The Future of the Journal?

Integrating research data with scientific discourse[1]

Anita de Waard

Anita de Waard has a background in experimental physics. She joined Elsevier as publisher in physics and neurology in 1988, and since 1997 has been employed as a Disruptive Technologies Director within the Elsevier Labs group. Her main focus is the development of innovative product concepts, with a specific interest in establishing collaborations between Elsevier and academic groups in information and computer science. Her interests include the application of Semantic Web technologies for scientific communication, and the development of a new, semantic form for the scientific article. She developed and led the Elsevier Grand Challenge for Life Sciences and the Killer App Award, both rewarding researchers for ideas pertaining to novel forms of science publishing. Other projects include running the W3C HCLS SiG Subtask on Rhetorical Document Structure, collaborations on representing scientific documents as hypotheses and evidence, and running a number of workshops that provide a platform enunciating the key possibilities for and defining the main impediments to changing scientific communications. From January 2006 onwards, de Waard has been working as a part-time researcher at the University of Utrecht, funded by a Casimir project grant by the Netherlands Organisation for Scientific Research. Her research focuses on discourse analysis of biological text, with an emphasis on finding key rhetorical components, and offers possible applications in the fields of hypothesis detection and automated copy editing.

E-mail: a.dewaard@elsevier.com
Website: http://elsatglabs.com/labs/anita

DOI: 10.1163/095796510X546878

Science is done using artifacts, elements of information that are created, shared, converted, concatenated, compared, and commented upon. For example: a problem in neuroscience is studied by raising and breeding a particular set of rats. Then, either a lesion or an injection is performed whereby a part of the rats' brain is damaged, and their behavior, ability to resist disease, or longevity is studied. All of the steps to perform this experiment are catalogued, recorded: the rats themselves and all of the conditions they are subjected to are numbered, described, and somehow stored. These measurements are then analyzed, interpreted, perhaps analyzed again. A new experiment is run; the results of the two are compared. Throughout this process, the various people involved in the experiment (researchers, analysts, managers, students, even cleaners (through the ubiquitous DO NOT SWITCH OFF! notes found in labs across the planet) communicate with each other: through email, at whiteboards, in meetings, via Skype, telephone or wikis: plans are forged, results are shared, thoughts are formulated.

After some time, a conclusion is reached: enough to publish as a paper. This story gets written, drafted, shared, edited; references are found and added; graphics are created and fitted in; the manuscript gets submitted to a conference, a journal. Editors acknowledge receipt, reviewers write reports, authors respond and amend their manuscript, the thing gets accepted and sent to a publisher. Now graphics need to be tweaked, words taken out, references stylized to fit an idiosyncratic journal format. The paper gets marked up in XML by typesetters in Manila, shipped to the Electronic Warehouse in Amsterdam, served up to the XML Content Server in Dayton, rendered into html (this part is invisible to the scientist) and appears in the journal: a nearly entirely electronic process, at the end of which the author has a link to a PDF to add to his or her name. A year, six months of science, that eventually results in a single DOI. But we are not done yet. The paper (hopefully) gets read, commented upon; perhaps a PowerPoint presentation is made with some of its figures; other scientists read it, they comment upon it in their blog or email, the paper gets cited. The main claims get reformulated as reference gets made, and appended to the DOI, to the author's H-Index. Perhaps the main points in the paper now get curated into a database; figures or phrases get used in a textbook. What was once a small thought in a lab, based on a set of rat behaviors following a particular injection of a chemical, is now an element in the canon of neuroscientific fact. Knowledge has been made.

And on the face of it everything sort of works. Papers get published, and read. People cite and read each other's work, and collectively build a temple of knowledge, brick by conceptual brick. But this system, which has served science for so many decades (one can even argue for centuries) is coming apart at the seams.

On the one hand, there is way too much knowledge. Of course, there are too many papers to read; a problem that has been addressed in many publications, and different solutions have been proposed for this problem. But there is also an avalanche of data created within a lab, a research group: all experimental data points exist as electronic files on some hard drive, all calculations, manipulations, renderings, interpretations; all conversations, cogitations and reviews; all presentations, publications, reviews and curations exist somewhere, stored in some format, by someone. Labs differ in the degree of rigor they impose on the structure, on the metadata of these data points. Some require the maintenance of an electronic lab notebook, where at least all steps performed in the preparation of the experiment, the settings for the measuring devices, and what the rats had for breakfast are recorded, labeled, stored, and backed up. Other labs are less neat: most of the knowledge is saved in emails or text files on individual hard-drives or network drives; unlabeled and inaccessible to anyone except their creator, or maybe one or two others. No lab has a perfect system for storing and accessing everything. Ideas, motivations for experiments are only stated in emails. Workflow steps exist as paper artifacts, or text files on idiosyncratic servers; for different reasons, not everyone follows the same workflow, but small deviations are ignored and not re-entered into the system. And no one has a full overview of what happens in the lab. Each researcher has enough trouble producing, locating and processing their own data and there is no time left over to make the information accessible to random strangers.

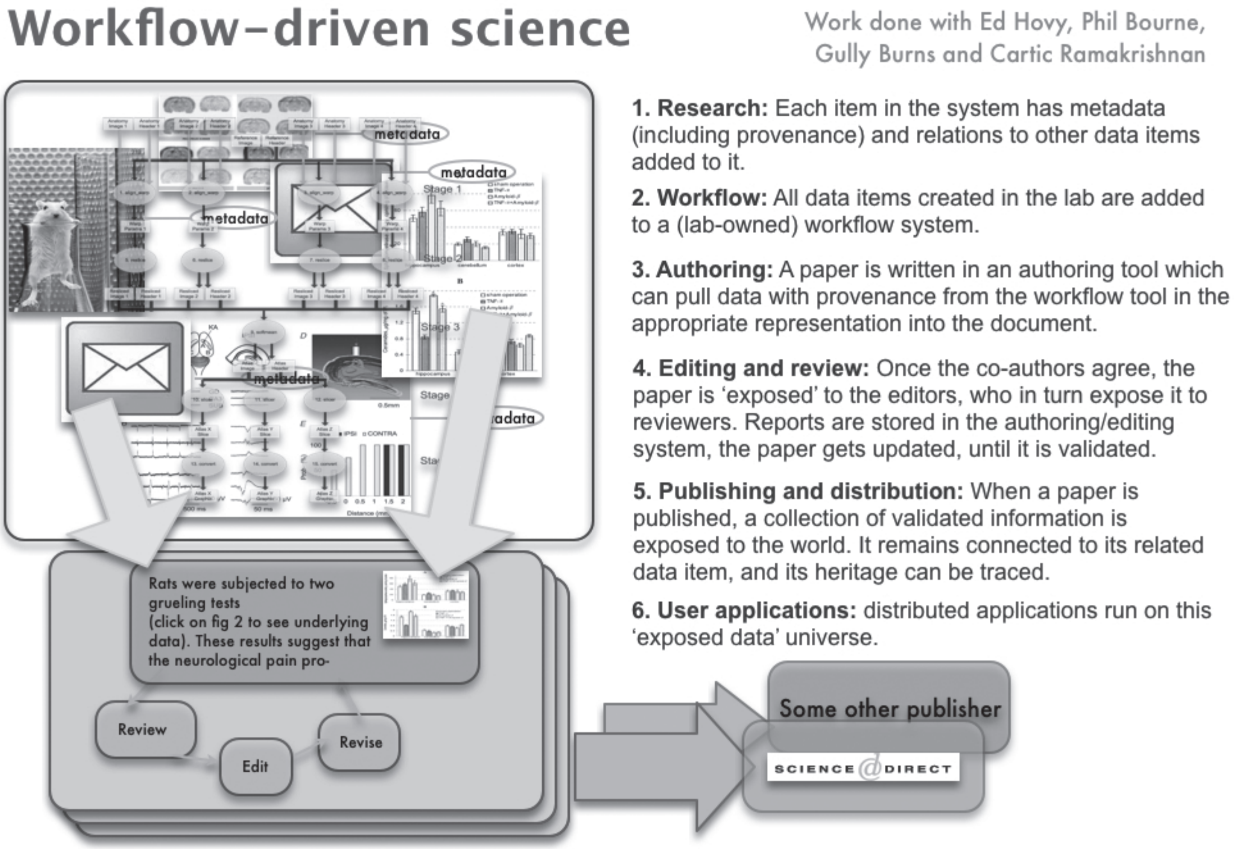
On the other hand, there is not enough knowledge. Papers, in their current format, are disjunct from experimental artifacts; they contain images that have been loosely derived from the research data, but there is no way for a reader to click on an image and see the spreadsheets, the calculations, the image bank or processing steps that went into producing that image. Experimental procedures are loose narratives that bear only an indirect relation to the actual processes that went into the research, making reconstruction of the experiment well-nigh impossible. Each paper is written with the author reproducing the experimental events from memory, tailored to support the argumentation, the scientific story. PowerPoint slide sets contain decontextualized images, taken from different papers, or random representations of current data. Each slide deck, and each paper, is reconstructed anew. Reuse is seen to be cheating. This means that there is no direct link between the information presented to a reader and the information created during the experiment, and no way to reconstruct what was done from the paper, save through a text, whose main goal is a persuasive one.

To alleviate these two problems and advance the pace of scientific discovery we propose a conceptual format that forms the basis of a truly new way of publishing science. In our proposal, all scientific communication objects (including experimental workflows, direct results, email conversations, and all drafted and published information artifacts) are labeled and stored in a great, big, distributed data store. Each item has a set of metadata attached to it, which includes (at least) the person by whom and the time at which it was created, the type of object it is, and the status of the object, including intellectual property rights and ownership. Every researcher can (and must) deposit every knowledge item that is produced in the lab into this repository. With this deposition goes an essential metadata component that states who has the rights to see, use, distribute, buy or sell this item. Into this grand (and system-wise distributed, cloud-based) architecture, all items produced by a single lab, or several labs in collaboration, are stored, labeled and connected. Each datum is connected to others, through a series of relationships: as part of the check-in process, a required step is to identify the data items that the newly entered one is related to. Is it a slide of a brain region? Then the tag number of the rat that used to own the brain, and the method of obtaining the slides, and the settings on the microscope are stored with the slide. Is it a proposal to apply for funding, or a comment on a recently run experiment? The grant number, the status of the comment need to be stored with the email trail.

And from this system, papers are created. Probably, at least in the foreseeable future, these will still consist mostly of persuasive, narrative, text, that is, a personal view on the research performed. But any figure added to the paper will be taken from the data store, and therefore it will contain the metadata that states how it was created, and the relations to the data from which it was derived. A table or figure contains a link to the numbers that went into their making. A reference is automatically imported as a claim gets cited; the claim was marked up as such from the cited paper, and bibliographic information got dragged along from the source. PowerPoint slides can be easily regrouped and reconstructed, since all previous slides have been stored and described in the system. Links are bidirectional: when a figure gets published in a paper, not only does a figure link back to the underlying data, but this fact is now added to the figure in the database, as well. Should there be an error found in the tools or samples that generated that figure, it is immediately clear what journal an erratum needs to be sent to. With proper provenance measures, published version of figures or text can always be found in the database, even when developing insights have prompted a change in the data representation or interpretation.

On the receiving end, each researcher creates his or her own view on this data store. Part of it is used every day, to access and retrieve experimental elements, to design and perform new experiments, to access interpretations and calculations on data. The same system, however, also contains the narratives created by other researchers, with meaningful relations to the data that went into them. Therefore, the boundary between a workflow tool, a data store, and a publishing platform blurs. A researcher can limit the viewing rights on her data to her immediate colleagues; on a draft paper to her co-authors; and on a published paper to everyone who has access to a given, coherent collection of papers, that used to be called 'a journal'. The system has tools for reviewing, rewriting, linking and commenting on papers. Papers can be sorted by keyword, because they refer to other papers, or because they are written by an author or a department of interest; the system simply alerts the scientist that new publications in any of these categories of interest have appeared. A social network of 'authors of interest' can be created, and since the authors are also members of the network, new collaborations can easily be started. When one scientist approaches another to suggest collaboration, the approached party can immediately look at not just the papers, but also the data, the workflows, and the research plans of the scientist who wants to collaborate (provided, of course, he or she allows access to this data). Portfolios can be made to apply for jobs by selecting papers and a trail of 'important ideas', or a download of a lab notebook, to see if the applicant is creative, hard-working, and a good communicator.

The tools that allow scientists to perform all these tasks are built on a shared, distributed platform, using a variety of software packages and custom-built applications. There is no single solution for a particular task, but (in iPhone-like fashion) there is an open marketplace of applications, which can be either bought or downloaded for free. Users download and install the apps that best suit their field, their data, their personality and work habits. Since the same system feeds work processes



Workflow–driven science

Work done with Ed Hovy, Phil Bourne, Gully Burns and Cartic Ramakrishnan

1. Research: Each item in the system has metadata (including provenance) and relations to other data items added to it.

2. Workflow: All data items created in the lab are added to a (lab-owned) workflow system.

3. Authoring: A paper is written in an authoring tool which can pull data with provenance from the workflow tool in the appropriate representation into the document.

4. Editing and review: Once the co-authors agree, the paper is 'exposed' to the editors, who in turn expose it to reviewers. Reports are stored in the authoring/editing system, the paper gets updated, until it is validated.

5. Publishing and distribution: When a paper is published, a collection of validated information is exposed to the world. It remains connected to its related data item, and its heritage can be traced.

6. User applications: distributed applications run on this 'exposed data' universe.

Rats were subjected to two grueling tests (click on fig 2 to see underlying data). These results suggest that the neurological pain pro-

Workflow-driven science

and literature research, the integration of new and existing knowledge is a seamless affair. You do not just mention a paper in an email to a colleague: you add it right there, and link the relevant passage or figure to the experimental feature you want to point out. Other apps are developed to mine, discover and visualize the relations that researchers create, within and between groups – provided the researchers checked the box that said their click-throughs could be recorded and the relationships they have made between various items mined. The system underlying all of these activities can be based on different operating systems, using different databases and query languages, and based on different interface, authoring and visualization tool preferences.

There are three things needed to make this vision a reality: first, the development of an exact, rich, future-proof set of metadata tags, which are versatile enough to handle all the tasks described above, but not so enormous that the system or the user are bogged down by them. Creative Commons have developed a 'no-rights-reserved' option, CC0 (see-see-zero; see http://wiki.creativecommons.org/CC0_FAQ) expressly for authors wanting to distance themselves from their data, and adding it to a common pool. Secondly, tools need to be developed that allow the efficient storage, markup, linking and retrieval of the multifarious data items that are to be added. Neither of these is very far from being available. Cloud-based systems offer more computing power than anyone knows what to do with, at negligible cost; for all of the tasks described, there are one or more tools available that can be developed and rolled out in a distributed fashion. Metadata standards for almost all of the items are being developed, or already in place, and proposals exist for proper provenance descriptors and intellectual property rights licenses.

Thirdly, and probably paramount, a social change is needed. To achieve this mythical future, scientists need to store their research workflow in a system, to structure their work habits around such tools, to take the time to ensure the metadata added is appropriate, relevant, and true. Reviewers and editors must be motivated to check, store, add this data, and allow interoperable publication formats. Quite probably, this revolution will not happen in one fell swoop. Rather, as with semantic web and linked-data initiatives, small patches of community will pop up using tools such as VisTrails,[2] MyExperiment[3] or Wings/Pegasus,[4] and decide they want to add their meticulously crafted metadata to the paper they are submitting, or insist on seeing 'under the hood' of their colleagues' data representations. A single image databank (e.g., The Journal of Cell Biology with its associated JCB DataViewer)[5] or a conference (e.g. the Sigmod Conference Repeatability requirements)[6] can nucleate a tradition of shared data; a data center (e.g. Pangea in Earth Sciences)[7] or common data format (e.g., the Cambridge Structural Database for chemistry)[8] can spawn habits of precision and submission, that collect into connectable troves of data.

Scientists will want to ensure that their data remains theirs; that any tools and pieces of software are not owned by a commercial vendor, publisher or data center, and that they can at any time decide to withdraw, edit or replace their data. Not all experiments are suitable for storing in a workflow tool: some data is one-off, erratic, and stored in an idiosyncratic format. Different subfields of science have very different needs and standards, both ethical and technical, of how to share, display, review and share research data. But it seems that now that papers are, for the most part, ubiquitously available, it's time to make them work, to make them able to contain a greater part of the scientific process than the post-hoc narrative. It is fascinating to see these many different efforts develop, and complementary components click into place. Without any hint of a doubt – these are exciting times for science publishing. □

Notes

[1] This paper reflects a discussion carried on in May of 2010 at ISI between Phil Bourne, Eduard Hovy, Gully Burns, Cartic Ramakrishnan and myself. The exact formulations in this piece are mine, but the main ideas are a direct outcome of this inspired conversation, and a distinctly collaborative effort. A lot of these visions and ideas were presaged and are being brought into practice by Tim Clark, Carol Goble, Larry Hunter, and many others in the bioinformatics workflow systems and semantic web community. I am very grateful to Tim Clark, Rosalind Reid, David Shotton, Amanda Clare and Phil Bourne for their constructive comments on this paper.

[2] http://www.vistrails.org/index.php/Main_Page#VisTrails_Overview.

[3] http://www.myexperiment.org/.

[4] http://pegasus.isi.edu/applications.php.

[5] http://jcb-dataviewer.rupress.org/jcb/.

[6] http://www.sigmod2010.org/calls_papers_sigmod_research_repeatability.shtml.

[7] http://www.pangaea.de/about/.

[8] http://www.ccdc.cam.ac.uk/products/csd/gen_information_content/.