

Greetings!

Please find my response to the above referenced RFI:

Patrick Durusau
patrick@durusau.net

Patrick Durusau (consultant)

Covington, Georgia 30014

Comments on questions (10) – (13), under “Standards for Interoperability, Re-Use and Re-Purposing.”

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

The goals of interoperability, reuse, and repurposing of digital scientific data are not usually addressed by a single standard on digital data.

For example, in astronomy, the FITS (<http://en.wikipedia.org/wiki/FITS>) format is routinely used to ensure digital data interoperability. In some absolute sense, if the data is in a proper FITS format, it can be “read” by FITS conforming software.

But being in FITS format is no guarantee of reuse or repurposing. Many projects adopt “local” extensions to FITS and their FITS files can be reused or repurposed, if and only if the local extensions are understood. (Local FITS Conventions (http://fits.gsfc.nasa.gov/fits_local_conventions.html), FITS Keyword Dictionaries (http://fits.gsfc.nasa.gov/fits_dictionary.html))

That is not to fault projects for having “local” conventions but to illustrate that scientific research can require customization of digital data standards and reuse and repurposing will depend upon documentation of those extensions.

Reuse and repurposing would be enhanced by the use of a mapping standard, such as ISO/IEC 13250, Topic Maps (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38068). Briefly stated, topic maps enable the creation of mapping/navigational structures over digital (and analog) scientific data, furthering the goals of reuse and repurposing.

To return to the “local” conventions for FITS, it isn't hard to imagine future solar research missions that develop different “local” conventions from the SDAC FITS Keyword Conventions (http://www.lmsal.com/solarsoft/ssw_standards.html). Interoperable to be sure because of the conformant FITS format, but reuse and repurposing become problematic with files from both

data sets.

Topic maps enable experts to map the “local” conventions of the projects, one to the other, without any prior limitation on the basis for that mapping. It is important that experts be able to use their “present day” reasons to map data sets together, not just reasons from the dusty past.

Some data may go unmapped. Or should we say that not all data will be found equally useful? Mapping can and will make it easier to reuse and repurpose data but that is not without cost. The participants in a field should be allowed to make the decision if mappings to legacy data are needed.

Some Babylonian astronomical texts(http://en.wikipedia.org/wiki/Babylonian_astronomy) have survived but they haven't been translated into modern astronomical digital format. The point being that no rule for mapping between data sets will fit all occasions.

When mapping is appropriate, topic maps offer the capacity to reuse data across shifting practices of nomenclature and styles. Twenty years ago asking about “Dublin Core” would have evoked a puzzled look. Asking about a current feature in “Dublin Core” twenty years from now, is likely to have the same response.

Planning on change and mapping it when useful, is a better response than pretending change stops with the current generation.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The work of the IAU (International Astronomical Union (<http://www.iau.org/>)) and its maintenance of the FITS standard mentioned above is an example of a successful data standard effort.

Not formally part of the standards process but the most important factor was the people involved. They were dedicated to the development of data and placing that data in the hands of others engaged in the same enterprise.

To put a less glowing and perhaps repeatable explanation on their sharing, one could say members of the astronomical community had a mutual interest in sharing data.

Where gathering of data is dependent upon the vagaries of the weather, equipment, observing schedules and the like, data has to be taken from any available source. That being the case, there is an incentive to share data with others in like circumstances.

Funding decisions for research should depend not only on the use of standards that enable sharing but awarding heavy consideration on active sharing.

(12) How could Federal agencies promote effective coordination on digital data standards with

other nations and international communities?

The answer here depends on what is meant by “effective coordination?” It wasn't all that long ago that the debates were raging about whether both ODF (ISO/IEC 26300) and OOXML (ISO/IEC 29500) should both be ISO standards. Despite being (or perhaps because of) the ODF editor, I thought it would be to the advantage of both proposals to be ISO standards.

Several years later, I stand by that position. Progress has been slower than I would like at seeing the standards draw closer together but there are applications that support both so that is a start.

Different digital standards have and will develop for the same areas of research. Some for reasons that aren't hard to see, some for historical accidents, others for reasons we may never know. Semantic diversity expressed in the existence of different standards is going to be with us always.

Attempting to force different communities (the source of different standards) together will have unhappy results all the way around. Instead, federal agencies should take the initiative to be the cross-walk as it were between diverse groups working in the same areas. As semantic brokers, who are familiar with two or three or perhaps more perspectives, federal agencies will offer a level of expertise that will be hard to match.

It will be a slow, evolutionary process but contributions based on understanding different perspectives will bring diverse efforts closer together. It won't be quick or easy but federal agencies are uniquely positioned to bring the long term commitment to develop such expertise.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Linking between publications and associated data presumes availability of the associated data. To recall the comments on incentives for sharing, making data available should be a requirement for present funding and a factor to be considered for future funding.

Applications for funding should also be judged on the extent to which they plan on incorporating existing data sets and/or provide reasons why that data should not be reused. Agencies can play an important “awareness” role by developing and maintaining resources that catalog data in given fields.

It isn't clear that any particular type of linking between publication and associated data should be mandated. The “type” of linking is going to vary based on available technologies.

What is clear is that the publication its dependency on associated data should be clearly identified. Moreover, the data should be documented such that in the absence of the published article, a researcher in the field could use or reuse the data.

Please don't hesitate to contact me if expansions or further explanations would be helpful.

Hope you are having a great day!

Patrick

--

Patrick Durusau

Chair, V1 - US TAG to JTC 1/SC 34

Convener, JTC 1/SC 34/WG 3 (Topic Maps)

Editor, OpenDocument Format TC (OASIS), Project Editor ISO/IEC 26300

Co-Editor, ISO/IEC 13250-1, 13250-5 (Topic Maps)

OASIS Technical Advisory Board (TAB) - member