Open Notebook Science and Not-so-open Notebook Science

Peter Murray-Rust introduced me to the term Open Notebook Science over lunch in June this year.

In this post I will look a little at open science, as practiced by chemists, and at a recent DLib paper by colleagues of mine from Monash about the continuum of data form when it is collected to when it is published, or curated, in the context of the forthcomi ng Australian National Data Service (ANDS).

There are two points to this post:

1. A cheeky suggestion for a new term "Shared Notebook Science." I assert that this is new because Google says:

   > Your search - **"Shared Notebook Science"** - did not match any documents.
   >
   > http://www.google.com.au/search?hl=en&q=%E2%80%9CShared+Notebook+Science%E2%80%9D

   Whereas "Open Notebook Science" returns over 12,000 hits.

2. And a proposal for a new bit of infrastructure to help with both Open and Shared varieties of data-driven research regardless of whether it's broadly or narrowly accessible.

Background

The term Open Notebook Science comes from Jean-Claude Bradley who has a presentation available in PDF and PowerPoint and audio and video. he talks about lots of tools that people can use for doing data-driven research:

Two slides stood out for me:

1. *The blog as an integrative tool*
2. *The wiki as the laboratory notebook*

There are lots of slides showing various data-driven research tools – but you need some way to make that research accessible to others and to provide the commentary that puts it in context. The presentation is worth watching. My comment is that the science would not have to be open for the tools to be useful.

The *Data Curation Continuum* doesn't discuss this aspect of the practice of Science but it does look at data-flows. The authors note that there's a requirement that access is **controlled, not open** for a lot of researchers:

> The work of investigators in the DART project has indicated that many researchers are extremely conservative when it comes to granting access to research data. This appears to be associated with increasing competiti on in attracting research funds and having articles accepted by high-value publicati ons. The recent move in Australia

> towards assessment of institutional research performance based on quality metrics (the Research Quality Framework – DEST 2007) is only intensifying this. As a result, many researchers want tightly controlled access prior to publication. It is theoretically possible to provide the levels of access control demanded by researchers in a repository that also hosts open-access content, but separation of the two types of repositories may be a preferred solution. Post publicati on, there is some evidence that open access leads to increased accessibility and increased citation rates. This may encourage researchers to be less restrictive about access.
>
> http://www.dlib.org/dlib/september07/treloar/09treloar.html

1 Shared Notebook science

I think that to cater to the access requirements noted by Andrew Treloar et al you might need a "Shared Notebook" approach as opposed to an Open Notebook approach. That is, the benefits of a unifying tools like Blogs and Wikis shared with colleagues but not open to all. (I do realize that there's a bit of baggage that comes along for the ride if you substitute Shared for Open, but I think the term fits.)

In fact I've worked on a "Shared Notebook" project myself. RUBRIC. RUBRIC used a shared space in which to work, which included a wiki, although our blogs were open-access. Read Kate Watson and Chelsea Harper's paper for more about how we got on.

So, some research needs to be kept under wraps, and some people are happier collaborating withing a trusted community.

But whether we're talking Open or Shared, there's a need to promote access and to communicate with peers. We need ways to do that on both sides of what Andrew Treloar and co. call "the curation boundary". Which brings us to my second point. Making it easy to talk about data, right across the curation continuum.

2 New tools needed

Back to the Australian National Data Service, ANDS.

I think one contribution that USQ could make, in collaboration with someone (and I'm thinking of chemists because I have made contact with some) is to work out a generic service interface for making data easy to embed into document, from the ephemeral through to published works. Here's a workflow. I'll base it around my earlier example of chemical markup language but it could apply to any kind of data, big or small.

1. I make a new molecule.

   Lets assume that it's a big one, so the Chemical Markup Language that describes it is a **big** lump of XML.

   (You have to suspend disbelief here. My first-year chemistry lecturer, Julia James suggested that chemistry might not be my vocation and pointed out that if I made up my mind quickly I could withdraw without penalty. I did.)

2. I commit all the data about my molecule to the ANDS data service, where I know it will be safe, and my colleagues can see it. This includes the CML but also observati ons I have made (I'm getting vague about the detail here, I know).
3. Now I want to tell people about my work. Some of them might find the data because they are subscribed to a feed of some kind. Atom seems like an appropriate protocol given the subject matter, but RSS if you really must. But what about others who aren't using those tools or who want to know a bit more about the context of my work?
4. In addition to automated dissemination I want to record stuff in my notebook, which could be a wiki, or I may choose to blog it, maybe using ICE.

   I think what we need here is a way for me to chuck a link into my document (for small amounts of data I would simply paste the data in) and have helper software do the rest, making a print view of the data and embedding live visualizat ion tools into the web view of the data.
5. Later, wehn I start a paper, and I can copy/paste bits from the blog and the Wiki into a document and have the ICE and the ANDS service look after the details.

To support this I reckon we need a standard way to ask for **any data** in the **range of views** we want, with software to make it easy to write about it or call it up in a presentation. This is as necessary as bibliography management via a tool like Zotero, isn't it?

The software service would need to do something like this:

- Look at my data and work out that it's CML (or any one of hundreds of other formats).
- Generate a 2D image for me to use in the print-view of my notes (may not be important for wikis and blogs but ICE makes PDF for everything and as you move towards publishing you want this view).
- Generate the HTML code needed to embed a useful visualizat ion in my web page.
- Insert all the right stuff in the right places without me having to do anything, so I get a suitable live link from my notebook/blog to the data with a useful visualization

ICE can already do some of this, and version 2, due in a couple of weeks will be a good platform for this kind of service.