A. Unwin (eds.), Handbook of Computational Statistics: Data Visualization, vol. III, Springer-Verlag, Heidelberg, 2006.

[50] H. Glaser, I. Millard, A. Jaffri, RKBExplorer.com: A knowledge driven infrastructure for linked data providers, in: ESWC Demo, Lecture Notes in Computer Science, Springer, 2008.

[51] B. C. Grau, B. Motik, Z. Wu, A. Fokoue, C. Lutz, OWL 2 Web Ontology Language: Profiles, W3C Working Draft, <http://www.w3.org/TR/owl2-profiles/> (Apr. 2008).

[52] L. Grdinoc. M. Sabou. M. d'Aquin. M. Dzbor. Querying RDF from the web, in: 3rd Latin American Web Congress, IEEE Press, 2005.

[59] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, J. Umbrich, Data summaries for on-demand queries over linked data, in: WWW, 2010.

[60] A. Harth, S. Kinsella, Topdis: Tensor-based ranking for data search and navigation, Tech. rep., DERI (6 2009).

[61] A. Harth, S. Kinsella, S. Decker, Using Naming Authority to Rank Data and Ontologies for Web Search, in: 8th International Semantic Web Conference (ISWC 2009), 2009.

[62] A. Harth, J. Umbrich, A. Hogan, S. Decker, YARS2: A Federated Repository for Querying Graph Structured Data from the Web, in: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, 2007.

[63] O. Hartig, C. Bizer, J. C. Freytag, Executing SPARQL Queries over the Web of Linked Data, in: International Semantic Web Conference, 2009.

[64] E. Hatcher, O. Gospodnetic, Lucene in Action, Manning Publications, 2004.

[65] P. Hayes, RDF Semantics, W3C Recommendation, <http://www.w3.org/TR/rdf-mt/> (Feb. 2004).

[66] B. He, M. Patel, Z. Zhang, K. C.-C. Chang, Accessing the Deep Web, Commun. ACM 50 (5) (2007) 94–101.

[67] J. Heflin, J. Hendler, S. Luke, SHOE: A Knowledge Representation Language for Internet Applications, Tech. Rep. CS-TR-4078, Dept. of Computer Science, University of Maryland (1999).

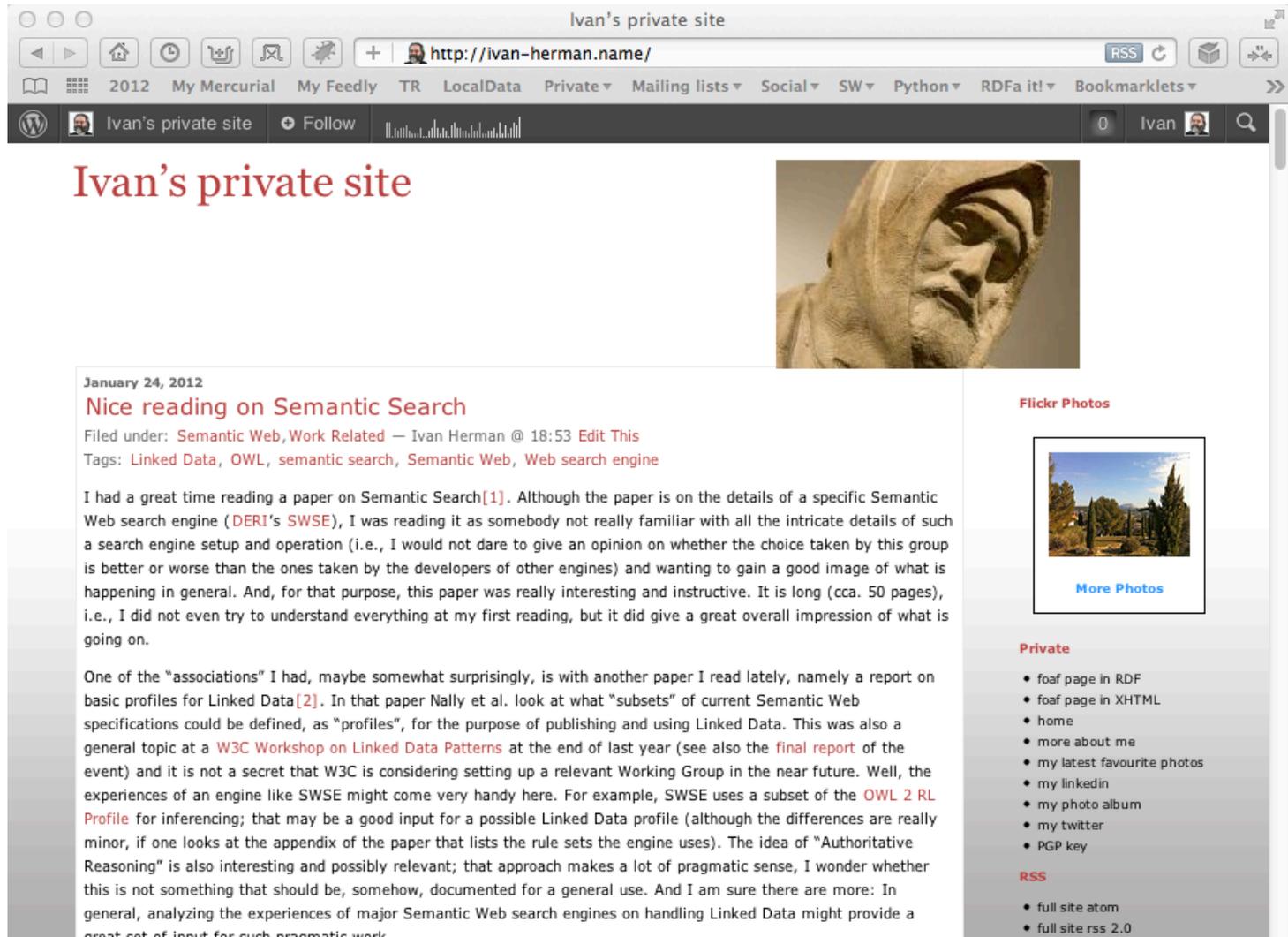
[68] P. Heim, J. Ziegler, S. Lohmann, gFacet: A Browser for the Web of Data, in: Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08), CEUR-WS, 2008.

[69] A. Heydon, M. Najork, Mercator: A Scalable, Extensible Web Crawler, World Wide Web 2 (1999) 219–229.

[70] J. Hirai, S. Raghavan, H. Garcia-Molina, A. Paepcke, WebBase: a repository of Web pages", Computer Networks 33 (1–6) (2000) 277–293.

[71] P. Hitzler, F. van Harmelen, A Reasonable Semantic Web, Semantic Web – Interoperability, Usability, Applications 1 (2006) 1–16.

# I liked the paper. So I wrote a blog...



The screenshot shows a web browser window titled "Ivan's private site" with the URL "http://ivan-herman.name/". The browser's address bar and navigation icons are visible. Below the browser, the blog post is displayed. The title is "Ivan's private site" in a large, red font. To the right of the title is a large, square image of a person's face, possibly a statue or a portrait. Below the title and image, the post is dated "January 24, 2012" and has the title "Nice reading on Semantic Search". The post content discusses a paper on Semantic Search and mentions a workshop on Linked Data Patterns. To the right of the main text, there is a sidebar with a "Flickr Photos" section showing a small image of a landscape and a "More Photos" link. Below that, there is a "Private" section with a list of links to various resources, and an "RSS" section with links to "full site atom" and "full site rss 2.0".

Ivan's private site

January 24, 2012

## Nice reading on Semantic Search

Filed under: [Semantic Web](#), [Work Related](#) — Ivan Herman @ 18:53 [Edit This](#)  
Tags: [Linked Data](#), [OWL](#), [semantic search](#), [Semantic Web](#), [Web search engine](#)

I had a great time reading a paper on Semantic Search[1]. Although the paper is on the details of a specific Semantic Web search engine (DERI's SWSE), I was reading it as somebody not really familiar with all the intricate details of such a search engine setup and operation (i.e., I would not dare to give an opinion on whether the choice taken by this group is better or worse than the ones taken by the developers of other engines) and wanting to gain a good image of what is happening in general. And, for that purpose, this paper was really interesting and instructive. It is long (cca. 50 pages), i.e., I did not even try to understand everything at my first reading, but it did give a great overall impression of what is going on.

One of the "associations" I had, maybe somewhat surprisingly, is with another paper I read lately, namely a report on basic profiles for Linked Data[2]. In that paper Nally et al. look at what "subsets" of current Semantic Web specifications could be defined, as "profiles", for the purpose of publishing and using Linked Data. This was also a general topic at a [W3C Workshop on Linked Data Patterns](#) at the end of last year (see also the [final report](#) of the event) and it is not a secret that W3C is considering setting up a relevant Working Group in the near future. Well, the experiences of an engine like SWSE might come very handy here. For example, SWSE uses a subset of the [OWL 2 RL Profile](#) for inferencing; that may be a good input for a possible Linked Data profile (although the differences are really minor, if one looks at the appendix of the paper that lists the rule sets the engine uses). The idea of "Authoritative Reasoning" is also interesting and possibly relevant; that approach makes a lot of pragmatic sense, I wonder whether this is not something that should be, somehow, documented for a general use. And I am sure there are more: In general, analyzing the experiences of major Semantic Web search engines on handling Linked Data might provide a great set of input for such pragmatic work.

**Flickr Photos**



[More Photos](#)

**Private**

- foaf page in RDF
- foaf page in XHTML
- home
- more about me
- my latest favourite photos
- my linkedin
- my photo album
- my twitter
- PGP key

**RSS**

- full site atom
- full site rss 2.0

# ...with exact references.

The screenshot shows a web browser window titled "Ivan's private site" with the URL <http://ivan-herman.name/>. The browser's address bar and navigation icons are visible at the top. Below the browser window, the content of the blog post is displayed. The main text discusses "associations" and "profiles" in the context of Linked Data, mentioning a W3C workshop and a report by Nally et al. The text also discusses the use of `owl:sameAs` and SKOS properties like `closetMatch`, `exactMatch`, and `broadMatch`. A red circle highlights a list of references at the bottom of the post. To the right of the main text, there is a sidebar with sections for "Private", "RSS", "Work Related", "Extra pages", "Categories", and "Archives per date".

going on.

One of the "associations" I had, maybe somewhat surprisingly, is with another paper I read lately, namely a report on basic profiles for Linked Data [2]. In that paper Nally et al. look at what "subsets" of current Semantic Web specifications could be defined, as "profiles", for the purpose of publishing and using Linked Data. This was also a general topic at a [W3C Workshop on Linked Data Patterns](#) at the end of last year (see also the [final report](#) of the event) and it is not a secret that W3C is considering setting up a relevant Working Group in the near future. Well, the experiences of an engine like SWSE might come very handy here. For example, SWSE uses a subset of the [OWL 2 RL Profile](#) for inferencing; that may be a good input for a possible Linked Data profile (although the differences are really minor, if one looks at the appendix of the paper that lists the rule sets the engine uses). The idea of "Authoritative Reasoning" is also interesting and possibly relevant; that approach makes a lot of pragmatic sense, I wonder whether this is not something that should be, somehow, documented for a general use. And I am sure there are more: In general, analyzing the experiences of major Semantic Web search engines on handling Linked Data might provide a great set of input for such pragmatic work.

I was also wondering about a very different issue. A great deal of work had to be done in SWSE on the proper handling of `owl:sameAs`. On the other hand, one of the recurring discussions on various mailing list and elsewhere is on whether the usage of this property is semantically o.k. or not (see, e.g., [3]). A possible alternative would be to define (beyond `owl:sameAs`) a set of properties borrowed from the [SKOS Recommendation](#), like `closetMatch`, `exactMatch`, `broadMatch`, etc. It is almost trivial to generalize these SKOS properties for the general case but, reading this paper, I was wondering: what effect would such predicates have on search? Would it make it more complicated or, in fact, would such predicates make the life of search engines easier by providing "hints" that could be used for the user interface? Or both? Or is it already too late, because the ubiquitous usage of `owl:sameAs` is already so prevalent that it is not worth touching that stuff? I do not have a clear answer at this moment...

Thanks to the authors!

1. A. Hogan, et al., "Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine", *Journal of Web Semantics*, vol. 4, no. December, pp. 365-401, 2011.
2. M. Nally and S. Speicher, "Toward a Basic Profile for Linked Data", IBM developersWork, 2011.
3. H. Halpin, et al. "When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data", Proceedings of the International Semantic Web Conference, pp. 305-320, 2010

Comments (6)

**Private**

- foaf page in RDF
- foaf page in XHTML
- home
- more about me
- my latest favourite photos
- my linkedin
- my photo album
- my twitter
- PGP key

**RSS**

- full site atom
- full site rss 2.0
- sw related rss 2.0
- work related rss 2.0

**Work Related**

- cv, publ. list, ...
- me at W3C

**Extra pages**

- Some goodies to download

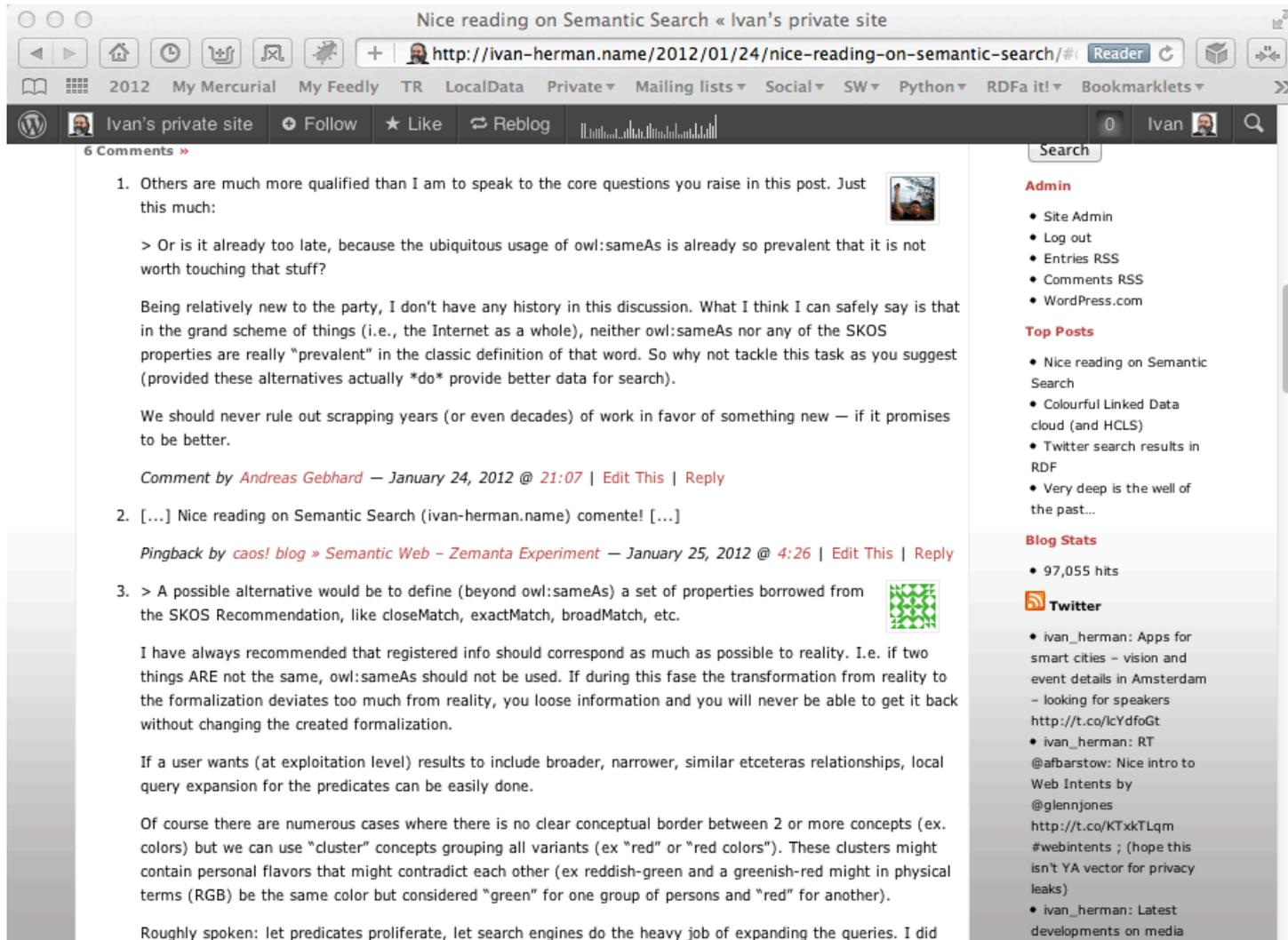
**Categories**

- Private (43)
- General (33)
- Hungary (14)
- Mac (1)
- Work Related (134)
- Code (15)
  - Python (10)
- Semantic Web (120)
- Social aspects (5)

**Archives per date**

Select Month

# There were a bunch of comments...



Nice reading on Semantic Search « Ivan's private site

http://ivan-herman.name/2012/01/24/nice-reading-on-semantic-search/#: Reader

2012 My Mercurial My Feedly TR LocalData Private Mailing lists Social SW Python RDFa it! Bookmarks

Ivan's private site Follow Like Reblog 0 Ivan

6 Comments »

- Others are much more qualified than I am to speak to the core questions you raise in this post. Just this much:  
  
> Or is it already too late, because the ubiquitous usage of owl:sameAs is already so prevalent that it is not worth touching that stuff?  
  
Being relatively new to the party, I don't have any history in this discussion. What I think I can safely say is that in the grand scheme of things (i.e., the Internet as a whole), neither owl:sameAs nor any of the SKOS properties are really "prevalent" in the classic definition of that word. So why not tackle this task as you suggest (provided these alternatives actually \*do\* provide better data for search).  
  
We should never rule out scrapping years (or even decades) of work in favor of something new — if it promises to be better.  
  
*Comment by Andreas Gebhard — January 24, 2012 @ 21:07 | Edit This | Reply*
- [...] Nice reading on Semantic Search (ivan-herman.name) comente! [...]  
  
*Pingback by caos! blog » Semantic Web - Zemanta Experiment — January 25, 2012 @ 4:26 | Edit This | Reply*
- > A possible alternative would be to define (beyond owl:sameAs) a set of properties borrowed from the SKOS Recommendation, like closeMatch, exactMatch, broadMatch, etc.  
  
  
I have always recommended that registered info should correspond as much as possible to reality. I.e. if two things ARE not the same, owl:sameAs should not be used. If during this fase the transformation from reality to the formalization deviates too much from reality, you loose information and you will never be able to get it back without changing the created formalization.  
  
If a user wants (at exploitation level) results to include broader, narrower, similar etceteras relationships, local query expansion for the predicates can be easily done.  
  
Of course there are numerous cases where there is no clear conceptual border between 2 or more concepts (ex. colors) but we can use "cluster" concepts grouping all variants (ex "red" or "red colors"). These clusters might contain personal flavors that might contradict each other (ex reddish-green and a greenish-red might in physical terms (RGB) be the same color but considered "green" for one group of persons and "red" for another).  
  
Roughly spoken: let predicates proliferate, let search engines do the heavy job of expanding the queries. I did

Search

**Admin**

- Site Admin
- Log out
- Entries RSS
- Comments RSS
- WordPress.com

**Top Posts**

- Nice reading on Semantic Search
- Colourful Linked Data cloud (and HCLS)
- Twitter search results in RDF
- Very deep is the well of the past...

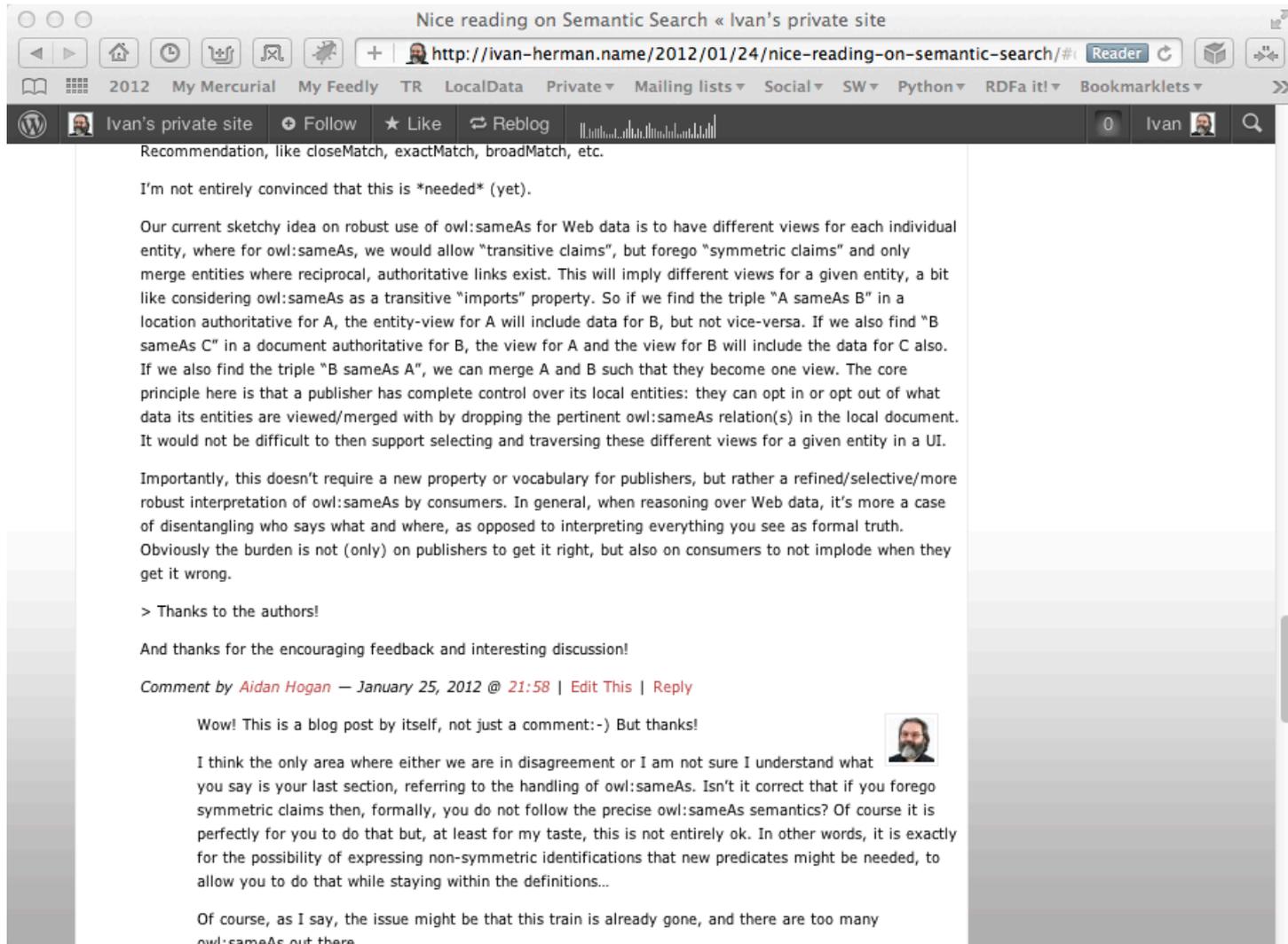
**Blog Stats**

- 97,055 hits

**Twitter**

- ivan\_herman: Apps for smart cities – vision and event details in Amsterdam – looking for speakers  
http://t.co/lcYdfGt
- ivan\_herman: RT @afbarstow: Nice intro to Web Intents by @glennjones  
http://t.co/KTxkTLqm #webintents ; (hope this isn't YA vector for privacy leaks)
- ivan\_herman: Latest developments on media

# ...including the author's, with my answer...



Nice reading on Semantic Search « Ivan's private site

http://ivan-herman.name/2012/01/24/nice-reading-on-semantic-search/# Reader

Ivan's private site Follow Like Reblog

Recommendation, like closeMatch, exactMatch, broadMatch, etc.

I'm not entirely convinced that this is *\*needed\** (yet).

Our current sketchy idea on robust use of owl:sameAs for Web data is to have different views for each individual entity, where for owl:sameAs, we would allow "transitive claims", but forego "symmetric claims" and only merge entities where reciprocal, authoritative links exist. This will imply different views for a given entity, a bit like considering owl:sameAs as a transitive "imports" property. So if we find the triple "A sameAs B" in a location authoritative for A, the entity-view for A will include data for B, but not vice-versa. If we also find "B sameAs C" in a document authoritative for B, the view for A and the view for B will include the data for C also. If we also find the triple "B sameAs A", we can merge A and B such that they become one view. The core principle here is that a publisher has complete control over its local entities: they can opt in or opt out of what data its entities are viewed/merged with by dropping the pertinent owl:sameAs relation(s) in the local document. It would not be difficult to then support selecting and traversing these different views for a given entity in a UI.

Importantly, this doesn't require a new property or vocabulary for publishers, but rather a refined/selective/more robust interpretation of owl:sameAs by consumers. In general, when reasoning over Web data, it's more a case of disentangling who says what and where, as opposed to interpreting everything you see as formal truth. Obviously the burden is not (only) on publishers to get it right, but also on consumers to not implode when they get it wrong.

> Thanks to the authors!

And thanks for the encouraging feedback and interesting discussion!

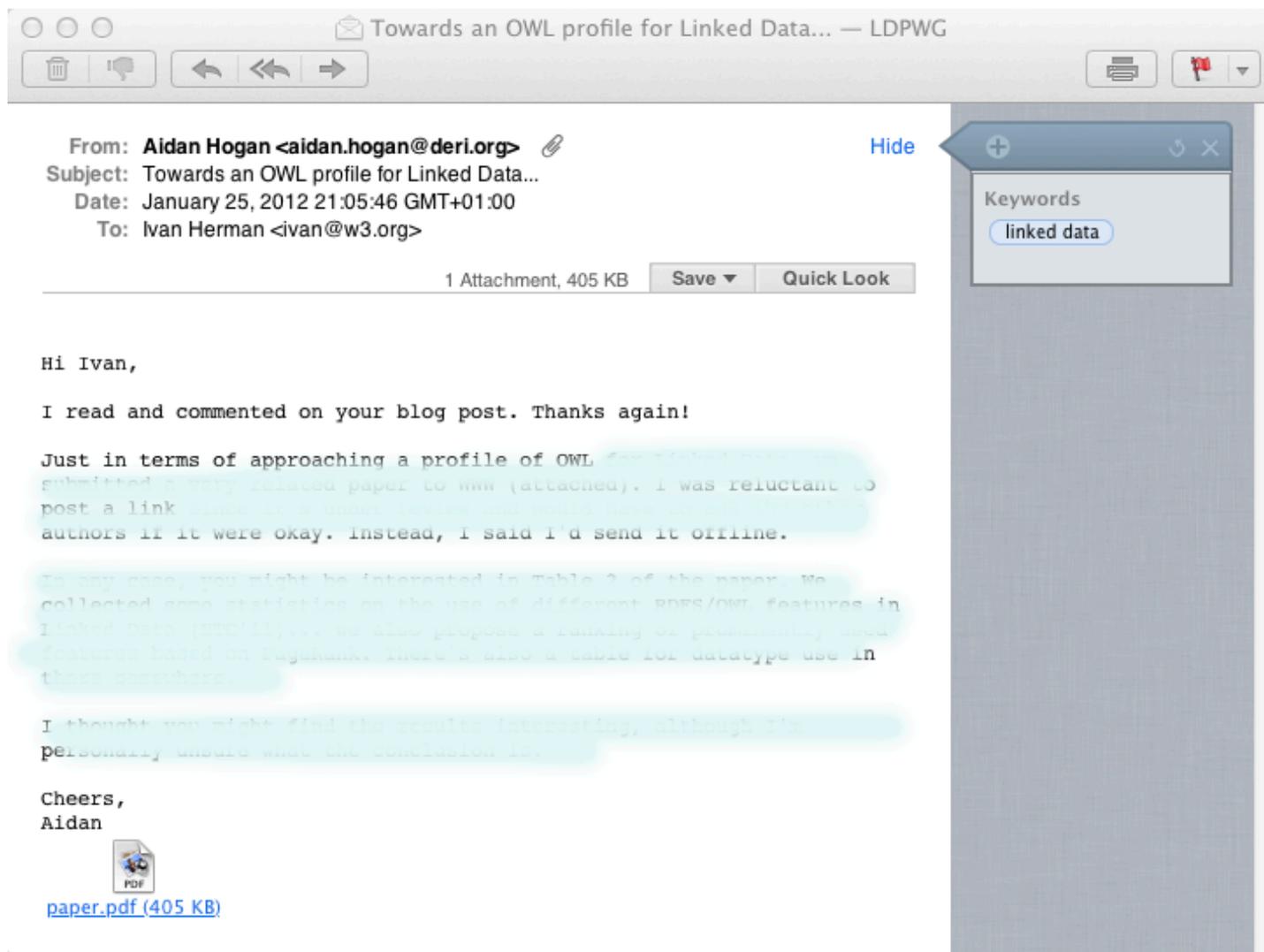
Comment by [Aidan Hogan](#) — January 25, 2012 @ 21:58 | [Edit This](#) | [Reply](#)

Wow! This is a blog post by itself, not just a comment:-) But thanks!

I think the only area where either we are in disagreement or I am not sure I understand what you say is your last section, referring to the handling of owl:sameAs. Isn't it correct that if you forego symmetric claims then, formally, you do not follow the precise owl:sameAs semantics? Of course it is perfectly for you to do that but, at least for my taste, this is not entirely ok. In other words, it is exactly for the possibility of expressing non-symmetric identifications that new predicates might be needed, to allow you to do that while staying within the definitions...

Of course, as I say, the issue might be that this train is already gone, and there are too many owl:sameAs out there

# ...and also private communications

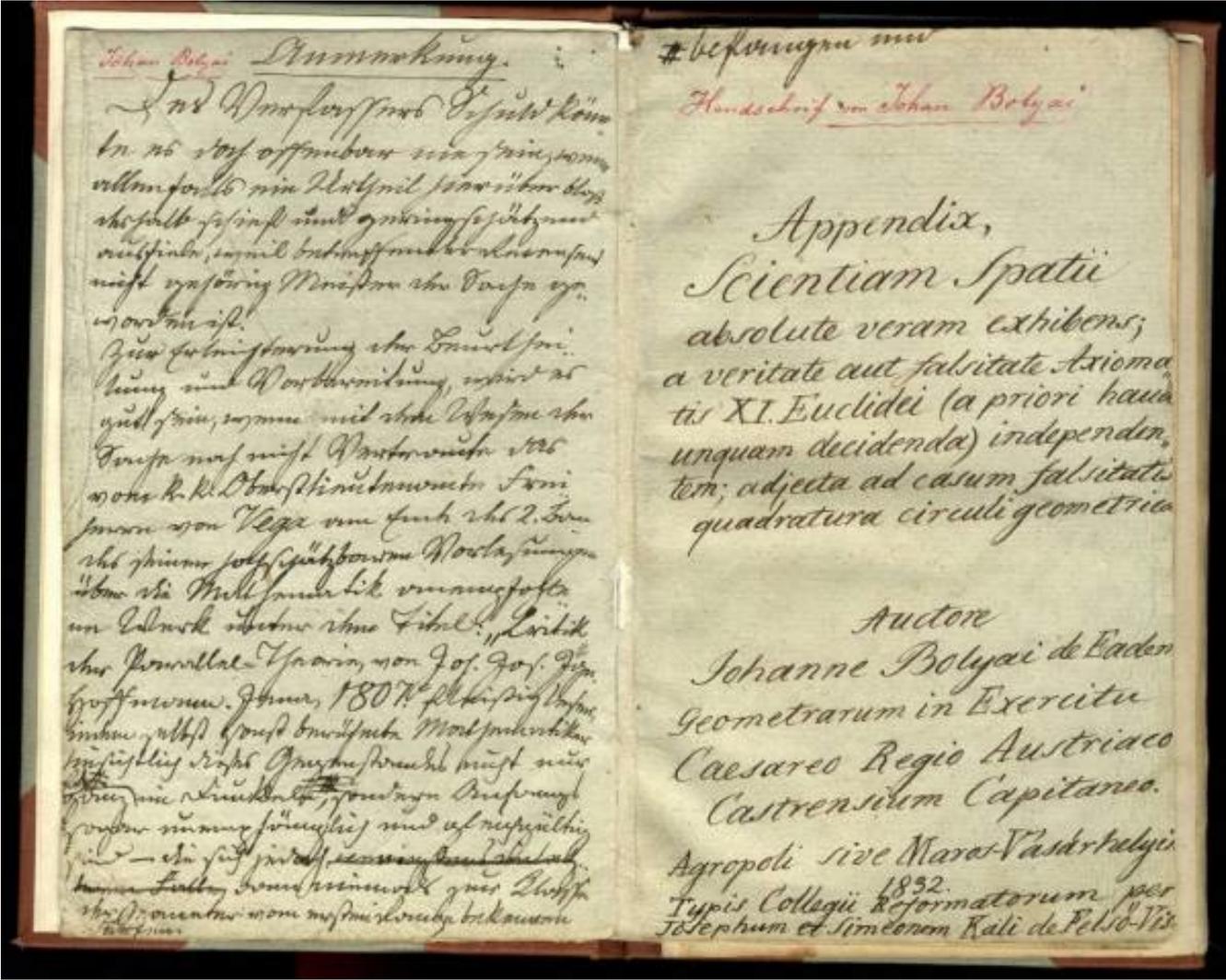


# This was an example of what scholarly communication is *really* all about

---

- ▶ Experts find one another's result
- ▶ They engage into private or public conversations, discussions
- ▶ They may lead to
  - new results
  - new, possibly common actions (that did happen in our case)
- ▶ They get to know and possibly influence one another's view
- ▶ Etc.

# Scholarly advances were always based on communication



# What is wrong with our story?

---

- ▶ I found out about the paper on a social site
  - *not* through a formal bibliography
- ▶ I could not get to the paper directly
  - though my Institute's library has a subscription to Elsevier's Web site...
  - I was not there! I.e., I had no access
  - I had to know (because I am part of the community) that there is a preprint server

# What is wrong with our story?

---

- ▶ I had access to PDF: like a paper printout, but on the screen
  - no access to higher resolution images
  - no access to the underlying data, so that I could check some of the statements
  - no access to the algorithm to really try it out (have you ever tried to *read* a complex computing algorithm?)
  - no direct link to all the other papers via the references
    - if I want to read them: for each paper the whole story starts all over again!

# What is wrong with our story?

---

- ▶ My blog led to
  - additional insights, possibly to both of us
  - maybe some practical results
    - submission of a specification to a standard body in this case
  - the paper certainly had an impact on me 😊...
- ▶ But... the whole series of communication, of references, etc., go unnoticed on the authors' official impact factors
  - and that is almost the *only* thing that counts for career advancement...

# Compare it with the Web age where...

---

- ▶ Web sites can offer you lively experiences on
  - images, interactive diagrams, video, audio
  - possibly illustrated algorithms running real time on demand
  - interactive control over remote program execution
- ▶ Hyperlinks are the norm: getting from one page to another is normal and expected

# Compare it with the Web age where...

---

- ▶ Storage is cheap: publishing data or images beyond pure text is common place
- ▶ Data mining is real: cross references, relationships, etc., become possible if the underlying content is “software friendly”

# Social behaviours are changing

---

- ▶ Experts communicate through emails, Twitter, Google+, Facebook, etc.
  - possibly more knowledge and information flows through these channels than through “official” scientific communications
  - this flow is not measured for scientific career purposes
- ▶ Pace of information exchange is higher, a publication must be made available almost instantaneously
  - compare it to the long publication delays through official channels

# Social behaviours are changing

---

- ▶ There is an information overload; people expect technical help in managing it
- ▶ Collaborative platforms come to the fore where scientific discourse may happen through common development and discussion
- ▶ etc.

*Scholarly communication should move away from its paper centric model and traditions, and join the information age!*





# SCHLOSS DAGSTUHL

Leibniz-Zentrum für Informatik

[About Dagstuhl](#)[Program](#)[Publications](#)[Library](#)

You are here: [Program](#) » [Calendar](#) » Seminar Homepage

<http://www.dagstuhl.de/11331>

**15.08.11 — 18.08.11, Seminar 11331**

## Perspectives Workshop: The Future of Research Communication

### Organizers

Timothy W. Clark (Mass General Hospital & Harvard Medical School, US)

Anita De Waard (Elsevier Labs - Jericho, US)

Ivan Herman (CWI - Amsterdam, NL)

Eduard H. Hovy (University of Southern California - Marina del Rey, US)



### Book exhibition

Books from the participants of the current Seminar

Book exhibition in the library, 1st floor, during the seminar week.

### Documentation

In the series Dagstuhl Reports each Dagstuhl Seminar and Dagstuhl Perspectives Workshop is documented. The seminar organizers, in cooperation with the collector, prepare a report that includes contributions from the participants' talks together with a summary of the seminar.

Download [overview leaflet \(PDF\)](#).

### Publications

Seminar participants may publish preprints within the scope of the seminar documentation as part of the Dagstuhl Preprint Archive.

Furthermore, a comprehensive peer-reviewed collection of research papers can

Dagstuhl Seminars  
Dagstuhl Perspectives  
GI-Dagstuhl Seminars  
Events  
Research Guests  
**Calendar**  
Seminars  
Events

# Dagstuhl workshop on “Future of Scholarly Communication”



- ▶ Took place on 15-18 August, 2011
- ▶ Large participation of people with very different scientific backgrounds
  - biologists
  - computer scientists
  - librarians
  - publishers
  - astrophysicists
  - ...

# The goal was *not* to

---

- ▶ Solve all the issues
- ▶ Come up with ready solution
- ▶ Develop new software
- ▶ Define full-blown new business models
- ▶ ...

# Instead...

---

- ▶ Experiences were exchanged among communities
- ▶ A high level synthesis of the issues was provided
  - A “Dagstuhl Manifesto”  
<http://dx.doi.org/10.4230/DagMan.1.1.41>
  - A separate Dagstuhl report:  
<http://drops.dagstuhl.de/opus/volltexte/2011/3315/>
  - A shorter version in the “Force11 Manifesto”:  
[http://force11.org/white\\_paper](http://force11.org/white_paper)
- ▶ *Build a community around the common goal*

*Goal: go beyond everyone's  
respective disciplines and look at the  
issues around scholarly  
communications in general*

# Meeting Recommendations

---

- ▶ Three major areas were explored
  1. Mechanisms and methodologies of publication
  2. Finances: business models for the future
  3. Re-evaluate quality assessment of researchers

# Mechanisms and Methodology of Publications

---

- ▶ What does “publication” mean?
- ▶ Text, but increasingly together with:
  - associated datasets, video, images, software, workflow plans—anything used in research
  - related papers, possibly reflecting the evolution of the paper’s thoughts
  - social network of related researchers
  - possibly earlier versions of the paper, with commentaries
  - metadata (publication date, exact references, quality verification results,...)

# Mechanisms and Methodology of Publications

---

- ▶ Associated information should be stored, like
  - datasets, mathematical models, etc.
  - software/algorithms in downloadable and runnable form
  - images, video, audio, etc., in indexable formats
- ▶ All these should be stored and made available using public standards
  - storage and encoding formats for the data components
  - metadata standards for the additional information
  - etc.

# Mechanisms and Methodology of Publications

---

- ▶ New procedures, systems, services, etc., are needed
  - manage information, extract relationships, help collaboration
  - link research data to other research data on the Web
  - help in gaining further insights, finding trends
  - etc.
- ▶ *Some of this work is hard* and require funding, new business models

# Finances: Business model for the future

---

- ▶ Role of publishers should change
  - current subscription models, closed information storage, stringent copyright policies, etc., may not work for long
  - no longer in charge of textual content only; manage storage and systematization of all the data, including the text
  - focus on long term archiving, preservation
  - provide extra valuable services on the research data; that should be the major source of future income

# Finances: Business model for the future

---

- ▶ Role of libraries also change
  - focus not on preservation but on access to information on the Web at large
  - user services of all kinds become more important
- ▶ New players (e.g., Google) appear on the market
  - their presence changes the role of both publishers and libraries

# Re-evaluate quality assessment of researchers

---

- ▶ New impact measurement should be developed
  - assess *usage* of publication (including ancillary material)
  - assess on-line discussions, cross-references, additional data, impact via Social Web
- ▶ Extended notion of “publication” brings new forms of contributions:
  - data collection creation, statistical processing of others’ results
  - creation of software that users others’ data
  - production of new type of software and assess its usage by the community

# Since the Workshop

---

- ▶ Produced the separate [white paper](#) (beyond the Dagstuhl report)
- ▶ Provided inputs to different governmental initiatives on scholarly communications
- ▶ Set up a separate [Web Site](#)
- ▶ Maintain a separate mailing list
  - seeded with the workshop participants' list
- ▶ Try to get extra funding to develop the ideas, software, web site, etc., further
- ▶ Spread the word...

# FORCE 11

[Contact](#) | [RSS Feed](#) | [Login](#) | [Join](#)

*the Future of Research Communications and e-Scholarship*

[About](#) | [Force 11 Publications](#) | [Editors' Picks](#) | [Blogs](#) | [Events](#) | [Tools and Resources](#) | [Members and Community](#)



Force11 is a virtual community working to transform scholarly communications through advanced use of computers and the Web.

We invite you to join us.



## Force 11 Publications

[Force 11 Manifesto](#)

[Read More](#)

## Editors' Picks

[Google Scholar Citations Open To All](#). [Google Scholar Blog](#), Wednesday, November 16, 2011 | 8:30 PM

[Computational biology resources lack persistence and usability.](#)

[URL decay in MEDLINE--a 4-year follow-up study.](#)

[A Gathering Storm of Scholarly Transformation](#)

[An open annotation ontology for science on web 3.0.](#)

[Read More](#)

## Popular Content

[The 33rd Annual IATUL Conference, Singapore, 4-7 Jun 2012](#)

[International Provenance and Annotation Workshop, Santa Barbara, 19-21 Jun 2012](#)

[SePublica2012 Workshop, May 27-31, 2012, Heraklion, Greece, part of the ESWC 2012 Conference](#)

[About Force 11](#)

[Semantic science and its communication - a personal view.](#)

# A final remark...

- ▶ This workshop was part of a series of similar events
- ▶ Everybody was charmed by the facilities of Dagstuhl!



# Thank you for your attention

These slides are also available on the Web:

<http://www.w3.org/2012/Talks/0319-Dagstuhl-IH/>

