Blog post: https://www.esipfed.org/esip-interviews/making-data-matter-with-bob-sandusky

Blog title: Making Data Matter with Bob Sandusky

Interviewee: Bob Sandusky, University of Illinois
Date: July 17, 2018
Interviewer: Arika Virapongse

Blog highlight: “Those of us working in the Library and Information Science space today agree that data should be free and open for people to use, and not commercialized.”

Arika: Could you tell me about when and how you got started working in data and informatics, particularly as it pertains to Earth Science?

Bob: I got involved through the DataONE project. I was teaching at the University of Tennessee at the School of Information Science. That was when the NSF (National Science Foundation) solicitation for the DataNet projects came out (in 2007). The solicitation asked that librarians or archivists be part of the project because the projects were about long term preservation of data--with the horizon of several decades. NSF wanted the projects to have multidisciplinary teams.

At the School of Information Science, we had a visit from someone from Oak Ridge National Labs to talk about the solicitation. As a result, several colleagues from Oak Ridge and I participated in writing the proposal, as well as folks from University of New Mexico and University of California-Santa Barbara--it was a large team. Bill Michener from University of New Mexico was the PI. He had been working with Matt Jones (who was also part of the core group) in ecoinformatics before the proposal, so they already had a strong conception of the general shape of the proposal. That was very helpful.

The proposal was submitted in 2008. To make a long story short, we won one of the first two awards; the award came through in summer 2009. This funding was the start of DataONE.

I have a background in Library Science and Computer Science, so I was able to participate in two sides of the group, working with the cyberinfrastructure team and the social science portion of the project. For the first 3-4 years of the project, I was involved in designing the software (cyberinfrastructure). For the social science portion, I did some research into libraries and librarians in regards to how to better involve them into research data management from the library perspective.

In comparison to how things are today, what was the political, science, and technology climate like at the time that allowed the project to come to fruition?

At the time that NSF put out that solicitation, people weren't thinking too much or too widely about data re-use. The National Institutes of Health (NIH) were a little farther along than NSF in terms of thinking about re-use, because the NIH already had some requirements for proposals about how research would be shared. But the momentum built pretty quickly. In 2010, NSF came out with their revised guidelines stating that proposals had to include a data management plan. In subsequent years, federal agencies also had to support data preservation and re-use in projects that they were funding.

Part of what DataONE was trying to do was to encourage the community to develop through outreach and direct training of scientists at professional meetings, develop educational materials, and look at how librarians could teach data management practices on their campuses. Librarians always saw themselves as playing a role in all of this because libraries traditionally steward and curate scholarship, which are traditionally in the form of journal articles, books, and special collections. People started viewing research as a new kind of special collection that had value. There was a huge acceleration in the amount of data that you could capture and store, and it still continues to expand very rapidly.

Data preservation and re-use are still so important. How do you think we are doing in reaching these goals, and what is the next step?

Funding resources are still such a challenge. It's also a shift in culture in research disciplines of all kinds. That sort of change occurs slowly, particularly in academic settings where people are trying to achieve tenure. There is not much recognition yet as to the value of collecting data and describing data, as opposed to publishing a research article. Academics are trained in the mentor/apprenticeship model, so it takes time for that type of culture to change. So things like mandates from NSF and NIH around data management plans help. Librarians are trying to educate undergraduate and graduate students to get these ideas planted in the next generation, as well as working with people who are mature and senior researchers. They are trying to work on all fronts to change culture by making data management and preservation a standard practice. There is still a long way to go.

Staying on the theme of data preservation and re-use, what are some other major milestones that have occurred to push these along?

There are commercial entities that are building vertical, integrated product structures. They are starting to move into the data management and preservation space. Over the past 10-15 years, they haven't been doing a good job with helping researchers manage data, but now they seem to be sensing more economic opportunity to build systems that provide support for that. But one of the big concerns is that commercial publishers will put access to data behind a paywall. Those of us working in the space today universally agree that data should be free and open for people to use, and not commercialized. It is going to be important to see in the next 2 to 5 years what happens with respect to these open, public projects, like DataONE, and how that relates to these commercial entities that are seeking to monetize data.

What do you mean by vertical, integrated product structures?

There has been a lot of consolidation of publishers over the last 15 years. There are fewer large scientific publishers. In addition to the typical article products, publishers also are creating products that help universities collect and understand research productivity within their university, and now they are moving into data. They want products that they can sell to a big university. For example, a university may pay $1 million to Elsevier for access to articles, so they are looking for opportunities to increase what they can charge. The concern is that this divides the world into the haves and have-nots in terms of access to data. We already see this happening with journal articles, so there is potential for this same thing to happen with data. There are some institutions that just can't afford to spend that money on 3 to 4 publishers a year. This is a problem in the US, so its a bigger problem for most of the world.

So what is the future for all of this?

It is hard to predict. In my field of Library and Information Science, everyone is in favor of open access. We try to encourage researchers to maintain some of their rights when they publish a paper. It's actually pretty easy to do. They can agree to publish in a journal but reserve the right to publish in their institutional repository so that its open access. Sometimes you can negotiate this, and there is often an embargo period of 6 months to a year. There are also more open access platforms. This past year there have been some universities that are refusing to pay the high costs set by these publishers. That's difficult to do, but they are pushing back to try to fix the economics. Publisher prices are rising 5-8% a year, and university budgets are declining in terms of public support. It's just breaking down.

[Disclaimer: Any opinions or recommendations expressed in this interview are those of the interviewee and do not necessarily reflect the views of the University of Illinois or any other organizations listed. This interview also represents an "oral history" (a recollection of history), so its value is in the personal perspectives and insights of the interviewee, rather than specific dates, years, and titles for reference.]