Open access: why, where and how?

Publishing Integrity in the Digital Age
Workshop for Journal Editors
University Library in Split, University of Split, Split, Croatia
6-7 July 2012

Jadranka Stojanovski
University of Zadar, Ruđer Bošković Institute

Statements

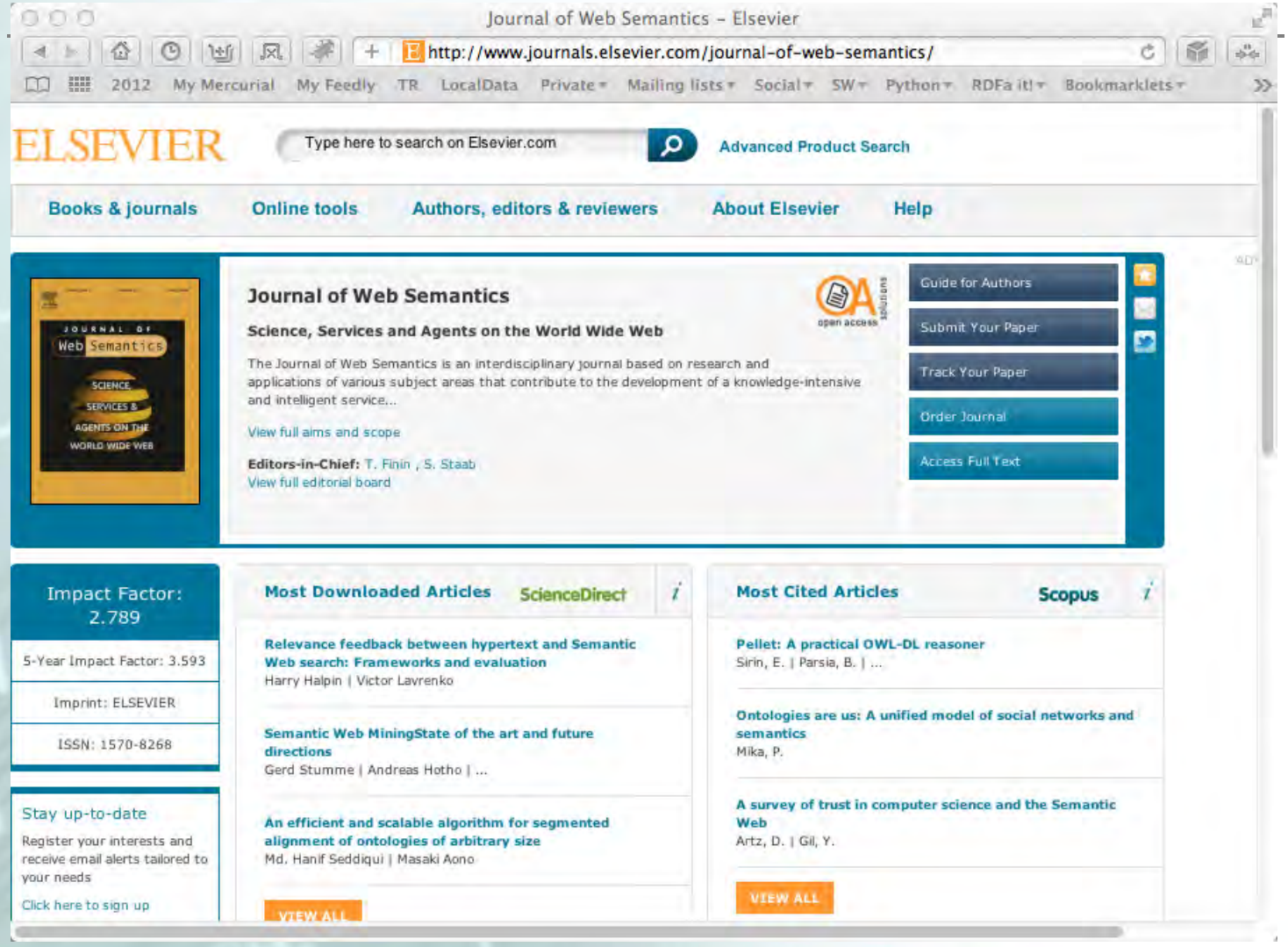
- no originality
- exaggeration
- no expertize
- not structured well
- not citing properly (but still citing)
- maybe boring (I'm a librarian)

A STORY FROM REAL LIFE

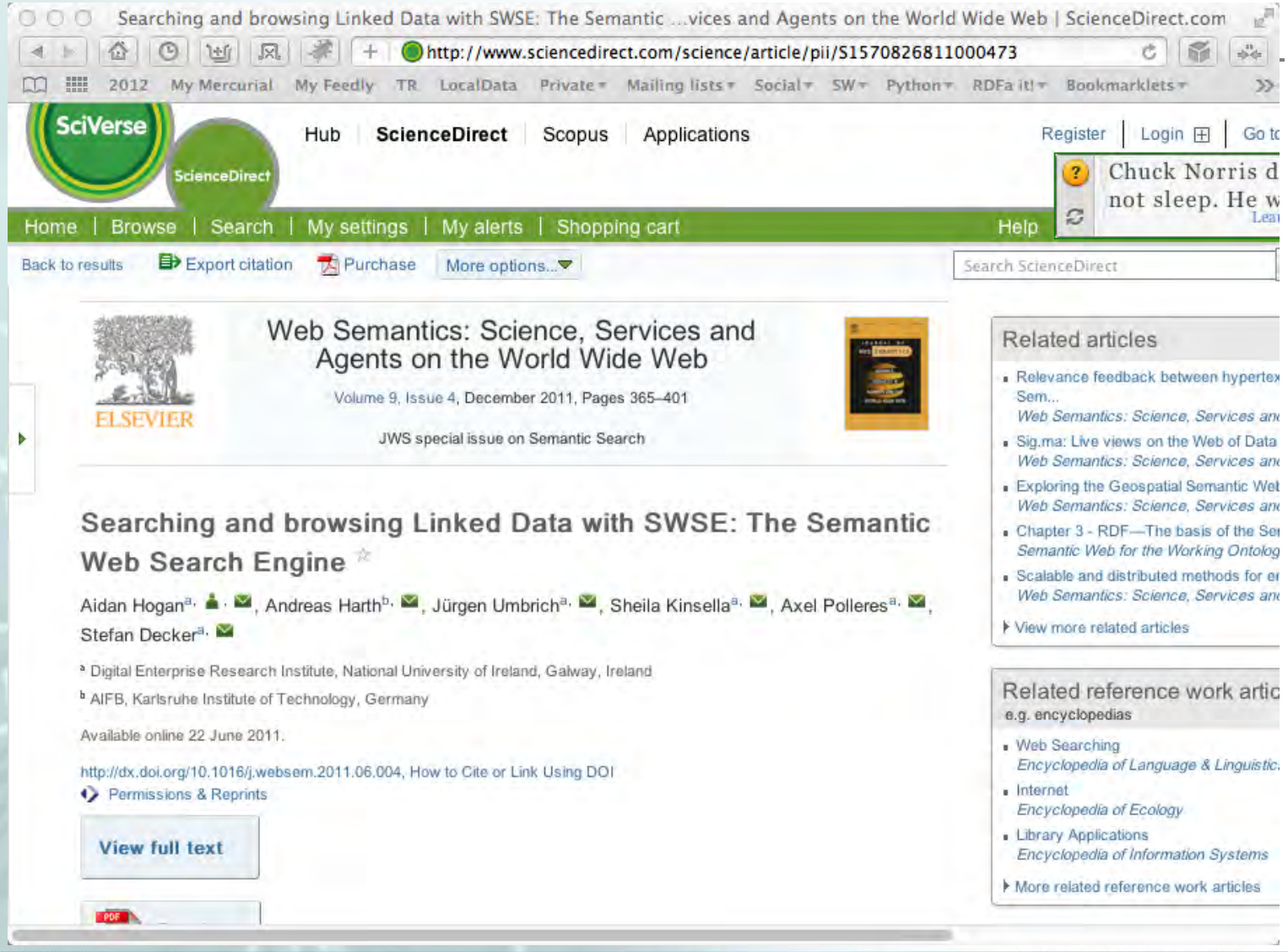
from Ivan Herman: Report on the “Future of Research Communications” Workshop, Dagstuhl, August 15-18, 2012

I. Herman heard about an interesting paper on a social site

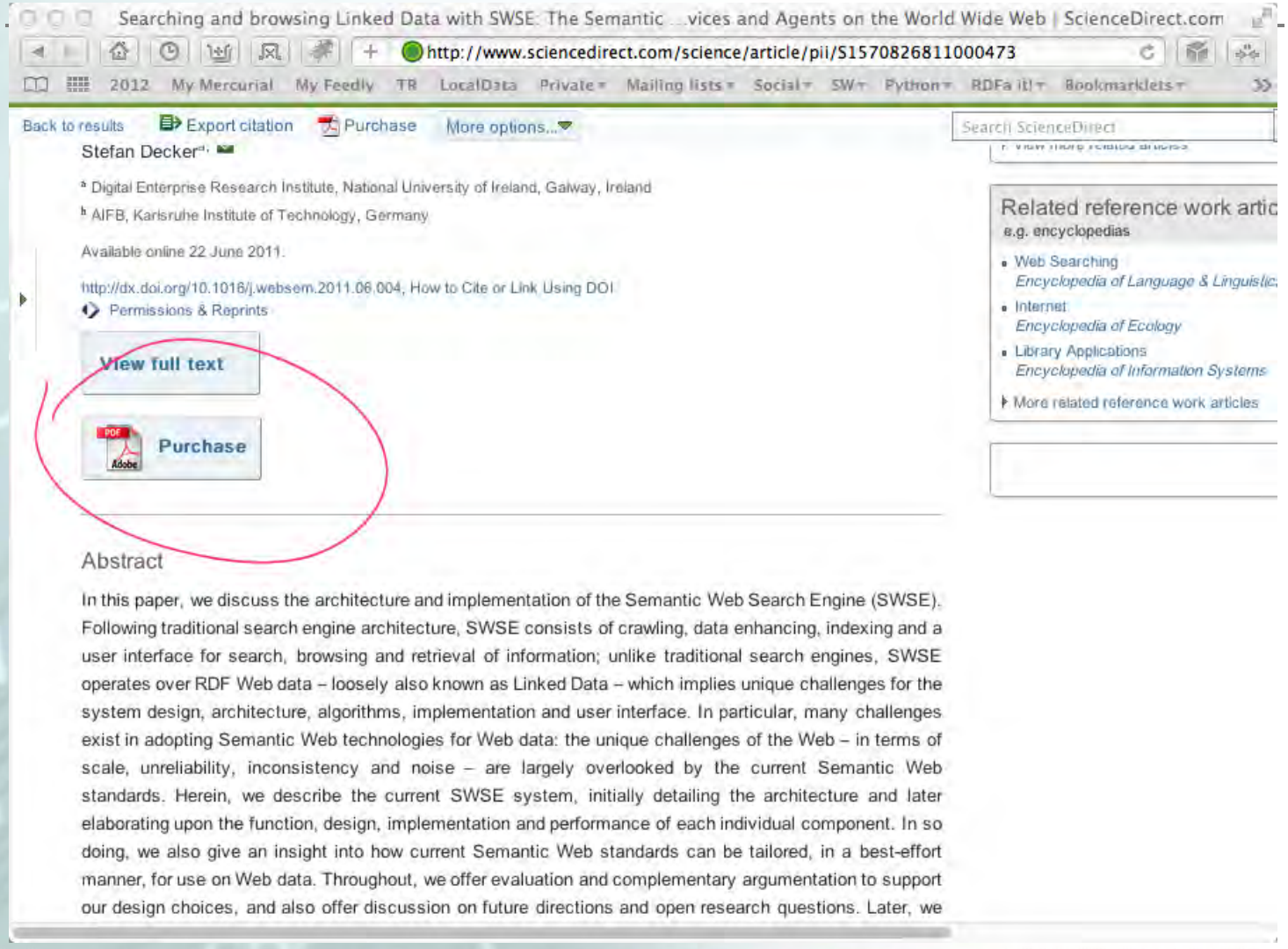
So he looked up the journal on the Web



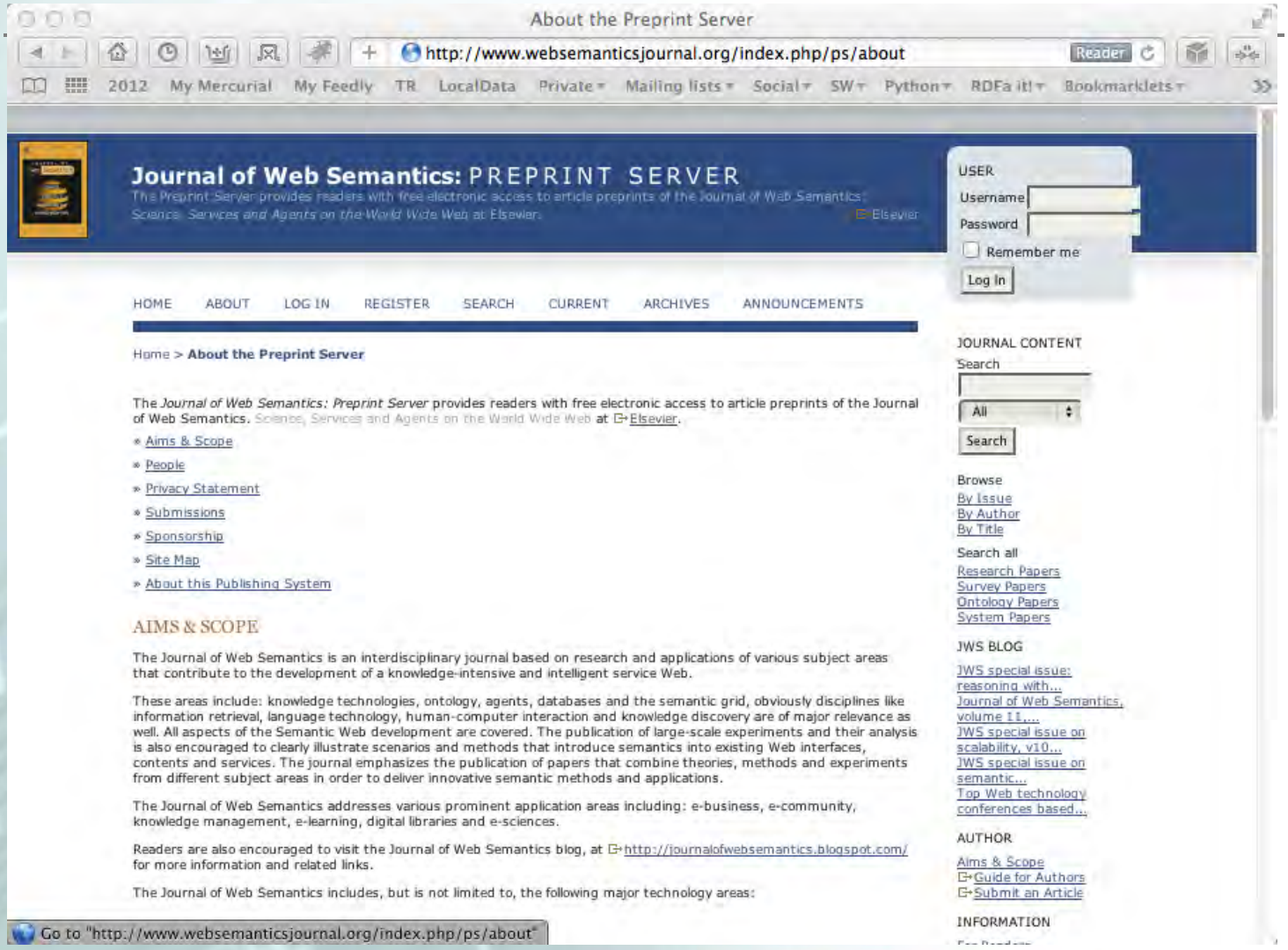
Found the paper



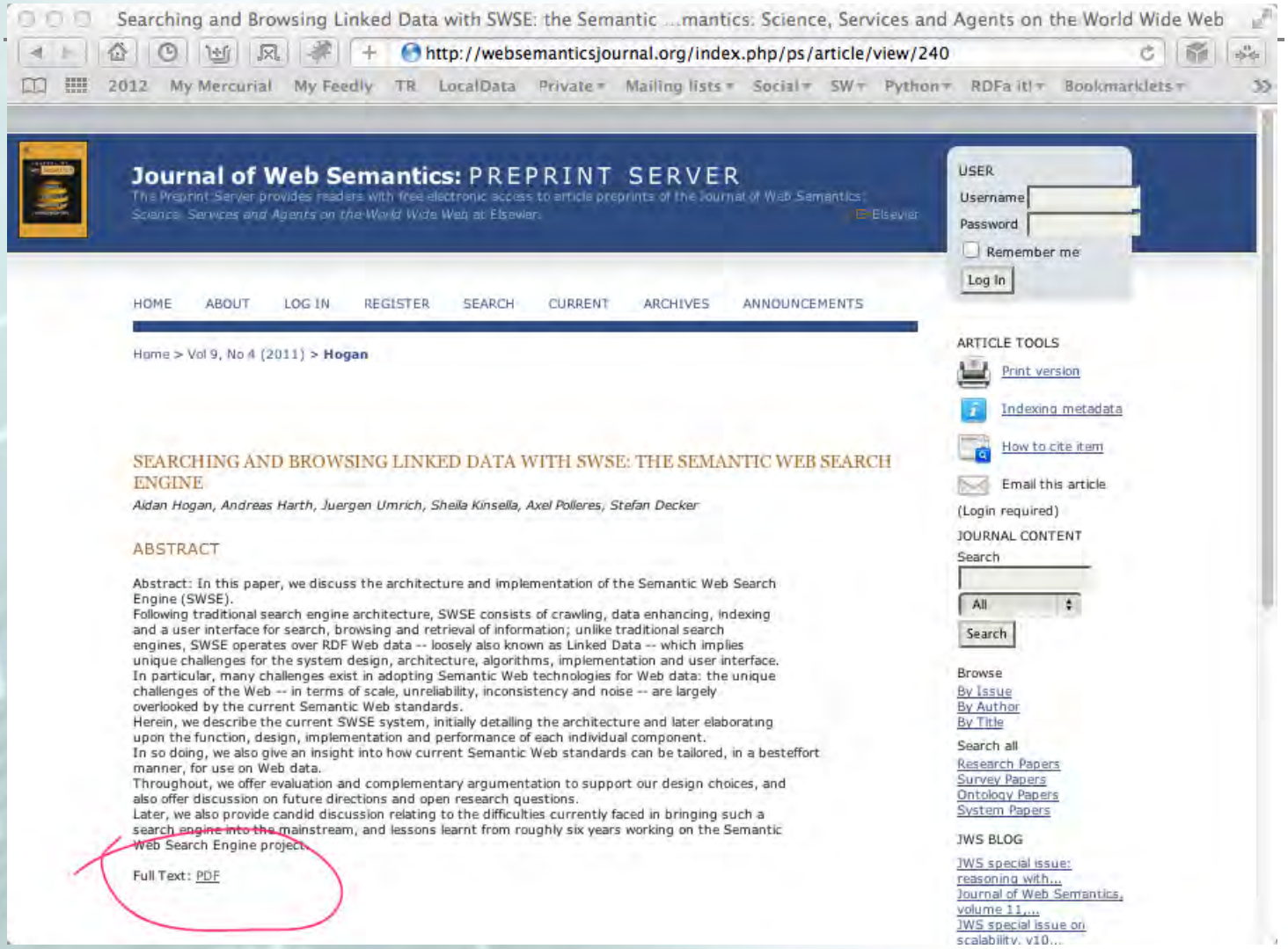
However, he was not at his institute…



But… He knew the secret!



But…He knew the secret!



So he 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 …

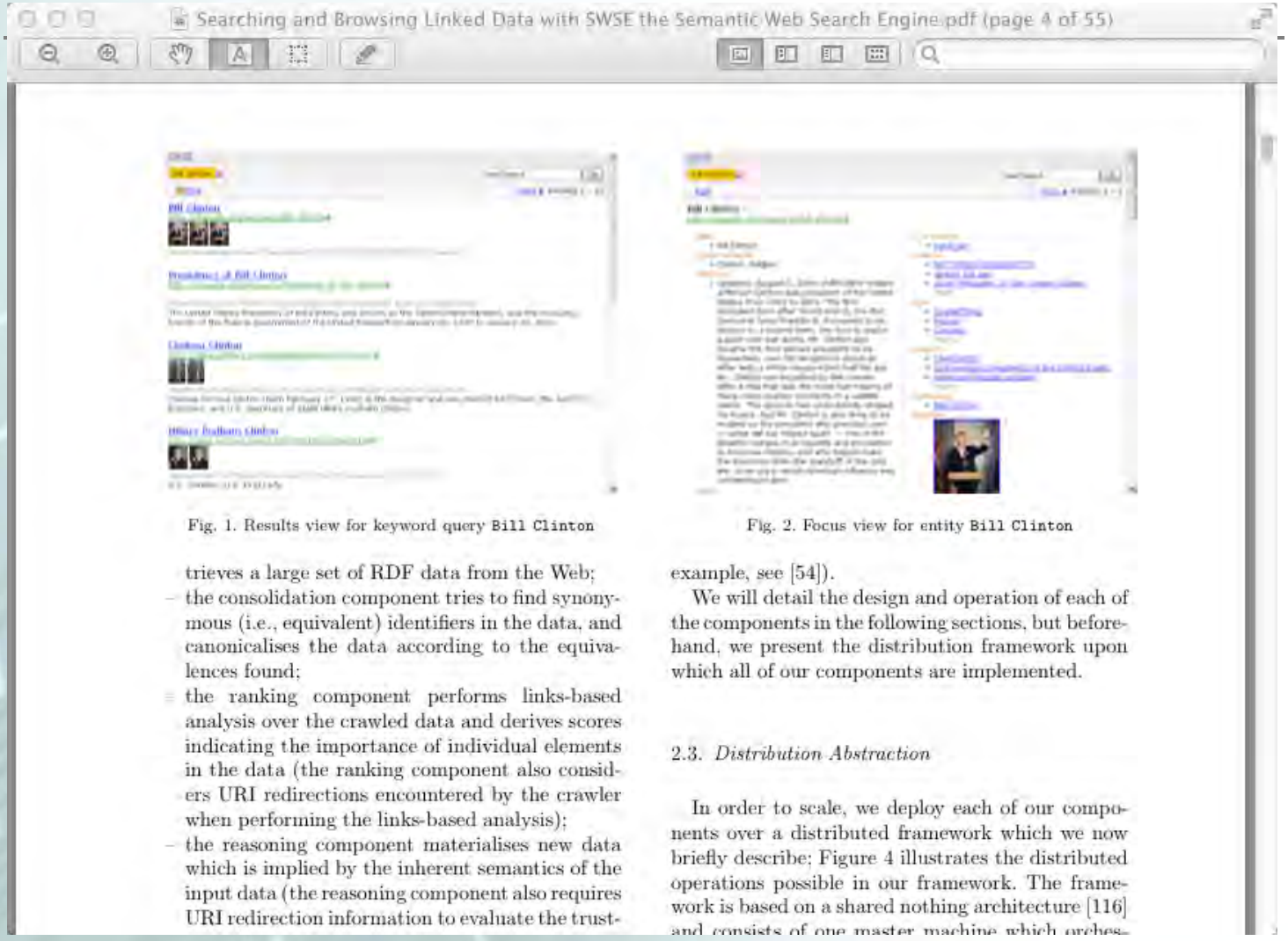


Fig. 1. Results view for keyword query `Bill Clinton`

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-

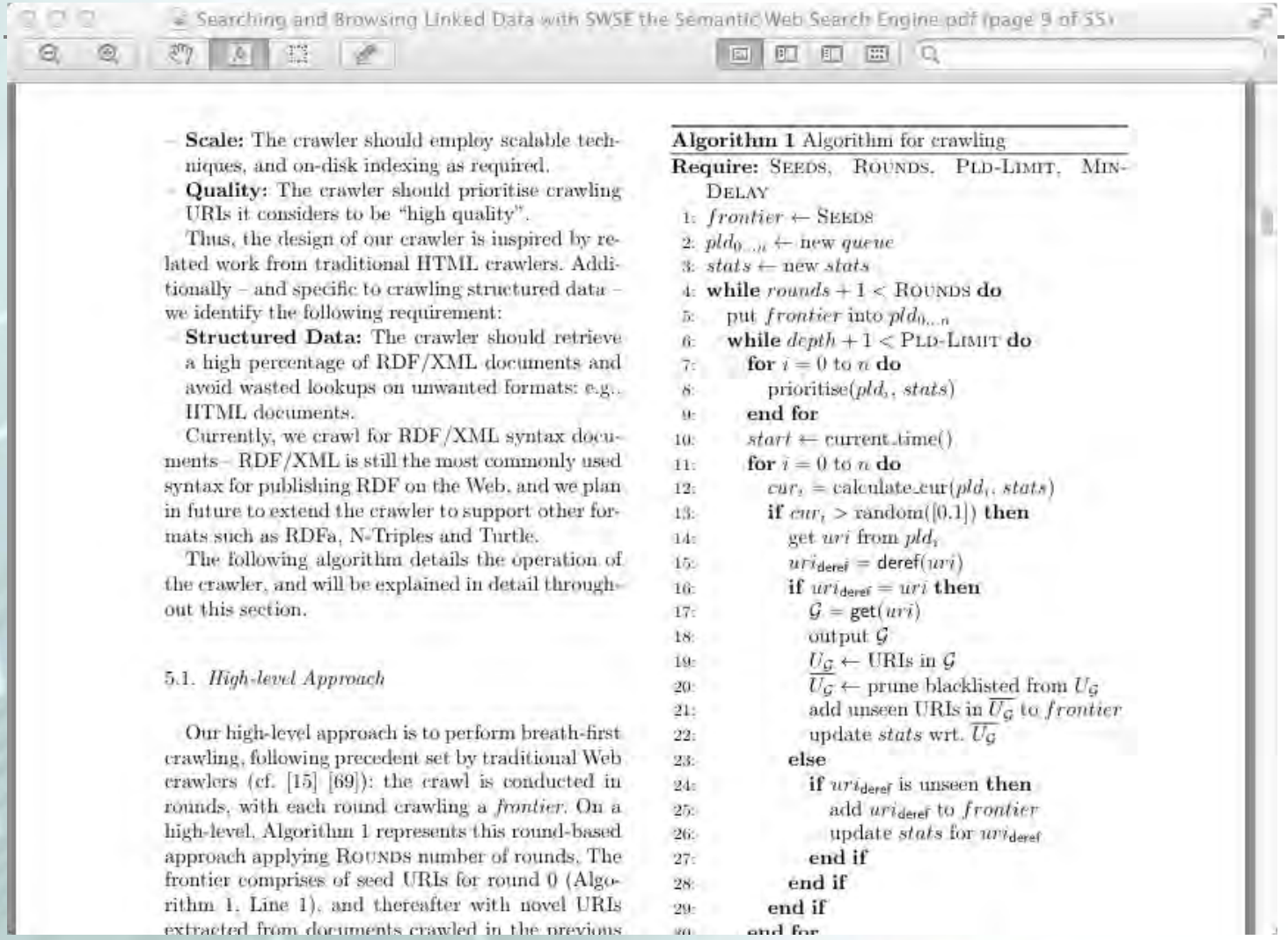
example, see [54]).

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*

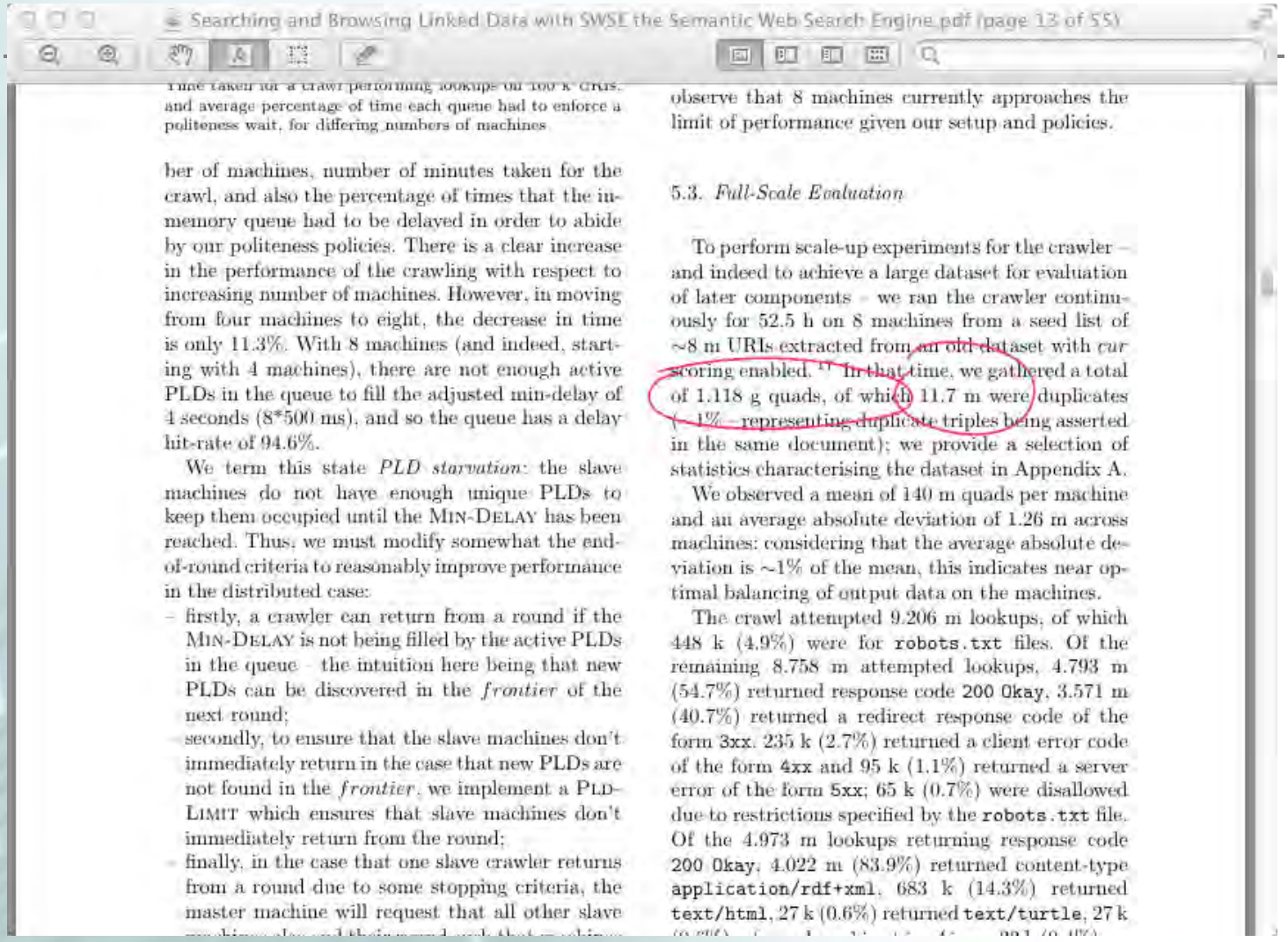
Our high-level approach is to perform breath-first crawling, following precedent set by traditional Web crawlers (cf. [15] [69]): the crawl is conducted in rounds, with each round crawling a *frontier*. On a high-level, Algorithm 1 represents this round-based approach applying ROUNDS number of rounds. 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…



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.

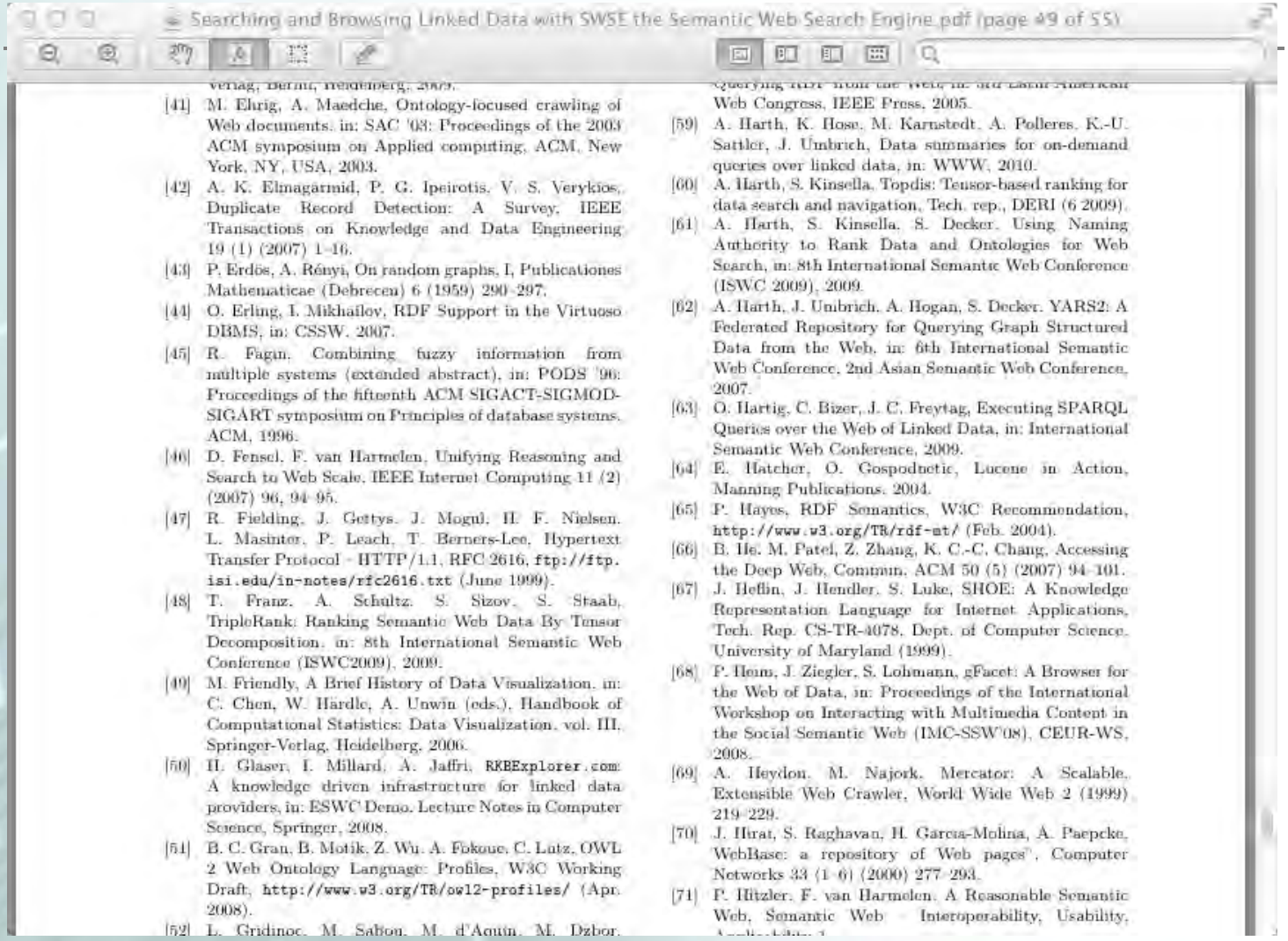
5.3. *Full-Scale Evaluation*

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. Gran, 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: 4th 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).

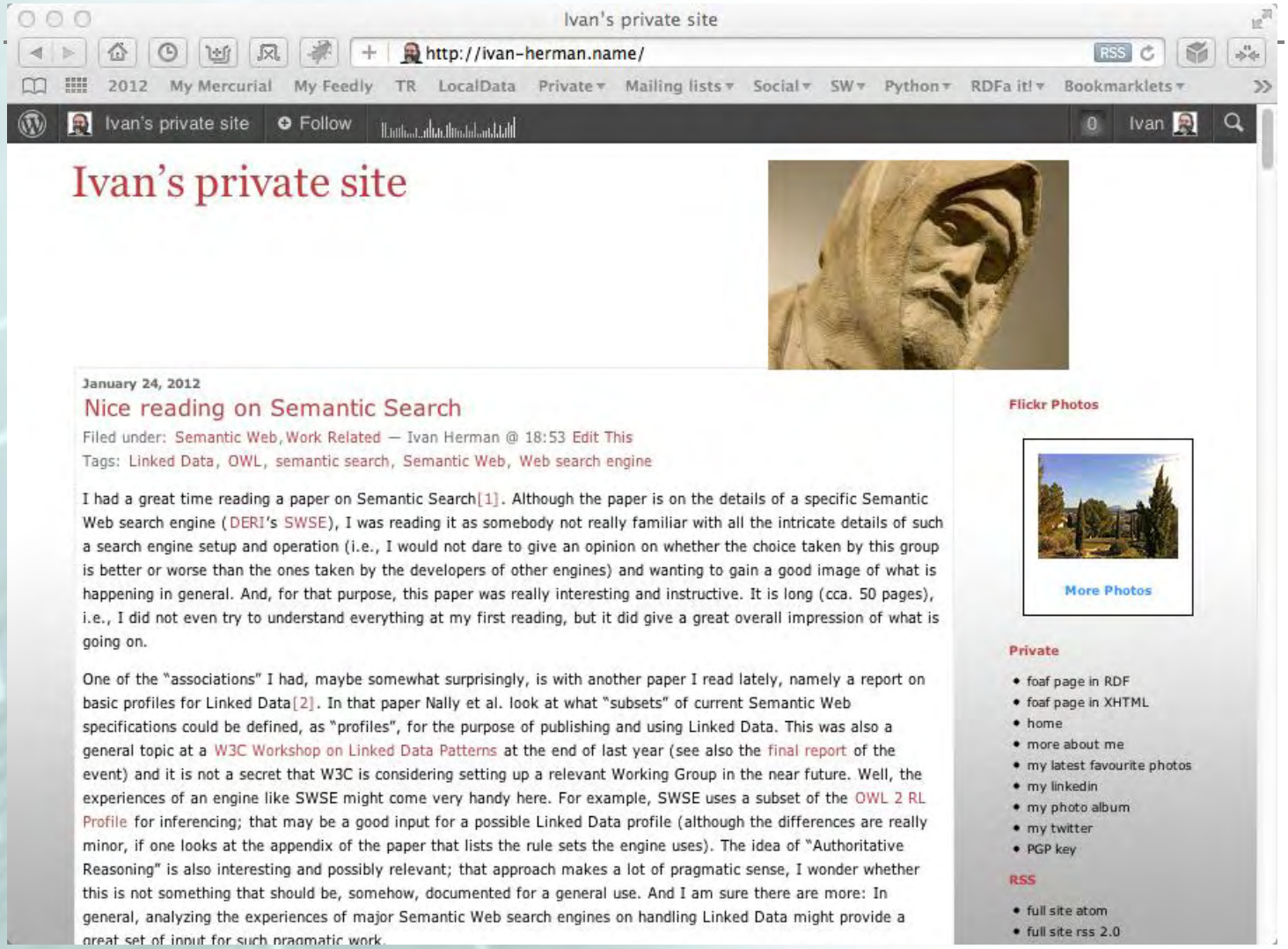
[68] P. Heim, J. Ziegler, S. Lohmann, gFacet: A Browser for the Web of Data, in: Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08), CEUR-WS, 2008.

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

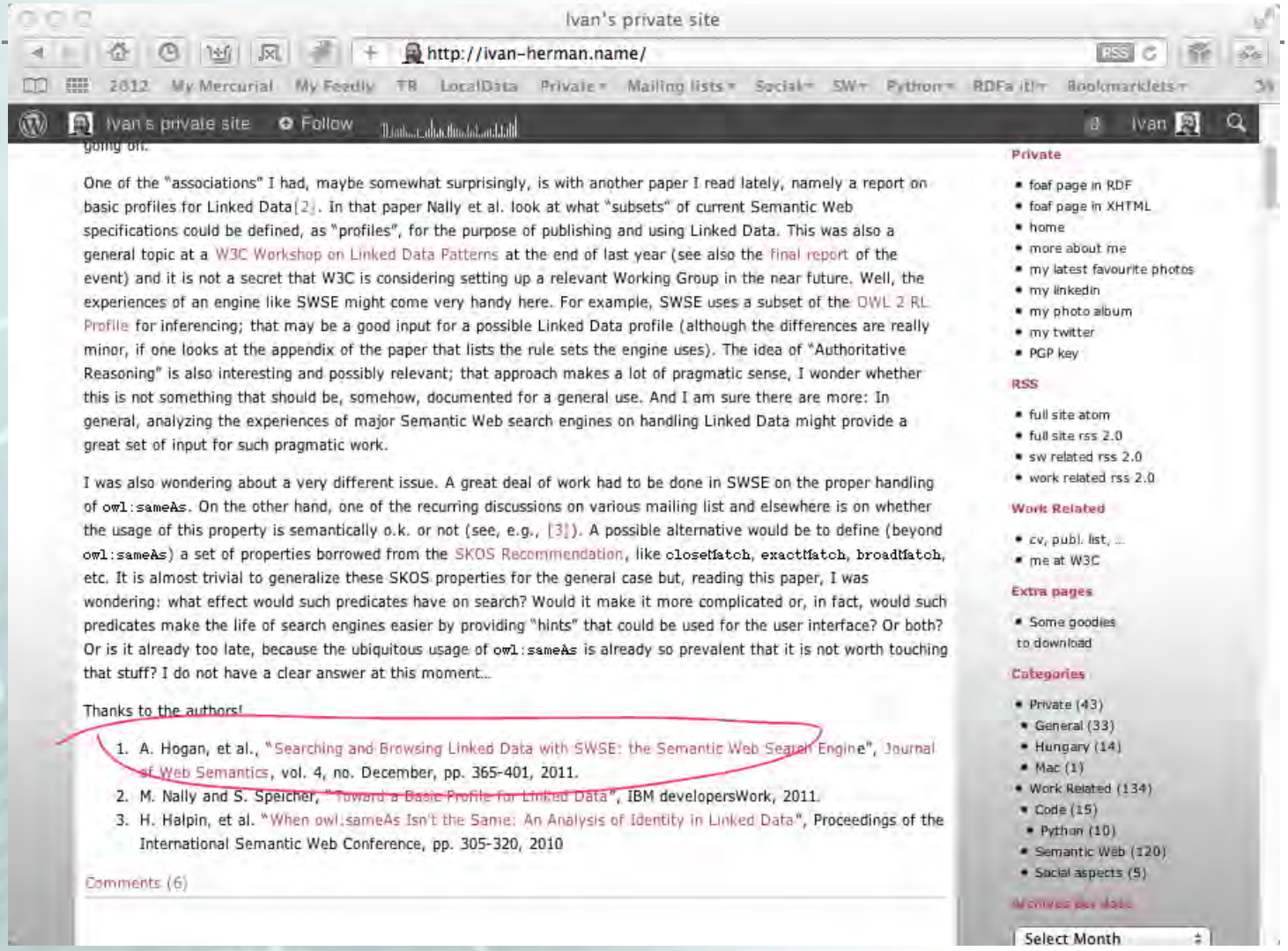
[70] J. Hirai, S. Raghavan, H. Garcia-Molina, A. Paepcke, WebBase: a repository of Web pages', Computer Networks 33 (1–6) (2000) 277–293.

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

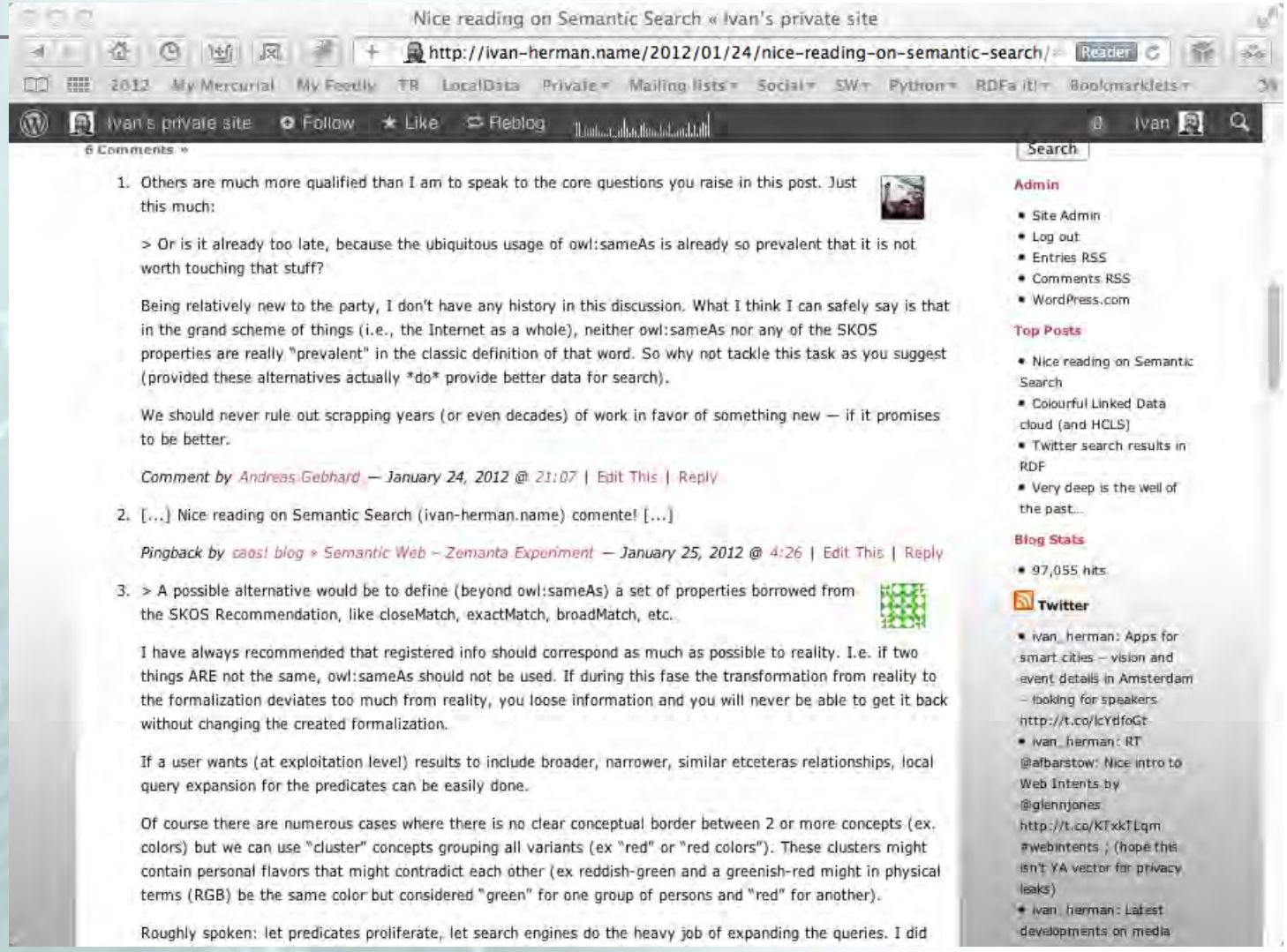
He liked the paper. So he wrote a blog…



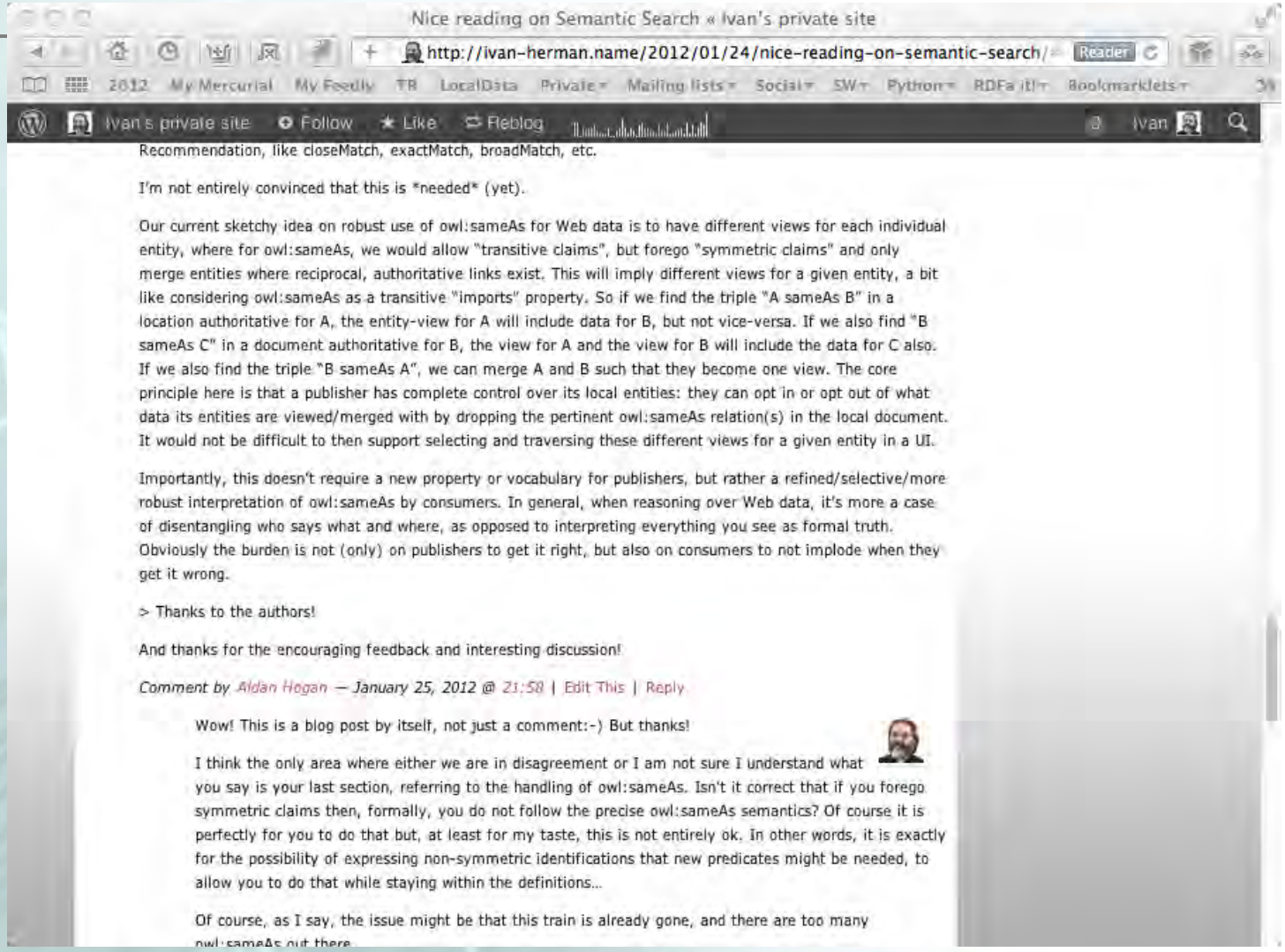
…with exact references.



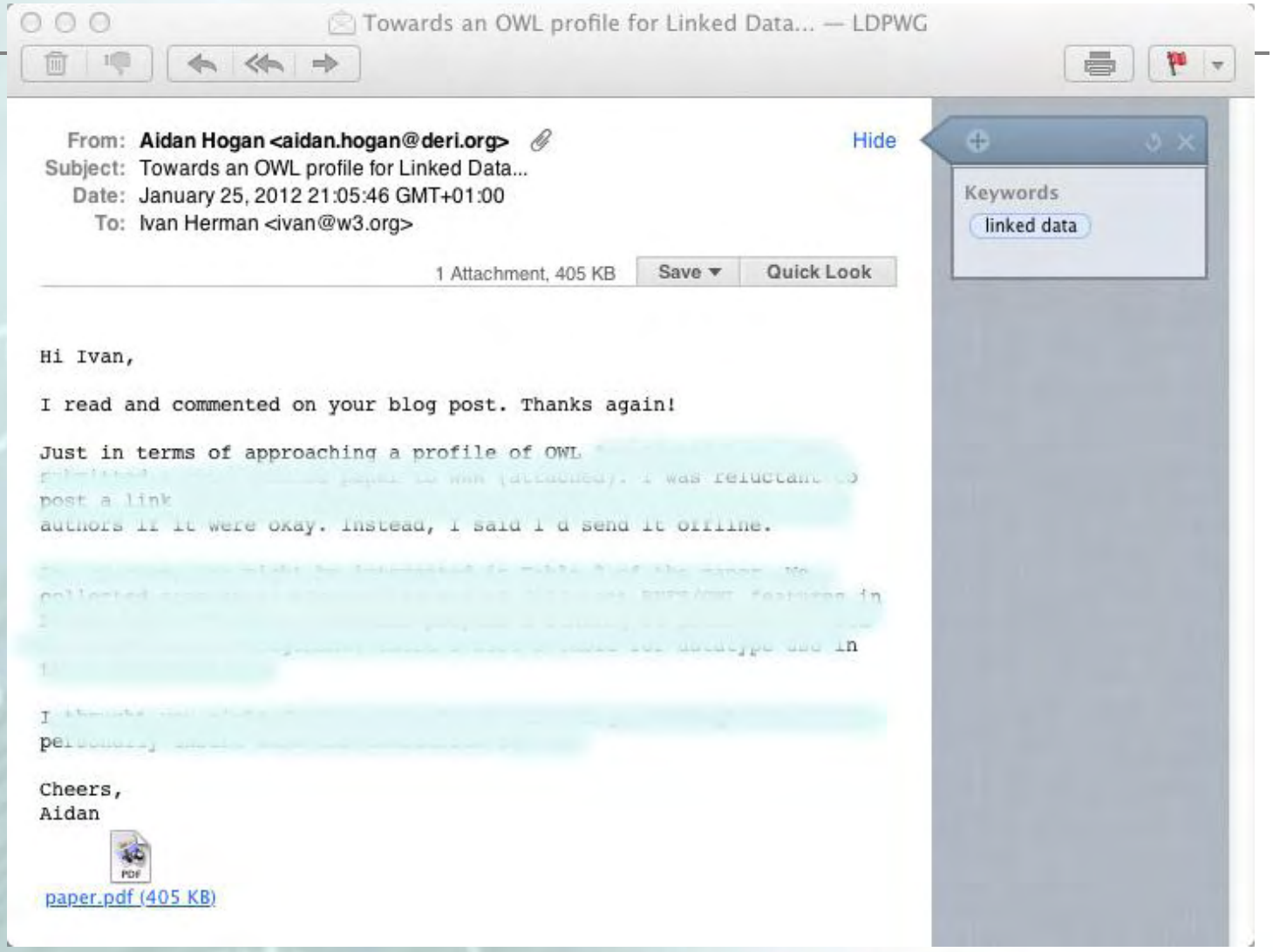
There were a bunch of comments…



…including the author's, with his 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
- they get to know and possibly influence one another's view
  - Etc.

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

Communication was important from the beginning

Denis de Sallo and Journal de Scavans

- Henry Oldensburg and Philosophical Transactions of the Royal Society

What is different today?

Today

- cheap storage
- Web advantages and hyperlinks
- machine readable articles (data mining)

- scholarly journals, on the other hand, are frozen as PDF files (yes, with some links sometimes...)

Changes

- Twitter, Google+, Facebook…as a communication tools should be assessed as an impact
  - Higher pace of information exchange
  - could we justify long publication delay?

Will scholarly communication move away from its paper centric model, and join the information age!

Article of the future

- not only text but:

- datasets, multimedia (video, audio, images) in standardized format
- interactivity with readers
- social network included
- full metadata set (date of application, date of acceptance, date of publishing, full metadata set for references list)
- related work, previous versions of the article
- hyperlinks usage
- re-usable

Publishers of the future

- more flexible subscription models, copyright policies
  - new business models
  - from text to other formats
    - research data!!
    - rich content
  - more long term archiving and preservation activities
- more added value – education, promotion of the new trends

Library of the future

- from archiving and preservation to **access** and **delivery**
  - from collections to user-centric **services**
- not to compete with, but use "big players'" products – Google, Amazon, etc., and learn from them


JOURNAL
YOU WRITE THE PAPERS,
YOU REVIEW THE PAPERS...
WHY SHOULD YOU PAY TO READ THEM ?

Science should be public good

- taxpayers
- created in non-for-profit institutions
- authors and referees don't expect financial rewards

Open access

- the best way to overcome scholarly publishing crisis
  - it's logical
  - it's natural
  - it's promising
- commercial publishers are upset – it should be good

What is open access?

- „By Open Access, we mean the free, immediate, availability on the public Internet of those **works which scholars give to the world without expectation of payment** – permitting any user to read, download, copy, distribute, print, search or link to the full text of these articles, crawl them for indexing, pass them as data to software or use them for any other lawful purpose. ”

The Scholarly Publishing & Academic Research Coalition .- SPARC

WHY?

Benefits

- opening the knowledge base to all - more researchers can build on it and there is less duplication of effort
- researchers can reach a greater audience and find that their work is more widely read and cited
- institutions gain an enhanced reputation as their research becomes more visible
- funding agencies see a greater return on their investment, and publishers find that the impact of their journals increases.

Greater visibility and accessibility, and impact

- according Harnad and Brody study open access doubles downloads and increases citations by an average of around 50% (from around 40% for biology to 250% for physics)
- 20% research in open access – a lot of impact (money invested in R&D) has been lost every year

Rapid and more efficient progress of scholarly research

- an example of arXiv with 768,539 papers
  - 1.5 million of connections per day
- time taken for articles to be cited is shorter every year

HOW?

Two roads

- network of institutional repositories (archives) + metadata exchange = green road
  - journals in open access = golden road

Institutional repositories...

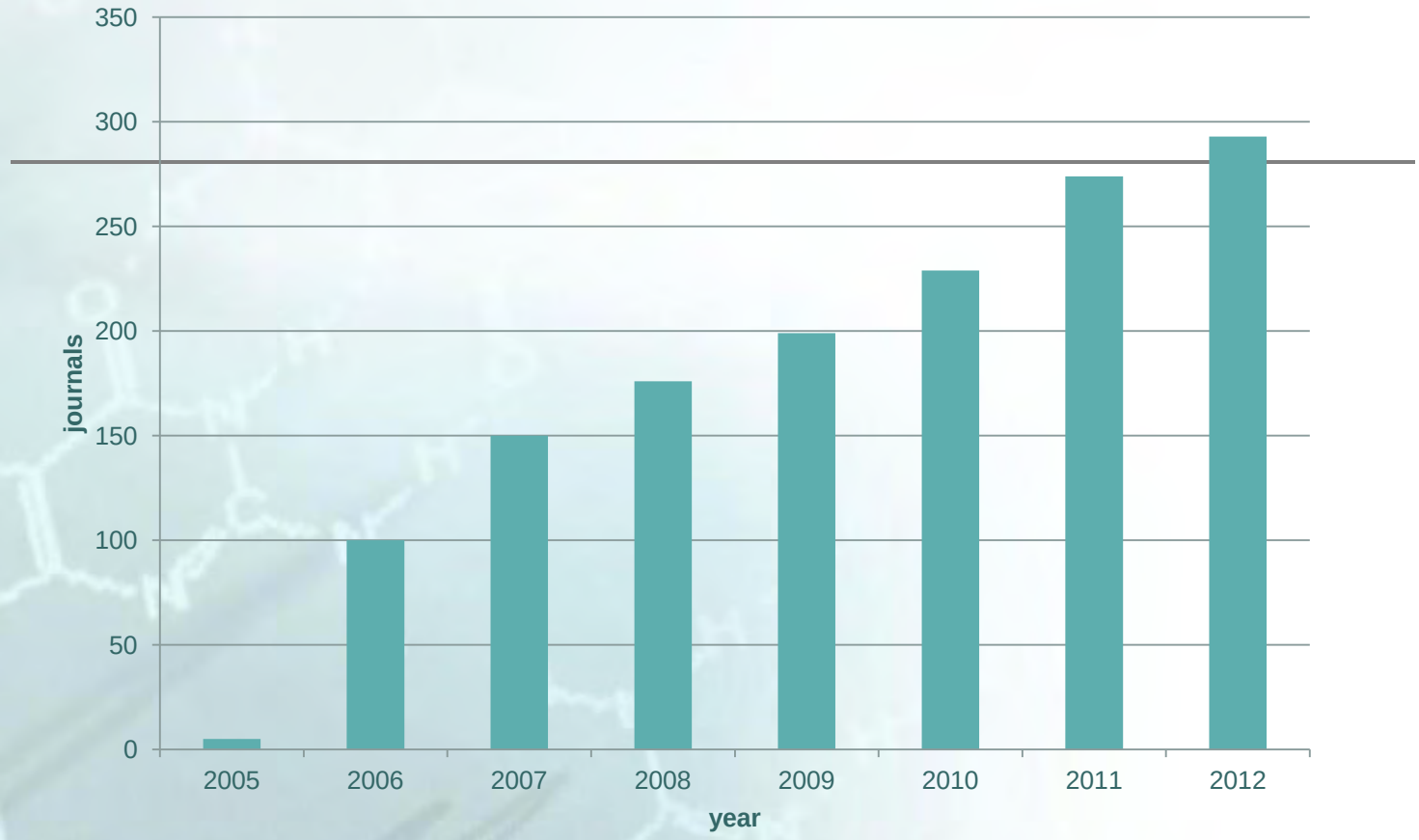
- "digital collections that preserve and provide access the the intellectual output of an institution."
  - selfarchiving
  - mandate

Raym Crow. *The case for institutional repositories: a SPARC position paper*. 2002

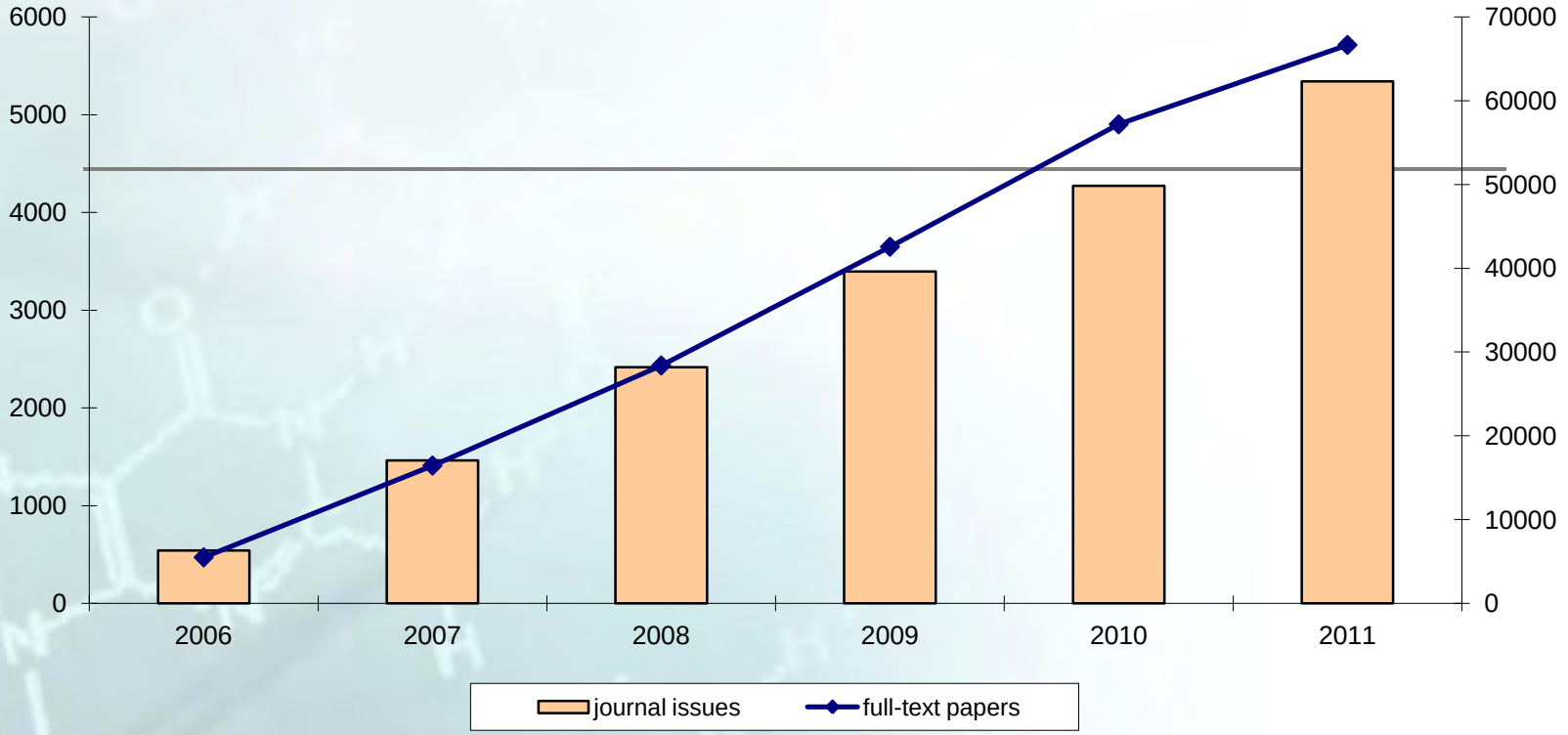
Open access journals

HRČAK – initial goals

- simple way to make online version of the (printed) journal
  - single access point for all Croatian open access journals (scholarly, professional and popular science)
    - metadata and full-text articles repository
- data sharing – with international repositories, databases, archives



number of HRČAK journals by years



Number of journal issues an full-text papers by year

Journal	Inclusion date	Full-text visits	Total visits
Bogoslovska smotra	10.1.2009	373465	715568
Društvena istraživanja	11.4.2006	300454	682859
Ekonomski pregled	6.2.2007	391306	582121
Collegium Antropologicum	10.7.2006	247850	388749
Građevinar	11.1.2007	248931	361871
Acta stomatologica Croatica	7.2.2006	209667	347778
Politička misao	20.2.2008	183359	335451
Narodna umjetnost: hrvatski časopis za etnologiju i folkloristiku	28.4.2006	182708	328320
MEDICUS	11.1.2008	208905	326598
ŽIVOT I ŠKOLA: časopis za teoriju i praksu odgoja i obrazovanja	14.3.2008	248036	320706
Acta Clinica Croatica	10.8.2007	118985	301218

2010-2012

Data harvesting

- OAI-PMH protocol
  - OAIster
  - BASE
  - Google Scholar
  - ...

Harvesters visited HRČAK on average:

- 2508 times per month in 2011
- 3543 times per month during the first four months in 2012

Cooperation with commercial database publishers

- Elsevier/Scopus – 105 Croatian journals
- ThomsonReuters/WoS – 65 Croatian journals
- Google Scholar – all HRČAK journals
- citation impact

Web 2.0 applications

Kemija u industriji, Vol.59 No.4 Travanj 2010.

Stručni rad

From Our Libraries:
Danko Škare (ed.)
Metrics of Scientific Journalism - Truths, Myths and Misconceptions

J. Stojanovski; University of Zadar, Zadar, Hrvatska; Ruđer Bošković Institute, Zagreb, Hrvatska

Puni tekst (Hrvatski) Str. 179 - 186 *(pdf, 1.09 MB) downloads: 482*

Sažetak

The evolution of the digital age has led to the development of so-called abstracting/indexing printed publications in online databases. A short survey of the four most popular databases in Croatia has been done, as Current Contents, Web of Science, Scopus and Google Scholar, with short history, coverage (included Croatian titles), features and possibilities of citation analysis. We discuss their strengths and weaknesses, and the importance of the efficient usage and appropriate interpretation of the data. The journal impact factor and other bibliometric measures are frequently misused to estimate the influence of individual papers and authors. Explaining potential benefits and limitations of bibliometric indicators, this article may help policymakers, librarians, administrators, and individual researchers to use the data provided by different databases more carefully and wisely in the evaluation of research.

Ključne riječi

metrics: scientific journalism; citation analysis; online databases; impact factor

[Hrvatski]

Posjeta: **1031** (od 01.01.2007.)

How many journals are in OA

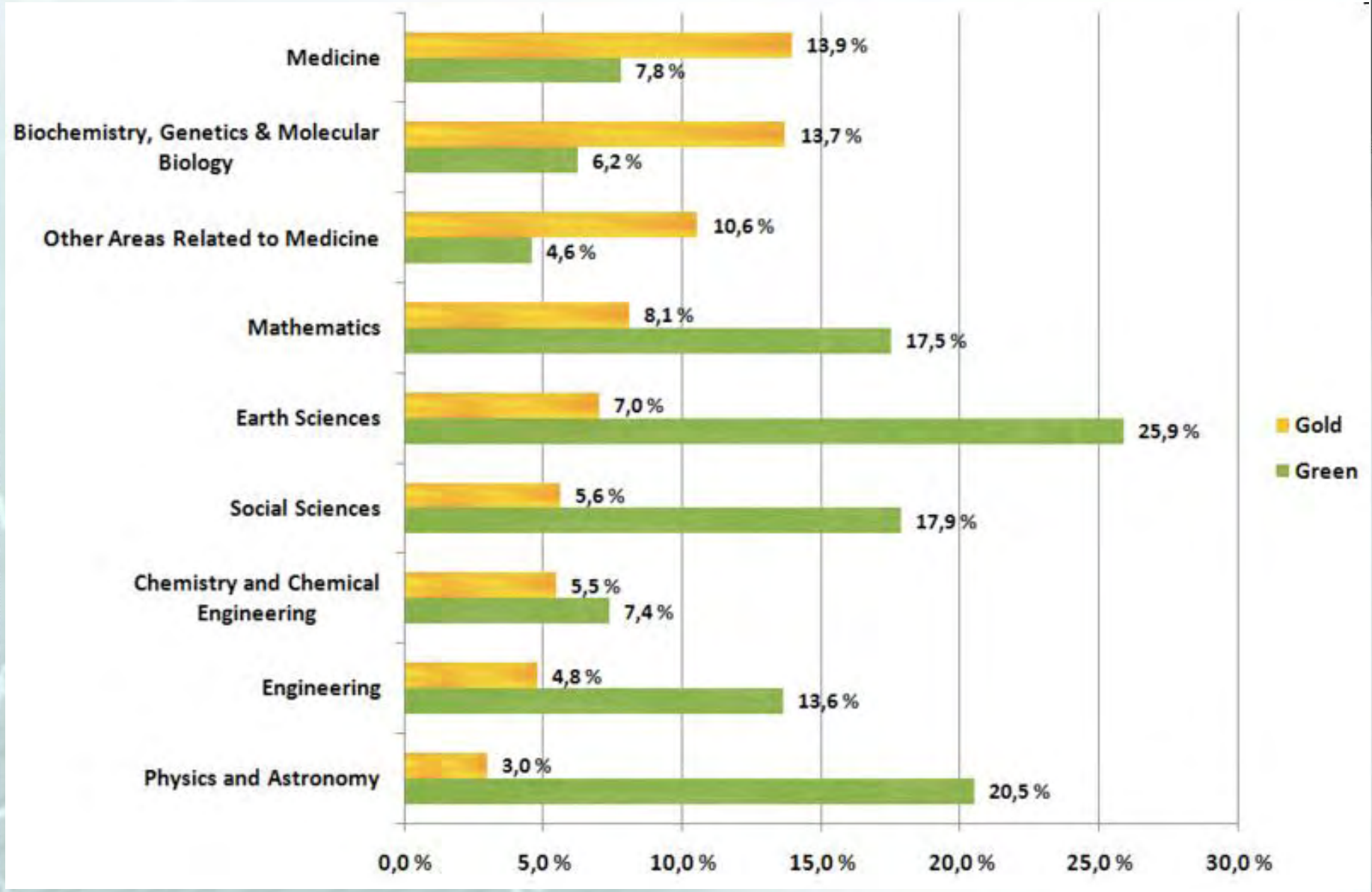
- **8.5%** free available at publishers web sites

  = **golden OA 8.5%**

- additional papers in the institutional repositories and authors' and institutions' web sites **= green OA 11.9%**

  = **20.4 %** annual scholarly output is freely available

Björk, Welling, Laakso, Majlender, Hedlund, & Guðnason (2010)



Björk, Welling, Laakso, Majlender, Hedlund, & Guðnason (2010)

WHERE? EVERYWHERE.

Where to publish or what makes great journals ?

- it is not about technology, big investments, or great promotion .....

- journals are based on the communities they serve
- living organism that rely on the editors, authors and reviewers that make up that community
- no value without the active support of high level scientists
- quality attracts quality

Anne Kitson, Executive Vice President, Health and Medical Sciences, Elsevier: Editor Seminar in Journal Publishing - Attaining Excellence in Scholarly Communication

How do authors choose a journal?

- subject coverage of their research paper and its quality and approach
- selection of the most appropriate journals in terms of subject coverage and readers
- matching the general quality of their paper (best, good, ok) to a class of journals (top, average, low)
- from that class they select a specific journal based upon experience
- recommendation from professor

Anne Kitson, Executive Vice President, Health and Medical Sciences, Elsevier: Editor Seminar in Journal Publishing - Attaining Excellence in Scholarly Communication

How do author choose a journal?

Journal metrics (IF etc.)

Reputation

Editorial standard

Publication speed

Access to audience

International coverage

Self evaluation

A&I coverage

Society link

Anne Kitson, Executive Vice President, Health and Medical Sciences, Elsevier: Editor Seminar in Journal Publishing - Attaining Excellence in Scholarly Communication

Peer Review?

Metrics?

- **existing indicators are often misused**
- **„wrapping paper” is more important than a content**
  - impact factor
    - SJR
    - SNIP
  - eigenfactor
  - h-index
  - article influence
  - number of citations
  - number of papers
  - ...

Possible metrics for an article

- total cites (Google Scholar, Scopus, WoS) – deduplicated
  - # visits
  - # downloads
  - # comments
  - # bookmarks, likes...
  - expert assessments...
  - # discussions (on well known blogs)
  - # appearance in news, blogs, etc.

Download: PDF | Citation | XML

EzReprint New & improved!

Metrics

Total Article Views: 48,368

Citations

CrossRef (14)
PubMed Central (6)
Scopus (15)
Web of Science® (15)

Social Networks

CiteULike (2)
Facebook (282)
Mendeley (19)

More

Related Content

Related Subject Categories

Chemistry, Genetics and Genomics, Molecular Biology

Related Articles on the Web

Google Scholar
PubMed

More

Share this Article

Email this article

OPEN ACCESS

d: The Identification of the Two Missing
ldren Using DNA Analysis

cs | Related Content | Comments: 3

	HTML Page Views	PDF Downloads	XML Downloads	Totals
PLoS	43,318	2,252	74	45,644
PMC	2,081	643	n.a.	2,724

CLE OPEN ACCESS

lved: The Identification of the Two Missing
Children Using DNA Analysis

letrics | Related Content | Comments: 3

mment on this article

As far as I understand, this conclusion in Discussion" is not based on DNA-tests. Probably, this only his opinion.

Posted by BorisRomanov on 02 Mar 2011 at 11:39 GMT

se on 02 Mar 2011 at 11:39 GMT

Additional support to GMI by the FWF Austrian cience Fund

Posted by coblem on 11 Mar 2009 at 22:28 GMT

se on 11 Mar 2009 at 22:28 GMT

Additional support to GMI by the FWF Austrian cience Fund

Posted by coblem on 11 Mar 2009 at 22:28 GMT

se on 11 Mar 2009 at 22:28 GMT

Parts of presentations (with permissions)

- The story from the beginning is from: Ivan Herman: Report on the "Future of Research Communications" Workshop, Dagstuhl, August 15-18, 2012
  - slides 52-54: Anne Kitson, Executive Vice President, Health and Medical Sciences, Elsevier: Editor Seminar in Journal Publishing - Attaining Excellence in Scholarly Communication

ttp://www.iavatuning.com/my-first-question-at-stackoverflow-com/

THANK YOU! ☺