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