ARIADNE

Issue 66 January 2011

At the Event

Beyond the PDF

Jodi Schneider reports on a three-day workshop about the future of scientific communication, held in San Diego CA, USA, over 19-21 January 2011.

Main Contents | Section Menu | Email Ariadne | Search Ariadne

Introduction

'Beyond the PDF' brought together around 80 people to the University of California San Diego to discuss scholarly communication, primarily in the sciences. The main topic: How can we apply emergent technologies to improve measurably the way that scholarship is conveyed and comprehended? The group included domain scientists, researchers and software developers, librarians, funders, publishers, journal editors - a mix which organiser **Phil Bourne** described as 'visionaries, developers, consumers, and conveyors' of scholarship. The workshop's vision was to identify key issues that must be overcome in order to satisfy consumers, create a plan (including responsible parties, a timeline, and deliverables), and to find a way to keep momentum.

Workshop Structure

The schedule was divided into two equal parts: presentations and working sessions. Momentum started building in advance of the gathering, with an active email discussion list. Based on these discussions, and publicly submitted abstracts, the organisers arranged the presentations into sessions on six key topics:

- Annotation
- Data
- Provenance
- New models
- Writing
- Reviewing and impact

Presentation sessions began with 4-6 short talks, followed by 30 minutes of discussion. I was amazed by the depth of these presentations, despite a tight timeframe of 10 minutes plus 5 minutes for questions for each talk. I was also impressed by the organisers' adeptness in responding to pressing topics and the group discussion by fitting in additional presentations, based on the group discussion and demos. In the second part of the workshop, participants divided themselves into breakout sessions. These groups met twice, and were allocated the bulk of the second day-and-a-half of the workshop. This worked towards the group's goal of producing deliverables, rather than a white paper.

Discussions, Backchannel, and Archives

Throughout the workshop, discussions proceeded during talks and in the dedicated discussion session through a Twitter backchannel (#beyondthepdf), which added to and also diverged from the voiced discussion. Because of reliable streaming and the use of a backchannel, offsite participation was feasible, and **Jason Priem** was a very active participant without travelling from the University of North Carolina. Much workshop material is available at or via the workshop Web site [1], including video archives [2].

Figure 1: Even in January, lunch outside is a good option in San Diego! Courtesy of Phil Bourne

Day 1: 19 January 2011

Introductory Talks from Session Moderators

The workshop opened with an orientation from Phil Bourne and introductions to each topic from the session moderators.

Annotation: Ed Hovy, ISI

Ed Hovy discussed the open issues in annotation, which he sees as central to the future of scholarly communication, for extracting and systematic organising information:

> *The paper of the future lives within a cocoon of data, annotations by various people, a social network of authors, another social network of relevant papers.*

He hoped that publishers could be paid to manage and systemise this, rather than 'to sit on top of our PDF with copyright'.

Data: Gully Burns, ISI

Gully Burns gave an example of blatant misuse of PDF as place for dumping supplemental data. He showed a screen capture of a page from a 40-page PDF of gene expression data. The font was so small that the images were only readable under x1200 magnification. To move data 'beyond the PDF,' he said we needed effective standards for interchanging data and terminology, better integration and coupling between data and the publication process, and automation for data processing and sharing.

Provenance: Paul Groth, VU University of Amsterdam

Paul Groth asked how far we want to get to full reproducibility of publications, raising questions about integrating experiments with publications through workflows and computable papers and drawing attention to the difference between reproducibility and reusability. Currently, he says, we have workflow systems to capture provenance (such as Wings and dexy), models for representing provenance (such as SWAN and OPM), and systems that integrate data into papers (such as GenePattern). Yet to enable computable papers and reproducible papers, we still need to determine best practices for connecting both the

captured data and its provenance to papers. He also drew attention to the recent W3C report on provenance [3] as a driver of future work for data provenance in the sciences.

New Models: Anita de Waard, Elsevier

Anita de Waard gave a historical perspective on new models in publishing, looking back 10 years to 2001 and 20 years to 1991, as shown in Figure 2. Despite the changes in the past 20 years, there's still a lot to be done.

New Models Then And Now

Issue	1991	2001	2011
Article Format	Modular papers?	Semantic papers?	Modular/semantic? W3C ORB; Nanopubs
Business Models	Everything will be free?	Author-pays model, for some journals	Author-pays model, for some journals
Research Data	Almost solved in astronomy	Solved in astronomy	Solved in astronomy! Data-driven papers?
Databases	Curators need help	Curators need help	Curators still need some help
Authoring tools	Need something besides Word + WordPerfect	Need something besides Word	Need something besides Word!
Annotation tools	Coming soon	Coming soon: Annotea	Errr… PDF?
Reviewing tools	Open Peer review coming soon	Open Peer review experiments galore	Errr… let's use a Wiki?
Search	Personal scientific search environment imminent	Semantic Desktop imminent	Errr… Google?
Interactive Math	LaTeX	MathML – just needs to be implemented	MathML – just needs to be implemented…
Chemistry	Almost solved (PMR)	Solved (PMR)	Still solved!

Figure 2: Plus ça change, plus c'est la même chose. Courtesy of Anita de Waard

Writing: Phil Bourne, UCSD

Phil asked three questions about writing: How should we write to be reproducible? What structure should the documents we write have? What tools should we write with in the future?

Reviewing and Impact: Cameron Neylon, Science and Technology Facilities Council

Cameron gave a provocative talk showing the irony in our widespread belief that 'science works because of peer review' even though we know of many flaws in the peer-review process. He condemned artificial scarcity, gave counter-examples which demonstrated that peer review does not guarantee technical quality and does not assess the importance or impact of work. Nor, Cameron reminded us, does it support discovery. In place of traditional, pre-publication peer review, Cameron argued for post-publication peer review.

Sessions

Annotation, Data, and Provenance were the topics for Day 1. In the annotation session, I was most struck by **Maryann Martone**'s talk on *Intelligent Handling of Resources for Human and Machines*, about the practical issues of integration and metasearch, drawing from her work at the Neuroscience Information Framework (NIF). Even having followed developments in scholarly communication closely over the past 3-4 years, I ran into many exciting projects that were new to me; so I was perversely pleased that professional curators are struggling with the overflow, too. Maryann said that even after four years of intensive searching, the NIF curators were still finding large caches of tools of which they were unaware.

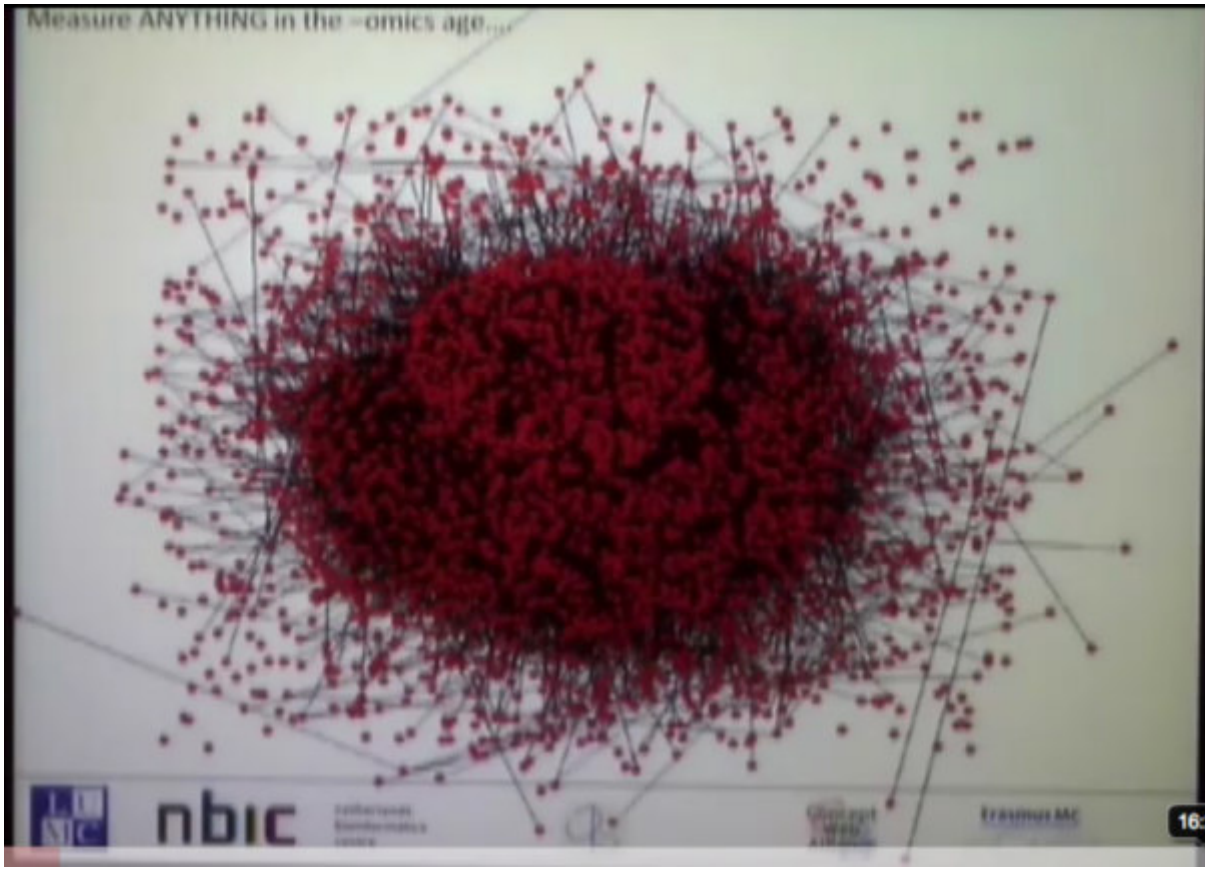

Figure 3: Barend Mons' depiction of 'hairball' of genes

In the Data session, **Barend Mons** asked, 'Is the (enriched) paper computer-reasonable?' For his work, there are too many papers to read: 90 papers on each of 200 genes with a 'hairball' graph of all genes that might be involved. Rather than this 'bignorance-driven research', he wants a semantic wiki where concepts have Universally Unique Identifiers (UUIDs). These concepts are envisioned as nanopublications with subject-predicate-object. Each nanopublication has been stated thousands of times on the Web. They simplify this multiplicity with a single UUID identifying the nanopublication, and calculate the 'Evidence Factor' with the evidence for and against the concept. **John Kunze** spoke about 'baby steps to data publication', envisioning a new standardised format of a 'data paper' as a cover sheet (with title, data, authors, abstract, and persistent identifier) with a set of links to archived artefacts. This would expose datasets to existing indexing engines and Google Scholar, facilitate citation of the data, and instill confidence in the data identifiers (since they are backed up by metadata). With incremental additional work, peer-reviewed and non-peer-reviewed overlay journals could publish data papers to publicise data that is scattered in distributed repositories.

In the Provenance session, **Juliana Freire** argued for provenance-rich, executable papers, which she said would lead to higher-quality publications that present more of the discovery process while supporting reviewers and would-be reproducers. She showed an example of an arXiv paper with 'deep captions' where results link to the experiment details in the VisTrails workflow-based visualisation tool. **Yolanda Gil** talked about enhancing reproducibility with semantic workflows and provenance of the software pipeline. Even with the data made available, and even when someone is willing to invest months of effort into reproducing a paper, it may be impossible. **Nick Encina**, who had attracted interest in the demo session, presented Wingu, a cloud-based startup application that allows display of science workflows and collaborative online editing of papers.

Day 2: 20 January 2011

Sessions

New models, Writing, and Reviewing and Impact were the topics for day 2.

New models was one of my favourite sessions. It included a great demo from **David Argue** of their prototype [4] of a next-generation paper format, which uses the idea of the view of the Model-View-Controller model, and a fun trailer 'Beyond the PDF-The Horror Movie' from Utopia's **Steve Pettifer**. There were also talks on Knowledge Engineering from Experimental Design (KEfED), an unmissable talk from the highly regarded **Peter Murray-Rust**, and a talk on 'Open access in developing countries'.

However, I was especially struck by **Michael Kurtz'** presentation on the NASA Astrophysics Data System (ADS), which focused on what has changed in the past 30 years, and how astronomy communicates. The costs involved in the various services were particularly fascinating, as was the history of ADS, which provides free universal access to the relevant literature. I wish that more disciplines would work together with publishers and funders to adopt this sensible approach!

In the Writing session, **Martin Fenner** and **Cameron Neylon** advocated HTML in their talk 'Blogging beyond the PDF or Copy by Reference', arguing that the best tool is one that is widely used. WordPress and its plugin architecture, and the new possibilities of HTML5, were particularly highlighted. **Michael Reich** described the challenges in reproducing one's own work, which motivated the creation of their reproducible researcher environment, GenePattern. GenePattern workflows can be now be embedded into Microsoft Word documents, to create a reproducible research document, allowing readers to reload and re-run a GenePattern analysis from within a document.

In the Reviewing and Impact session, **Anita de Waard** gave a very cogent talk about how, through hedging, claims become treated as facts. Hedging expresses uncertainty, with phrases such as 'suggest that' or 'imply that', yet citations remove this uncertainty, interpreting and solidifying these claims into generally accepted facts which 'enter the canon of knowledge'. She suggests that we develop systems that enforce 'legal lifting' to ensure that the experimental context and parameters are cited along with the claim. **Paul Groth** suggested aggregating alternative metrics-such as downloads, slide reuse, data citation, and Youtube views - to show impact. He showed **Jason Priem**'s mockup of what such a filtering system could look like, while noting that obtaining metrics from the ever-growing array of social media services is currently difficult.

A Lunchtime Treat

NASA executives were visiting the California Institute for Telecommunications and Information Technology (CalIT2) building, and we were invited to sit in on a video demo. With impressive network bandwidth and incredible video technology, they can broadcast in enough detail for this near-sighted viewer to pick out individual faces in a milling crowd from the back of a large lecture hall.

Breakout Sessions I

After lunch, breakout sessions were formed. We decided on four main areas: research objects of the future, writing and reading, business rights and intellectual property rights (IPR), and attribution, evaluation, & archiving. A draft final report of the workshop [5], linked from the workshop home page, summarises the work of each group.

Day 3: 21 January 2011

The last day of the workshop was devoted to the breakout topics. We began the day with a summary of the previous day's breakout sessions. Participants were invited to change sessions for the second day, or even be 'butterflies' moving between several groups.

Breakout Sessions 2

Three of the four groups left with specific deliverables. I can speak mainly about the writing and reading group, which discussed a system for collecting up and processing research data and files. The group plans to stitch together various storage utilities (e.g. Dropbox, Git) can draw in part from the work of **Peter Sefton**'s group on the 'Fascinator'. Prototype testers are being sought, and the group hoped such a system could help keep to retain organisational knowledge often lost through transitions such as students graduating or post-docs leaving a group.

Figure 4: The pool was closed, but what a view! Photo courtesy of Monica Duke

Outcomes and Future Planning

The workshop succeeded in gathering a diverse international group of participants and raised awareness and energy around scholarly communication. Vivid conversations are continuing on the email list and various social media sites. Smaller groups are collaborating more closely based on the face-to-face discussions during the event. Even with the partial and incomplete view I have of what's going on, I can see that the workshop produced a major impact.

The group is coming to a consensus on the 'La Jolla Manifesto'. An international Hackfest on Scholarly HTML, following from the 'reading and writing' breakout, is being planned for March. Beyond the PDF participants are maintaining a calendar of interesting conferences and events and collecting resources and tools lists [6]. EPUB is getting wider discussion as a format for scholarship. **Martin Fenner** has convened an email list about WordPress for scientists. The W3C Health Care Life Sciences group has taken up the Research Object of the Future discussion.

Many people are focusing attention on Spinal Muscular Atrophy (SMA); opening the literature has been one area of discussion, and even before **Maryann Martone** started her sabbatical at the SMA Foundation, there were concrete suggestions for an impact dashboard for started her sabbatical at the SMA Foundation, there were concrete suggestions for an impact dashboard for SMA, and an international team is developing a prototype system to help the SMA Foundation access scientific and medical knowledge better and faster via text mining and claim detection on the literature.

There has also been some engagement from the standards community; on the last day one discussion comment was that a new standard, PDF/E, is being developed for PDFs for the engineering and scientific community.

Conclusion

'Beyond the PDF' was an energising gathering for taking the next steps towards changing the way we convey and communicate science. I was delighted to put faces to names for many people I had been following online. This workshop was filled with interesting (and concise!) talks; I highly recommend dipping into the entire YouTube playlist!

References

1. Beyond the PDF Web site http://sites.google.com/site/beyondthepdf/
2. Beyond the PDF Video Proceedings at YouTube Web site http://www.youtube.com/view_play_list?p=BE627F48A0DB94FD
3. W3C Incubator Group, Provenance XG Final Report, 8 December 2010 http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/
4. David Argue, Next-generation-paper demo http://www.zfishbook.org/NGP/
5. Phil Bourne, 'Beyond the PDF' Draft Report, 30 January 2011 https://docs.google.com/document/d/1ZPkFvUxC94o4ekLvJwTIpingYm-7mBjnfoh89q-2vSI/edit?hl=en&authkey=CKGC5JML
6. Resources and tools http://neurolex.org/wiki/Category:Beyond_the_pdf

Author Details

Jodi Schneider
Ph.D. Student
Digital Enterprise Research Institute (DERI)
National University of Ireland

Email: jschneider@pobox.com
Web site: http://jodischneider.com/jodi.html

Jodi Schneider is a second-year Ph.D. student at the Digital Enterprise Research Institute (DERI), NUI Galway, Ireland. She holds an M.A. in mathematics, an M.S. in Library and Information Science. Her research interests are in argumentation, scientific and scholarly communication, and the Social Semantic Web. Before joining DERI, Jodi founded an open access journal for library technologists (Code4Lib Journal), was community liaison for the research summary wiki AcaWiki, and worked in academic libraries. At DERI, her current research is on argumentation on the Social Semantic Web, and she serves on W3C groups on Scientific Discourse in biosciences and Library Linked Data. At Beyond the PDF, she spoke on 'Supporting Reading'.

Return to top

Article Title: "Beyond the PDF"
Author: Jodi Schneider
Publication Date: 30-January-2011 Publication: Ariadne Issue 66
Originating URL: http://www.ariadne.ac.uk/issue66/beyond-pdf-rpt/

Main Contents | Section Menu | Email Ariadne | Search Ariadne

Ariadne is published every three months by UKOLN. UKOLN is funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.