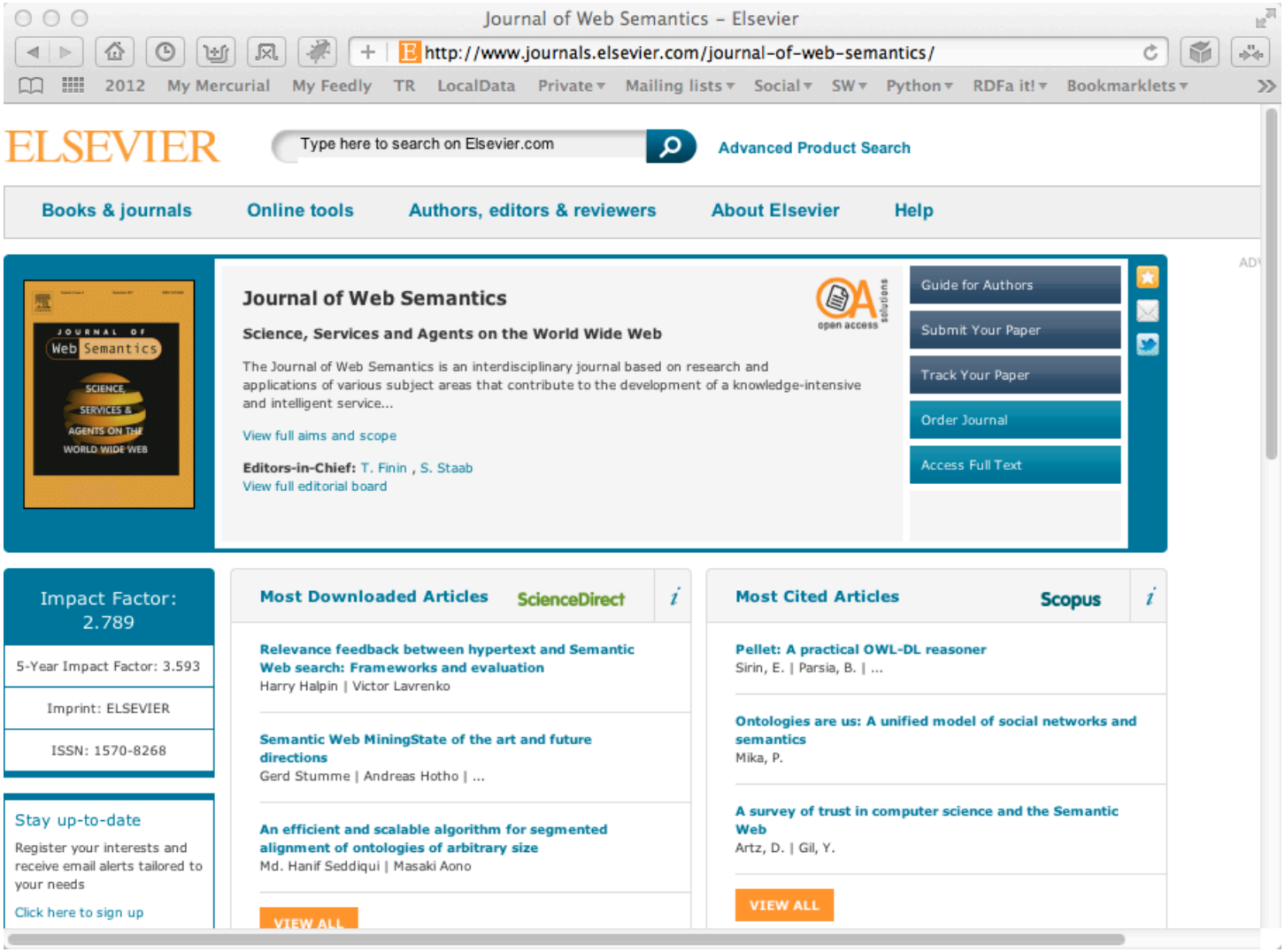
Report on the "Future of Research Communications" Workshop Dagstuhl, August 15-18, 2012

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

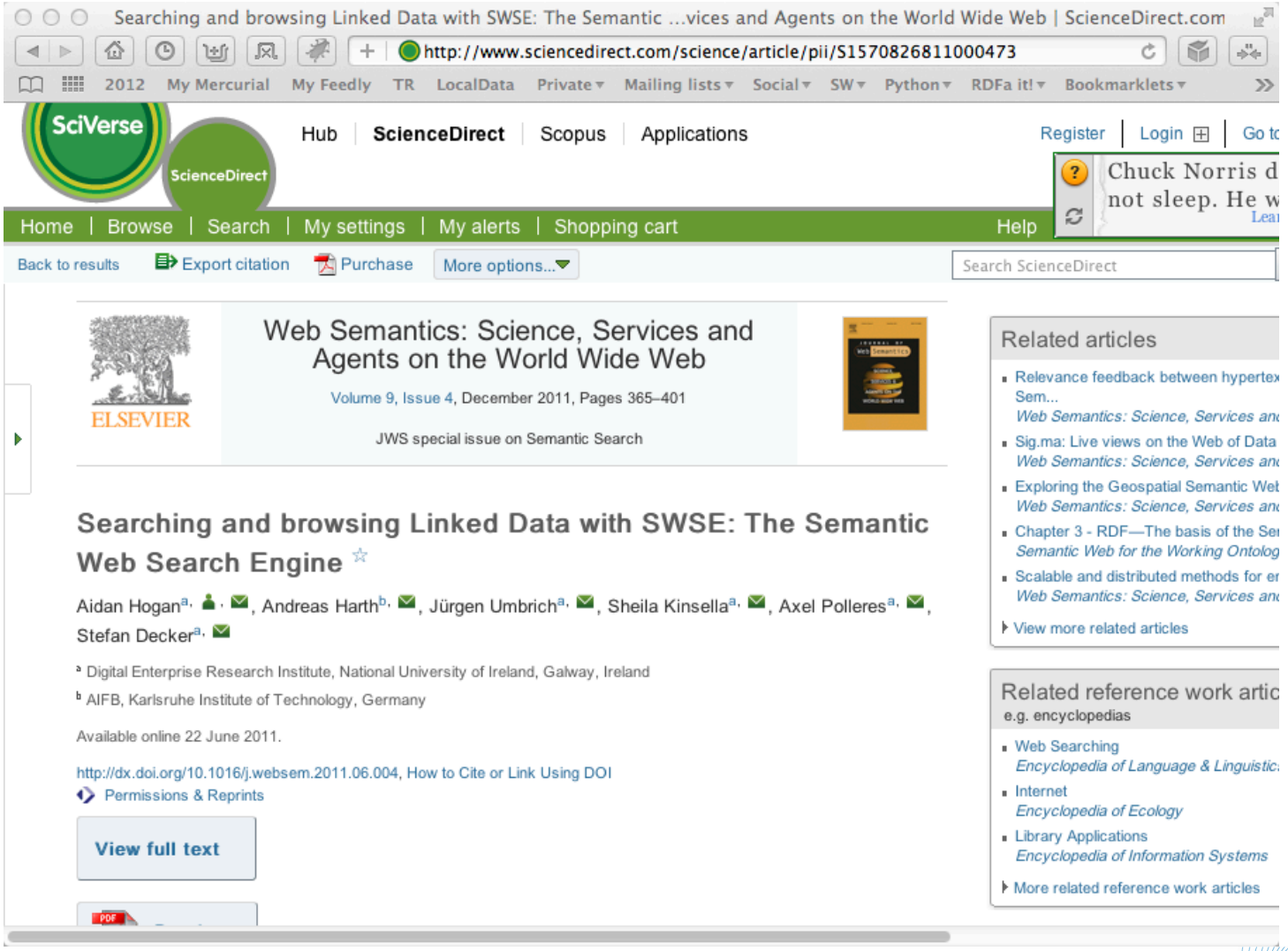
Let me tell you a story…

I heard about an interesting paper on a social site

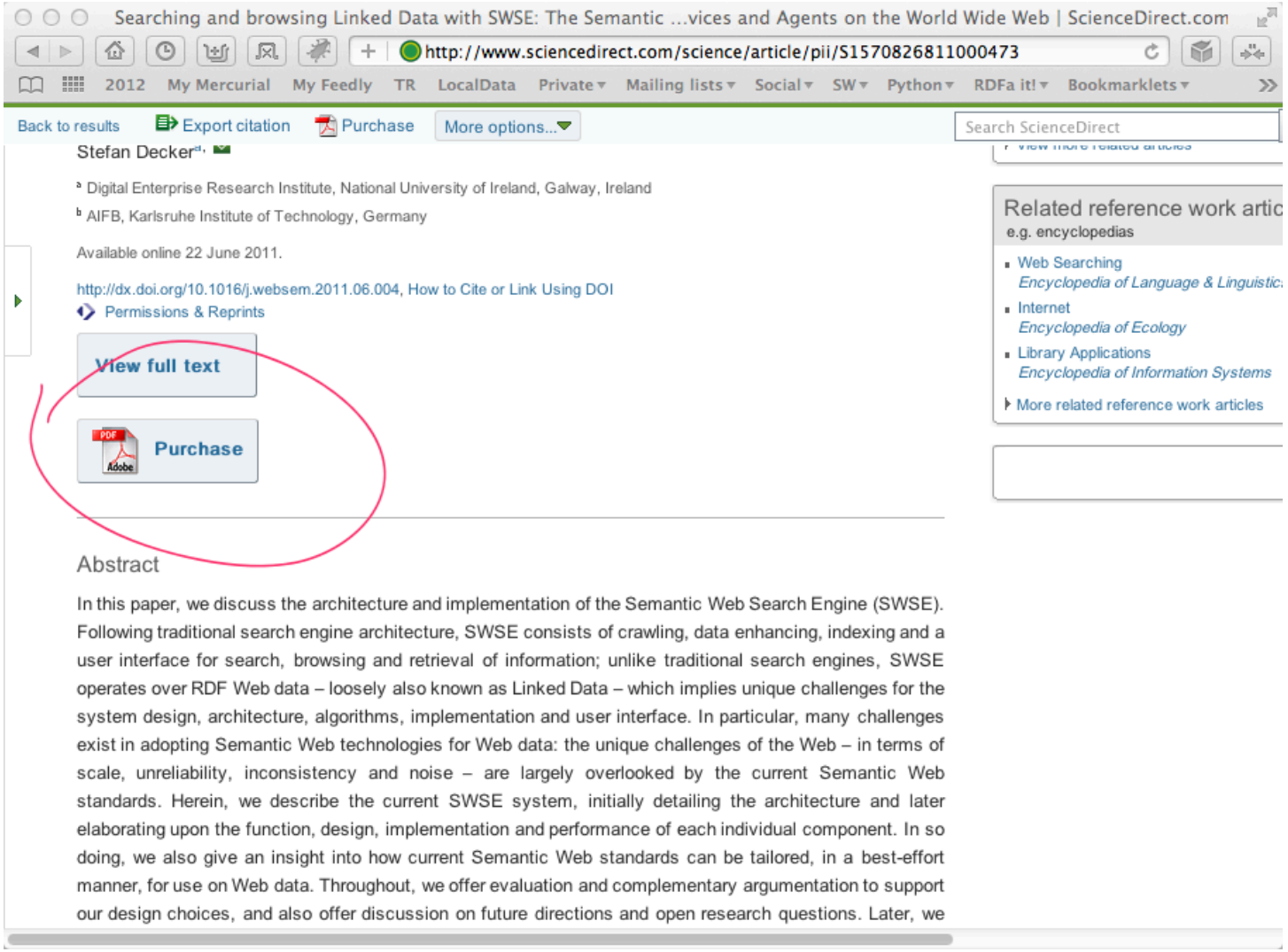
So I looked up the journal on the Web



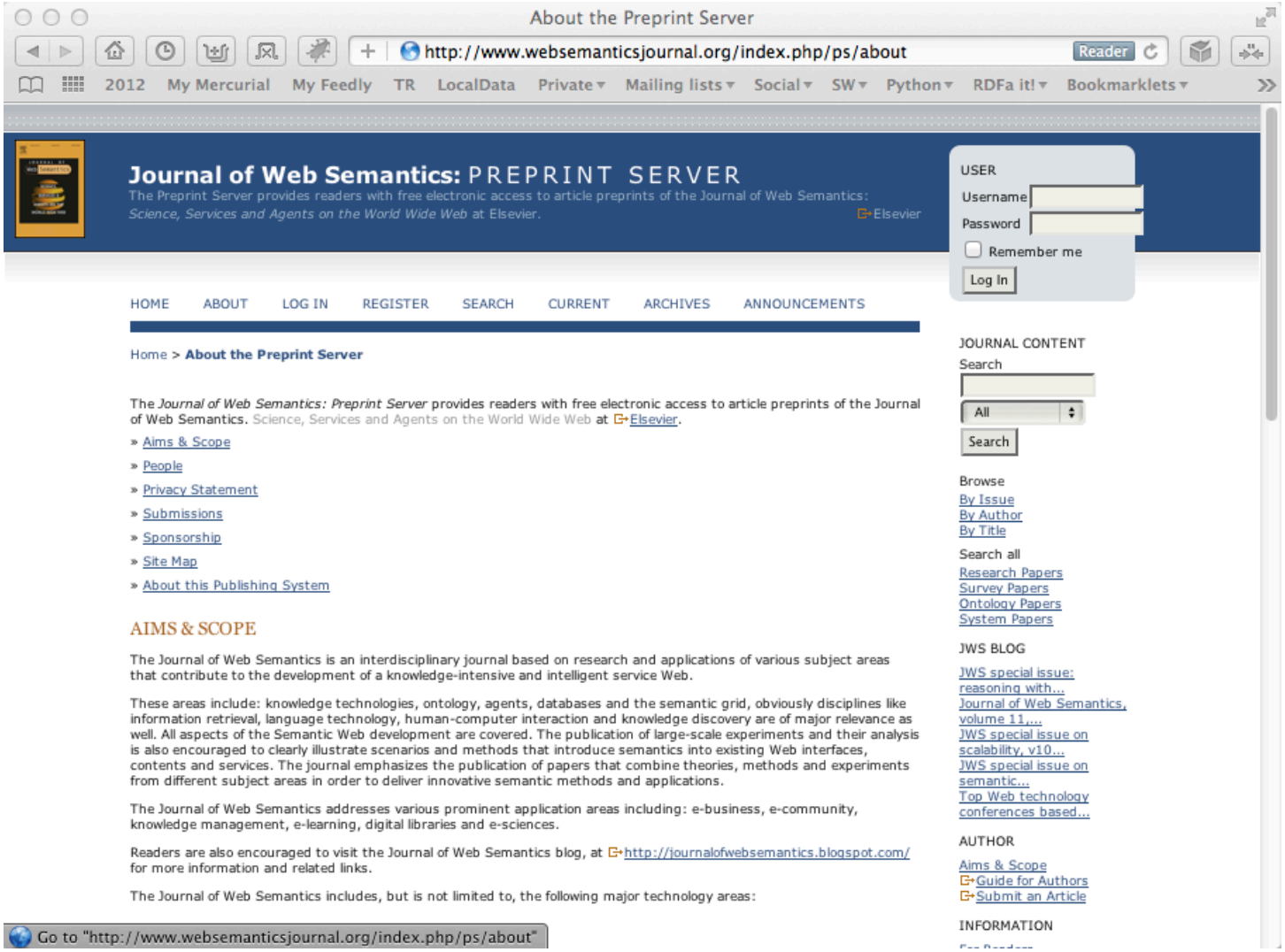
Found the paper



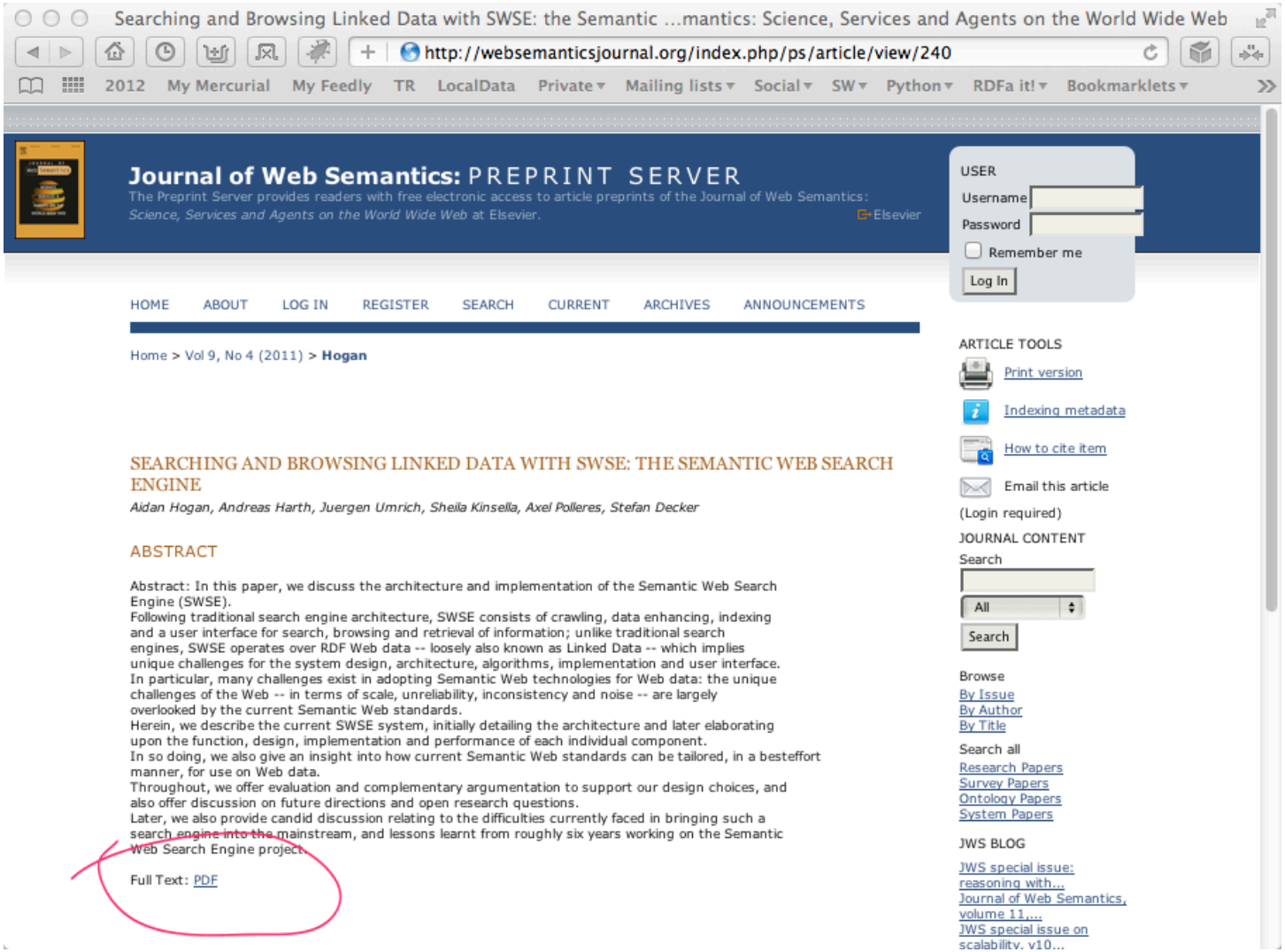
However, I was not at my institute…



But… I knew the secret!

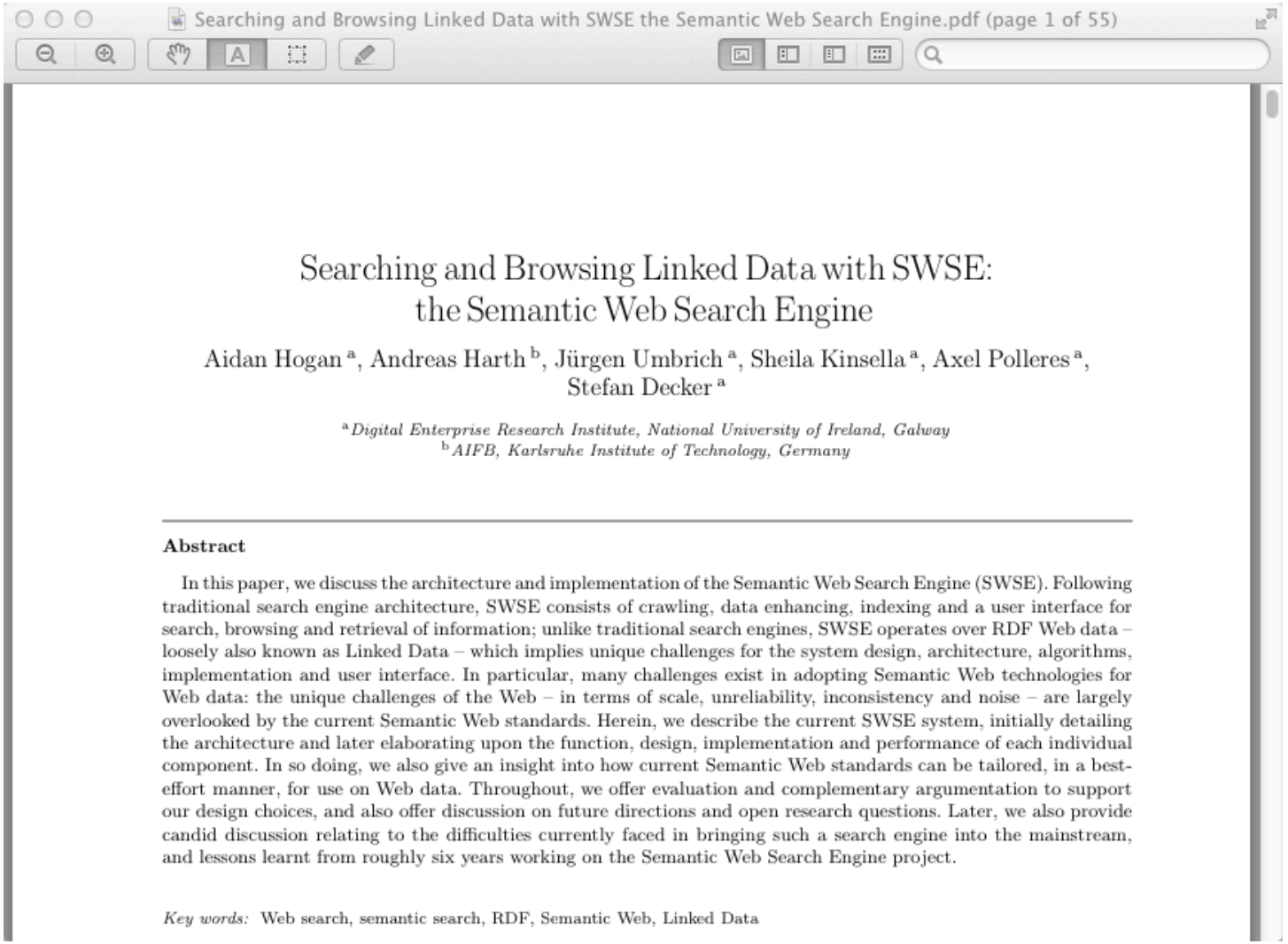


But… I knew the secret!

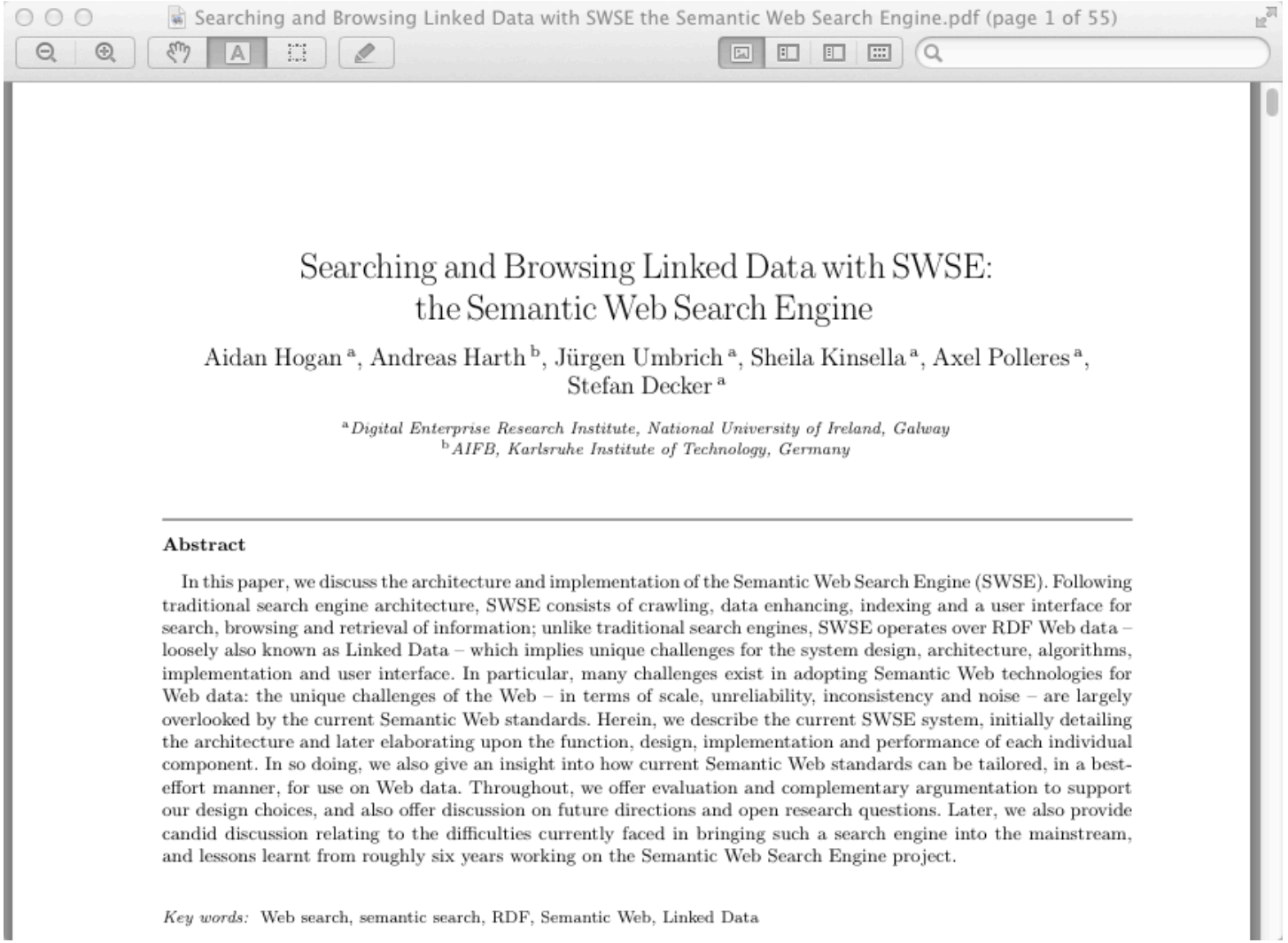


FORCE11

So I could read the paper.



The paper also had…



Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine

Aidan Hogan [a], Andreas Harth [b], Jürgen Umbrich [a], Sheila Kinsella [a], Axel Polleres [a], Stefan Decker [a]

[a] *Digital Enterprise Research Institute, National University of Ireland, Galway*
[b] *AIFB, Karlsruhe Institute of Technology, Germany*

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 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.

Key words: Web search, semantic search, RDF, Semantic Web, Linked Data

… (low resolution) diagrams, and …

Searching and Browsing Linked Data with SWSE the Semantic Web Search Engine.pdf (page 4 of 55)

Fig. 1. Results view for keyword query `Bill Clinton`

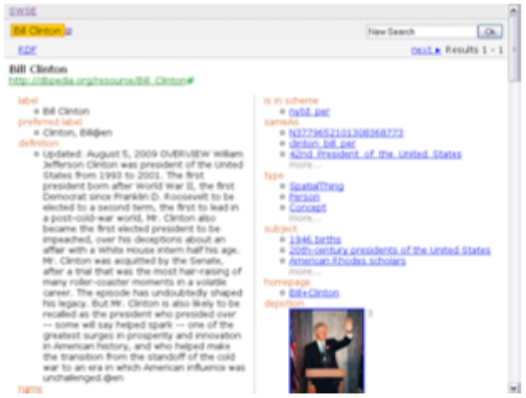


Fig. 2. Focus view for entity `Bill Clinton`

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-

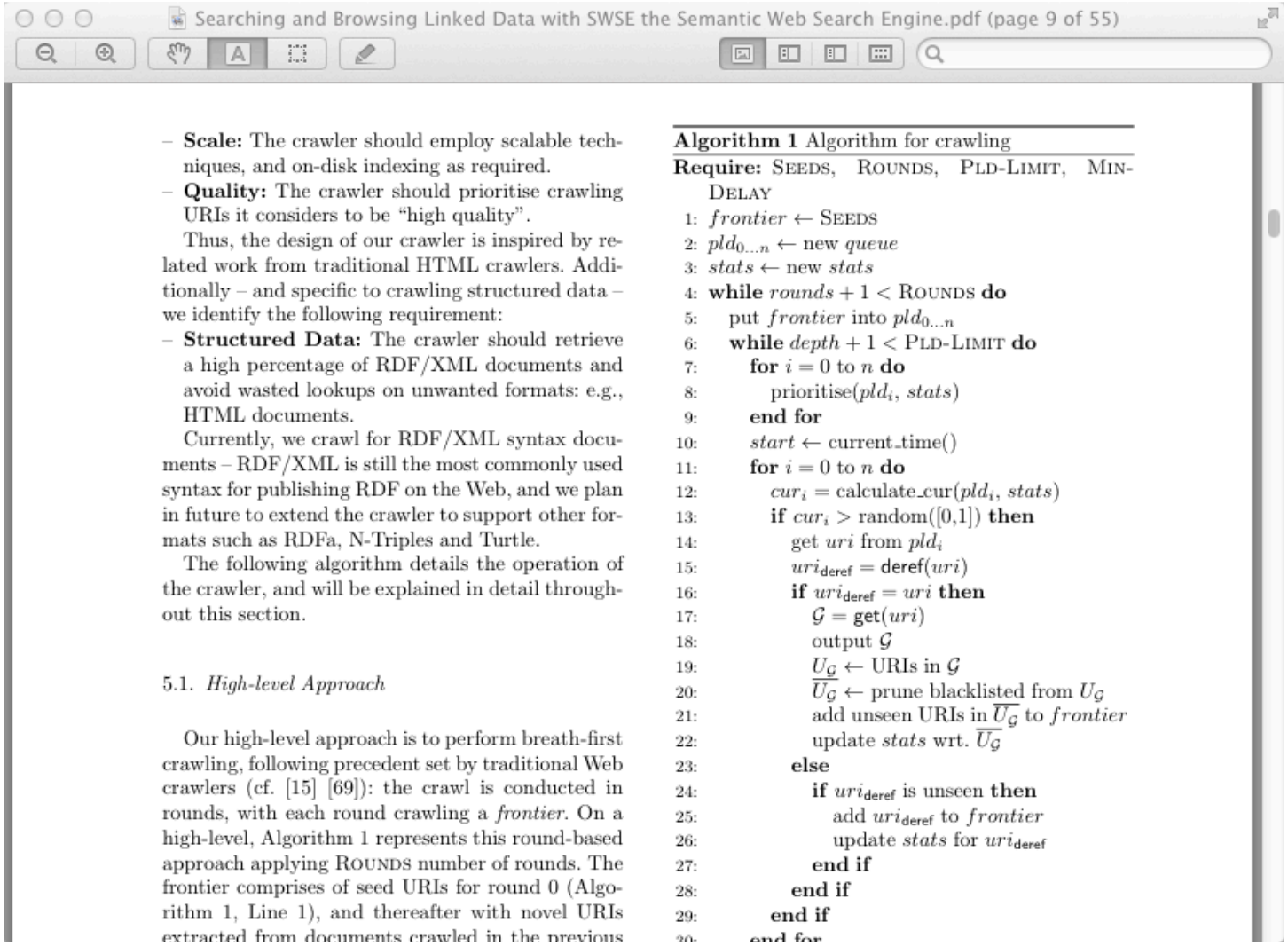
example, see [54]).

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

…lots of data; and of course…

Searching and Browsing Linked Data with SWSE the Semantic Web Search Engine.pdf (page 13 of 55)

Time taken for a crawl performing lookups on 100 k URIs, and average percentage of time each queue had to enforce a politeness wait, for differing numbers of machines

ber of machines, number of minutes taken for the crawl, and also the percentage of times that the in-memory queue had to be delayed in order to abide by our politeness policies. There is a clear increase in the performance of the crawling with respect to increasing number of machines. However, in moving from four machines to eight, the decrease in time is only 11.3%. With 8 machines (and indeed, starting with 4 machines), there are not enough active PLDs in the queue to fill the adjusted min-delay of 4 seconds (8*500 ms), and so the queue has a delay hit-rate of 94.6%.

We term this state *PLD starvation*: the slave machines do not have enough unique PLDs to keep them occupied until the MIN-DELAY has been reached. Thus, we must modify somewhat the end-of-round criteria to reasonably improve performance in the distributed case:

- firstly, a crawler can return from a round if the MIN-DELAY is not being filled by the active PLDs in the queue – the intuition here being that new PLDs can be discovered in the *frontier* of the next round;
- secondly, to ensure that the slave machines don't immediately return in the case that new PLDs are not found in the *frontier*, we implement a PLD-LIMIT which ensures that slave machines don't immediately return from the round;
- finally, in the case that one slave crawler returns from a round due to some stopping criteria, the master machine will request that all other slave

observe that 8 machines currently approaches the limit of performance given our setup and policies.

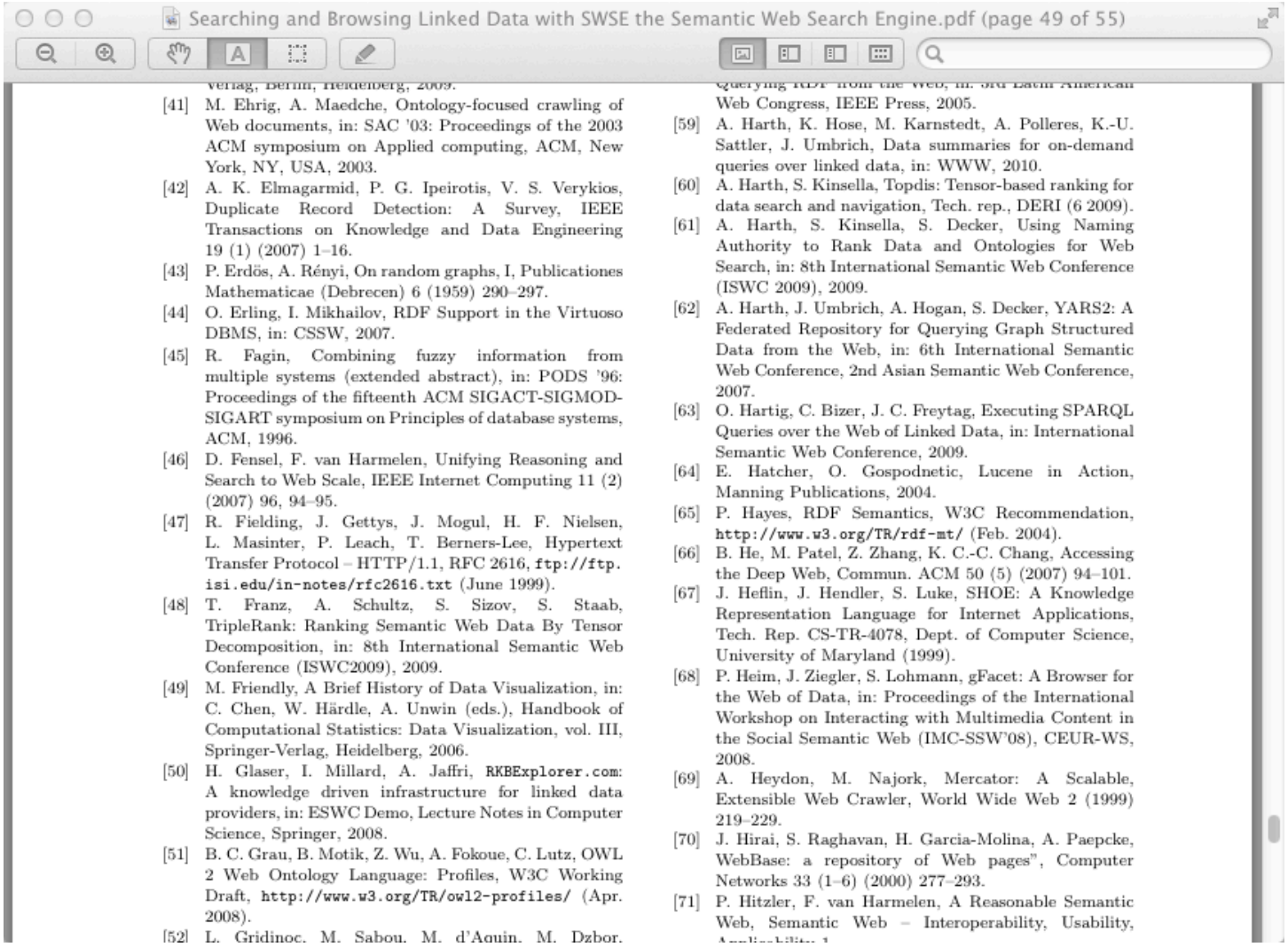
5.3. *Full-Scale Evaluation*

To perform scale-up experiments for the crawler – and indeed to achieve a large dataset for evaluation of later components – we ran the crawler continuously for 52.5 h on 8 machines from a seed list of ~8 m URIs extracted from an old dataset with *cur* scoring enabled. [17] In that time, we gathered a total of 1.118 g quads, of which 11.7 m were duplicates (~1% – representing duplicate triples being asserted in the same document); we provide a selection of statistics characterising the dataset in Appendix A.

We observed a mean of 140 m quads per machine and an average absolute deviation of 1.26 m across machines: considering that the average absolute deviation is ~1% of the mean, this indicates near optimal balancing of output data on the machines.

The crawl attempted 9.206 m lookups, of which 448 k (4.9%) were for `robots.txt` files. Of the remaining 8.758 m attempted lookups, 4.793 m (54.7%) returned response code `200 Okay`, 3.571 m (40.7%) returned a redirect response code of the form `3xx`, 235 k (2.7%) returned a client error code of the form `4xx` and 95 k (1.1%) returned a server error of the form `5xx`; 65 k (0.7%) were disallowed due to restrictions specified by the `robots.txt` file. Of the 4.973 m lookups returning response code `200 Okay`, 4.022 m (83.9%) returned content-type `application/rdf+xml`, 683 k (14.3%) returned `text/html`, 27 k (0.6%) returned `text/turtle`, 27 k

… *lots* of references.



Verlag, Berlin, 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. Erdös, 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. Gridinoc, 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).

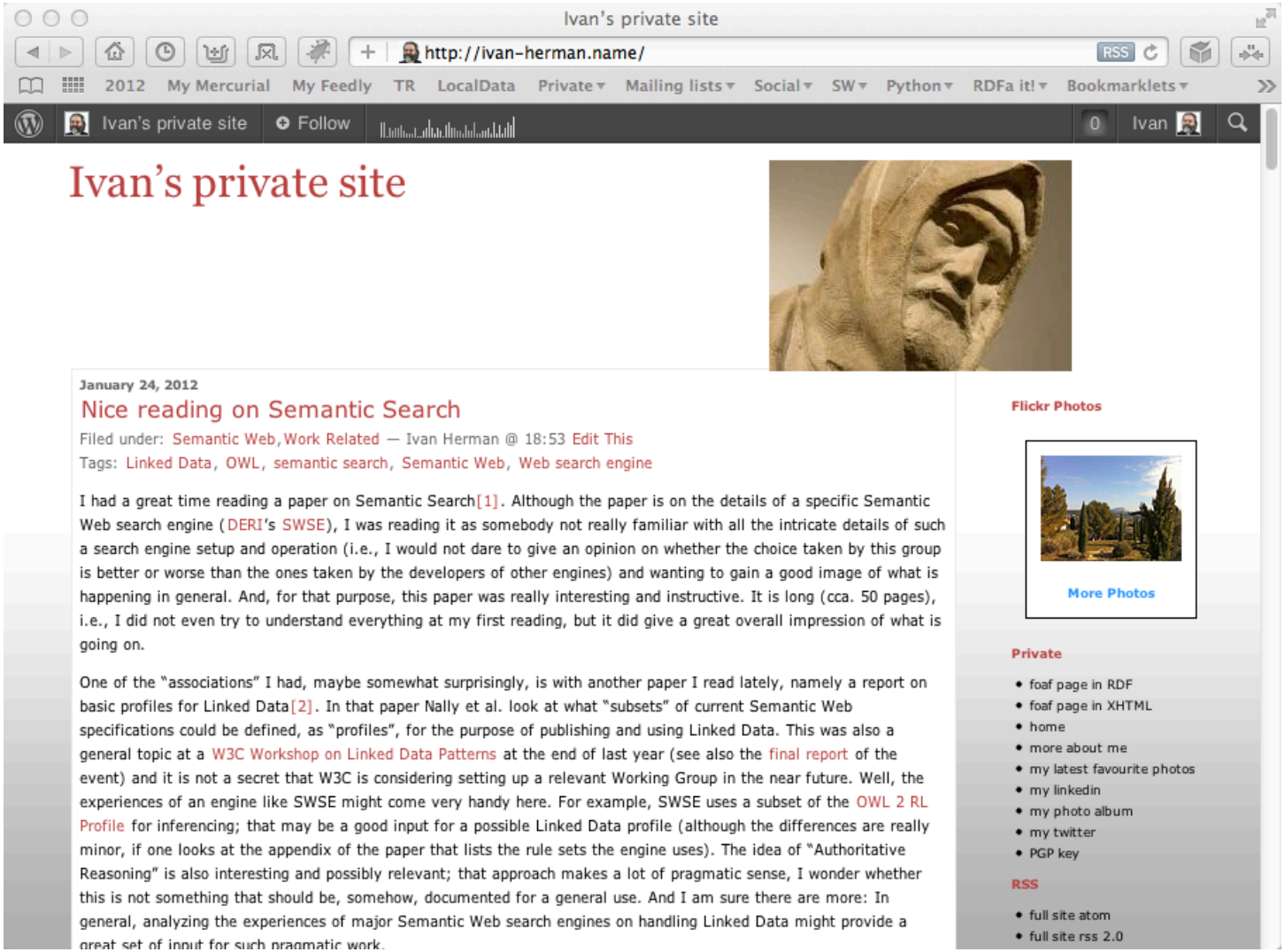
[68] P. Heim, J. Ziegler, S. Lohmann, gFacet: A Browser for the Web of Data, in: Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08), CEUR-WS, 2008.

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

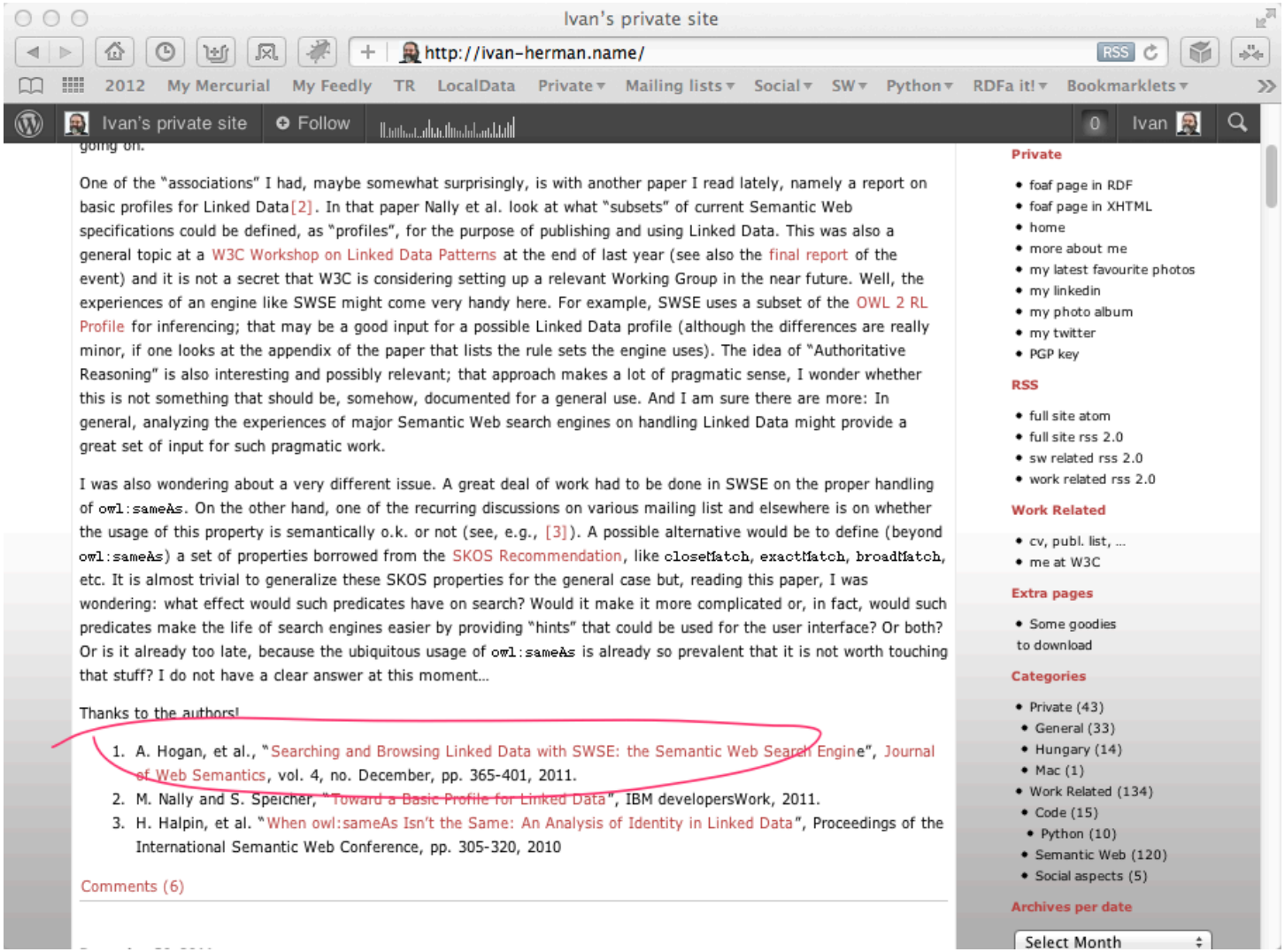
[70] J. Hirai, S. Raghavan, H. Garcia-Molina, A. Paepcke, WebBase: a repository of Web pages", Computer Networks 33 (1–6) (2000) 277–293.

[71] P. Hitzler, F. van Harmelen, A Reasonable Semantic Web, Semantic Web – Interoperability, Usability,

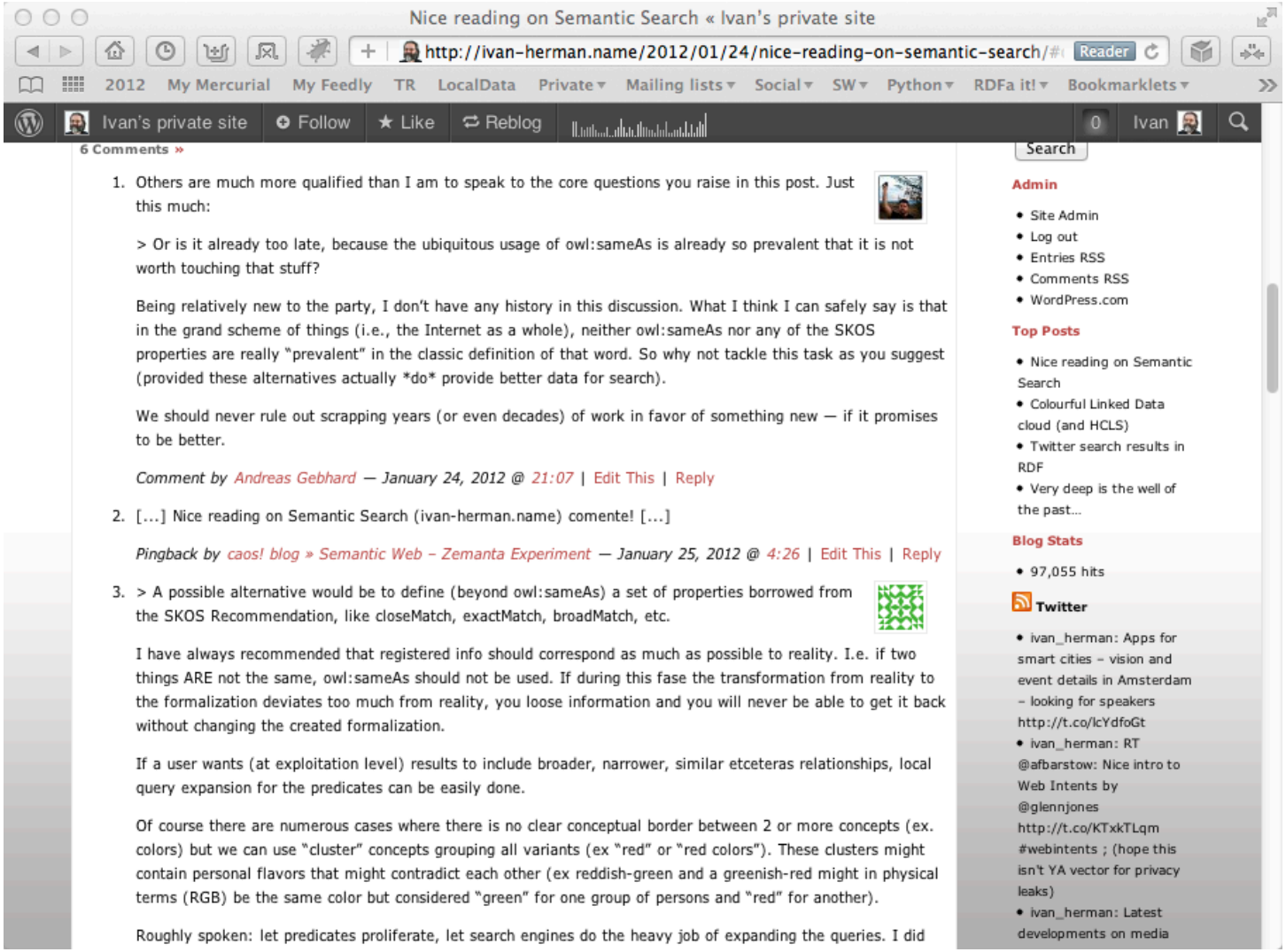
I liked the paper. So I wrote a blog…



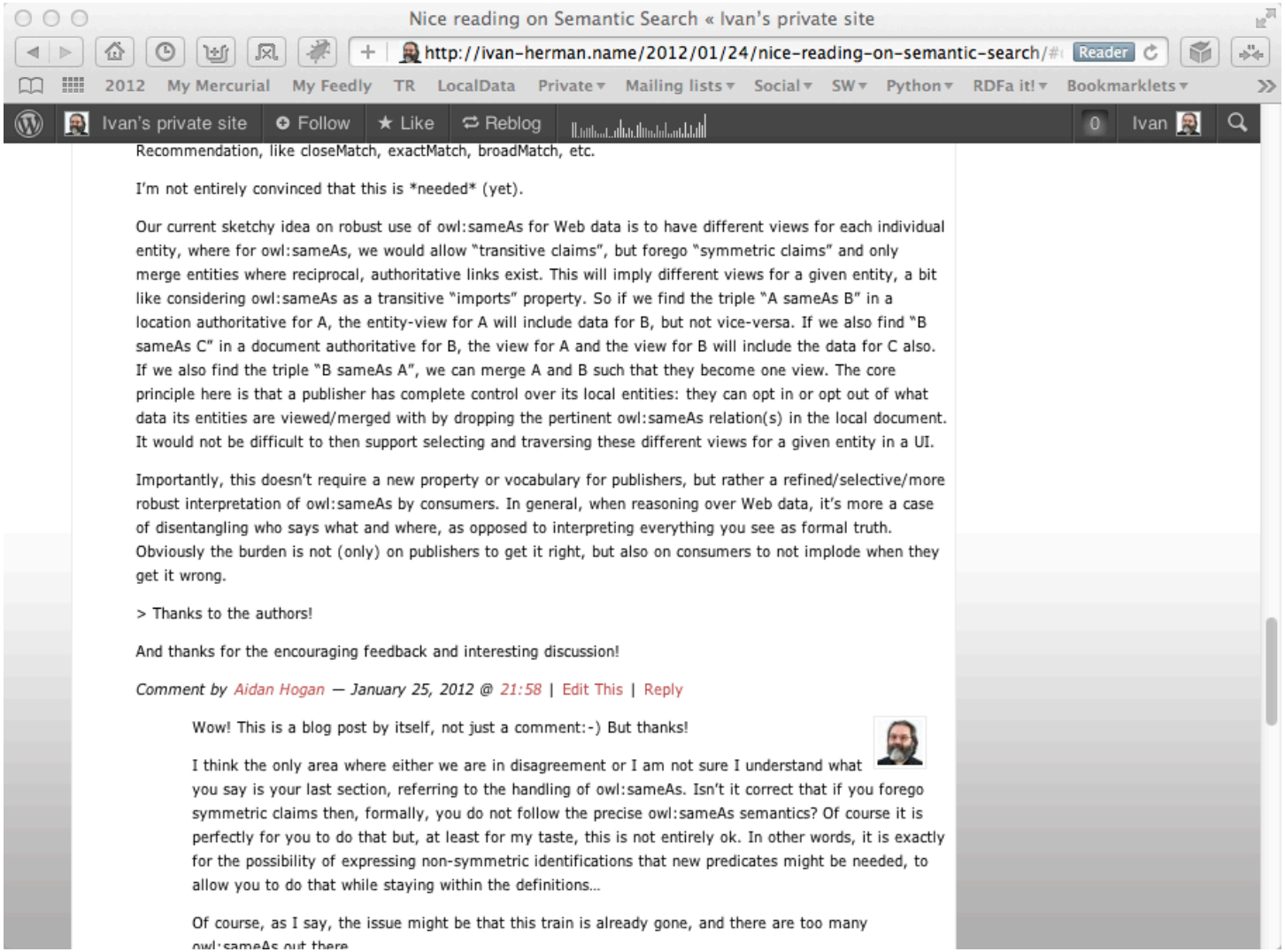
…with exact references.



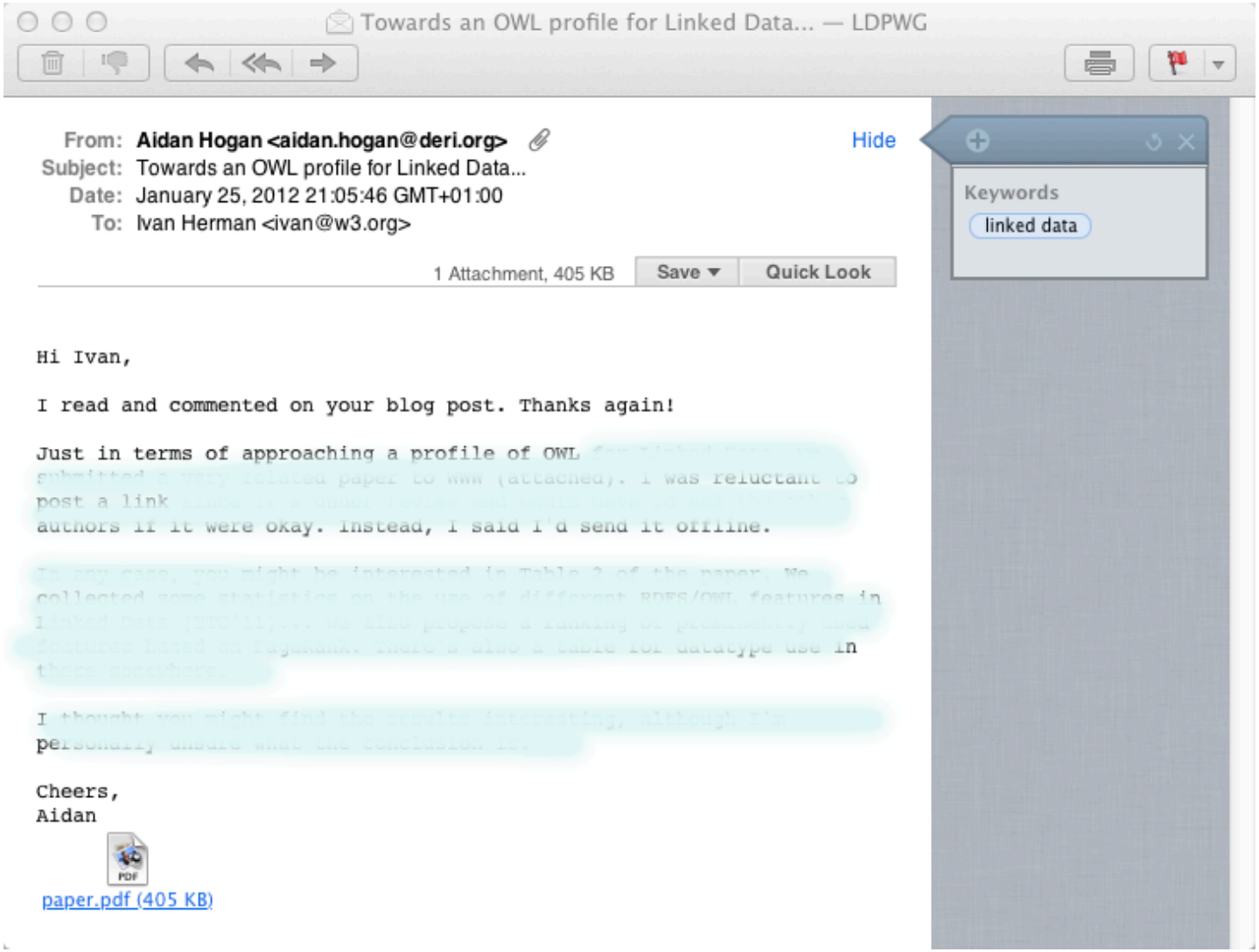
There were a bunch of comments…



…including the author's, with my answer…



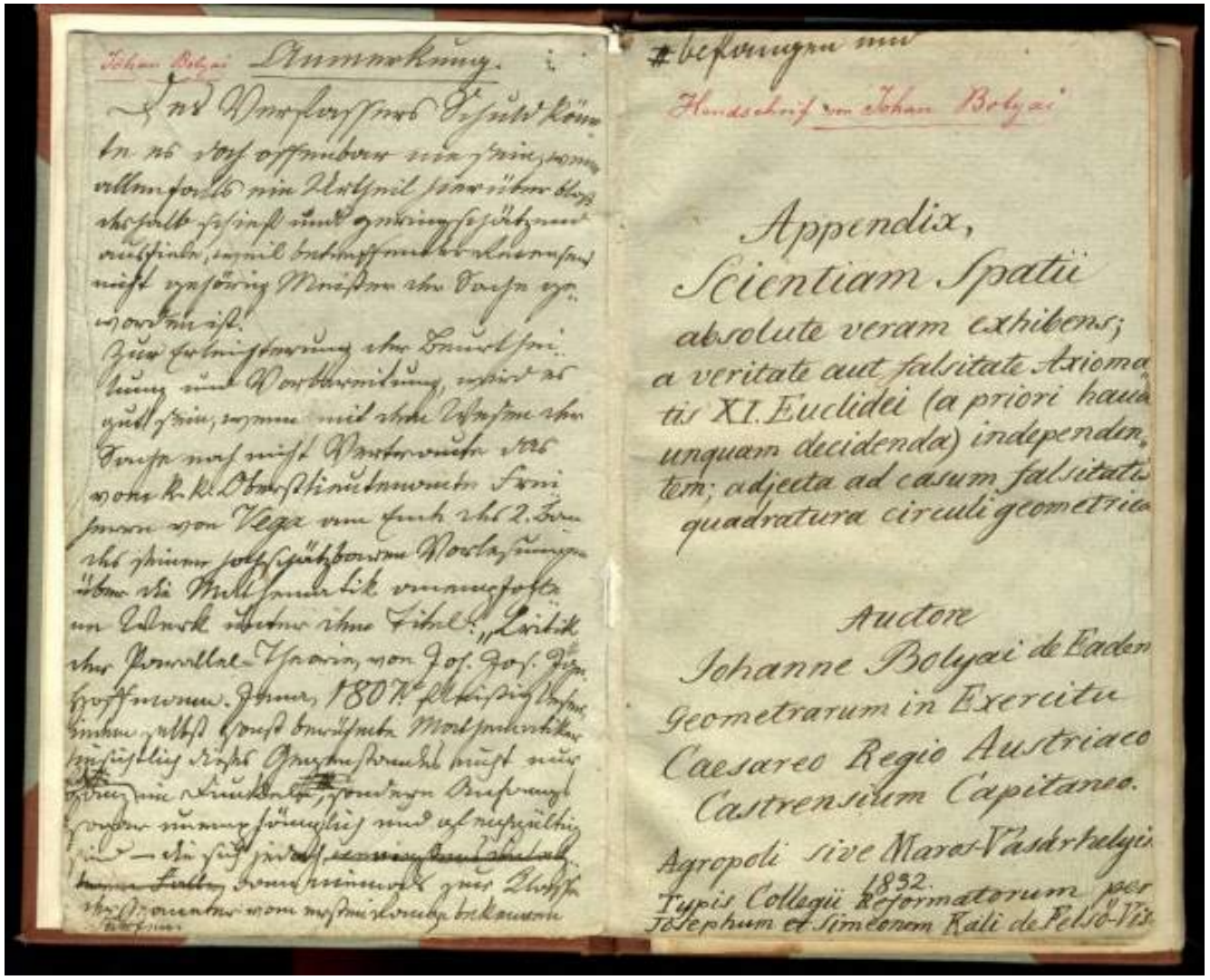
…and also private communications



This was and 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…

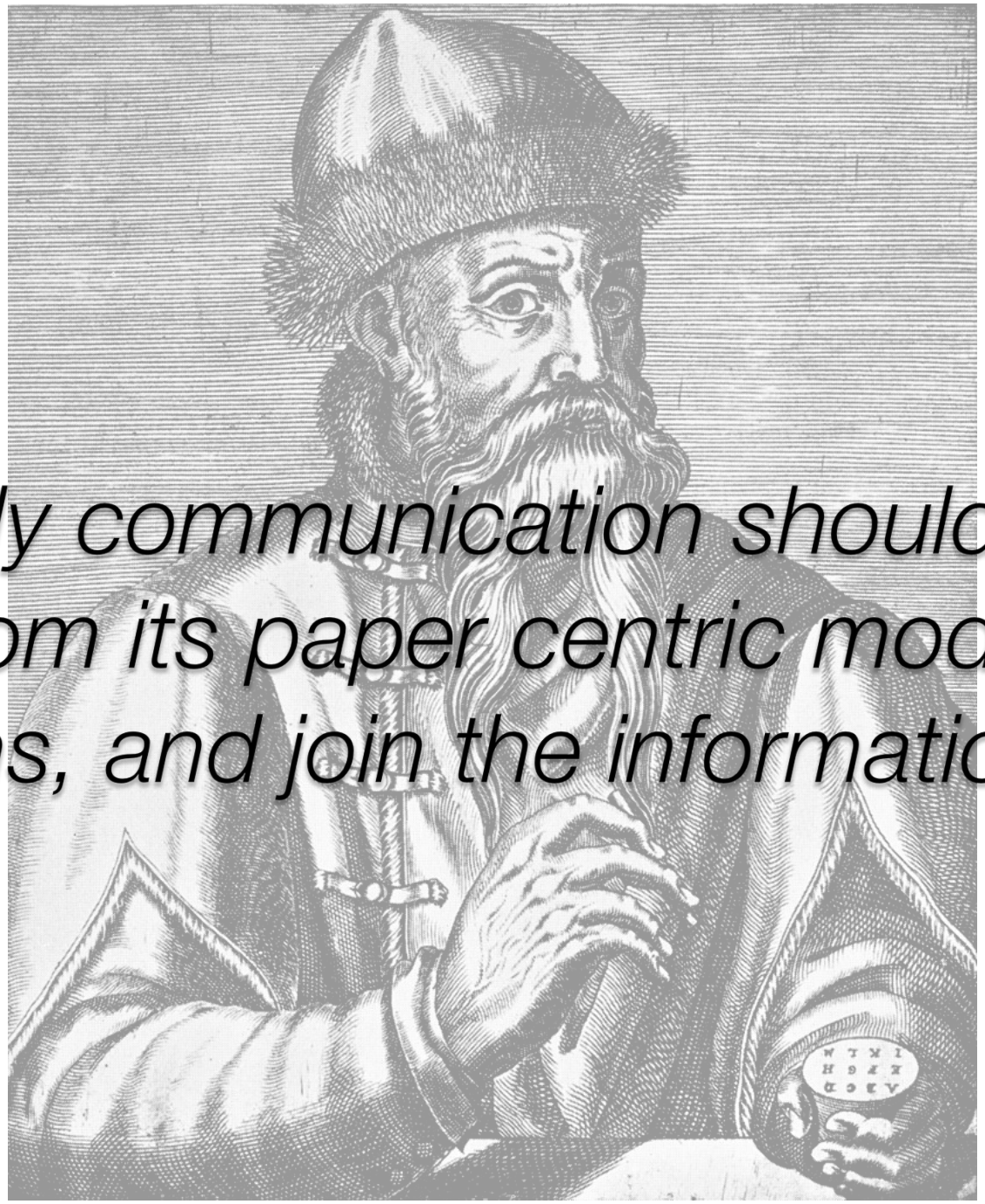
- Storage is cheap: publishing data or images beyond pure text is common place
- Data mining is real: cross references, relationships, etc., become possible if the underlying content is “software friendly”

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

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 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
- *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…

FORCE11

the Future of Research Communications and e-Scholarship

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/