

First Jewett data workshop - Observations and Impressions

The first data provenance workshop was a timely and necessary start to the understanding the problems and processes researchers will have in the near future especially for submitting manuscripts for publication which will require that data used in that article be included with the manuscript.

Although I am only involved on the periphery of data management, I have some observations from the workshop. I will try to itemize instead of discuss each observation in detail. Please take all these observations as positive and helpful, as they are intended.

Since this is not my area of expertise, I did not understand much of the terminology and acronyms which I have since taken the time to understand. But I was at a loss sometimes at the workshop.

The ctd example was a very good example since it is one of the common measurements performed in oceanography. It was obvious that the workshop had a hard time setting up a description for this basic measurement. I feel working on an extensible framework for describing the data would have been more useful than actually trying to describe the data that was not very well understood by many at the meeting.

I felt the workshop shifted back among publication data, data publications, and field data, which was confusing for me. I feel the items specifically listed on the agenda should have been pursued a little more. It seemed the workshop was very ambitious (which does not mean that is a bad thing).

I still feel that the original raw data need not be referenced in the publication. For example: the ctd data that Peter was trying to define was at, say, a level 3. The initial data file (level 1) contains an ascii metadata header and binary data that is seldom useful to anyone and need proprietary software to convert to a usable form. After conversion (level 2), the data is saved in ascii text form which includes the metadata header. The ctd data example from the workshop was a modified version (level 3) with the metadata header removed for easier plotting. Which level should 1) be included in the manuscript, 2) be referred to in the document, 3) be made available?

I would have liked the workshop to define recommendations for publishers and editors as well as those for those who are submitting it..

It might have been useful to include a publisher and/or editor of an existing scientific journal, perhaps one that actually has not been set up for this format. There are many available in the area (JASA, IEEE JOE, etc). This would allow us another perspective for what they may be worried about converting to this format and to be able to address some of their concerns.

I would suggest more time in smaller break out groups or follow up groups. Large groups seldom come to a consensus and a smaller group may have a better chance of understanding and defining the data.

What was the exact intention(s) to accomplish? This was not well-defined for me.

Journals are peer reviewed and should be able to stand on their own. Which comes first published manuscript or published data? I feel metadata should not include the methods to get to the scientific end, that is for the manuscript to describe. This might have been discussed and defined a little more.

I learned a lot and hope that maybe I can more useful in the future.