**International Resource Registry for National Metrology Institutes**
**Next Steps**

*Robert J. Hanisch*
*Director, Office of Data and Informatics*
*Material Measurement Laboratory*
*National Institute of Standards and Technology*
*Gaithersburg, MD USA*

28 December 2015

Following our kick-off telecon of 7 December, I agreed to write down a draft plan for how to proceed in the coming months and prior to our planned face-to-face meeting at BIPM, Paris, on 15 April 2016.

We have two major issues to discuss: metadata and software. Since our telecon the issue has also been raised regarding policy, that is, how do we decide just what kind of data or ancillary information gets entered into the Registry? The policy issue is actually closely related to the metadata schema design.

***Metadata***

It is common in many metadata schemas to start with a well-known and widely used schema as a core, and expand upon that to deal with special use cases. I suggest that we use the Dublin Core metadata schema[1] as our starting point. Dublin Core originated in the library community and thus was designed primarily to describe books and publications, but its basic metadata elements are relevant to data. These are as follows:

*Contributor*
*Coverage*
*Creator*
*Date*
*Description*
*Format*
*Identifier*
*Language*
*Publisher*
*Relation*
*Rights*
*Source*
*Subject*
*Title*
*Type*

Let me give you an example of what these fields might look like for one of the NIST Standard Reference Data databases, SRD #20.

[1] http://dublincore.org/documents/dces/

| | |
|---|---|
| *Title:* | NIST X-ray Photoelectron Spectroscopy Database XPS, Version 4.1 |
| *Type:* | database |
| *Subject:* | photoelectron and Auger-electron spectral lines |
| *Description:* | The NIST X-ray Photoelectron Spectroscopy (XPS) Database gives easy access to the energies of many photoelectron and Auger-electron spectral lines. Resulting from a critical evaluation of the published literature, the database contains over 22,000 line positions, chemical shifts, doublet splittings, and energy separations of photoelectron and Auger-electron lines. A highly interactive program allows the user to search by element, line type, line energy, and many other variables. Users can easily identify unknown measured lines by matching to previous measurements. |
| *Creator:* | Alexander V. Naumkin, Anna Kraut-Vass, Stephen W. Gaarenstroom, Cedric J. Powell |
| *Contributor:* | Charles D. Wagner |
| *Date:* | 2012-09-15 (ISO 8601 format) |
| *Publisher:* | National Institute of Standards and Technology |
| *Format:* | website |
| *Identifier:* | ECBCC1C130062ED9E04306570681B10712 (could be DOI) |
| *Rights:* | copyright 2012 by the U.S. Secretary of Commerce on behalf of the United States of America |
| *Language:* | en |
| *Coverage:* | — (not applicable) |
| *Relation:* | NIST SRD 100 Database for the Simulation of Electron Spectra for Surface Analysis |
| *Source:* | A.Y. Lee et al., Data Science Journal 1, 1 (2002) |

This is the sort of general bibliographic information that one would expect to be available for NMI data products. Some additional metadata elements could be useful, such as

*Keywords*
*ContactName*
*ContactEmail*
*Website*
*Version*

The idea is to have enough metadata to support data discovery and allow users to locate information quite accurately, but without putting too large a burden on data providers to populate a metadata schema with many tens or hundreds of elements. We need to discuss other possible metadata terms and consider if we should drop some of the Dublin Core terms that might rarely be used (such as *Coverage*).

***Policy***

As national metrology institutes, the NMIs are expected to be providers of high quality data. On the other hand, data that is not good enough for one purpose may be perfectly adequate in another, and it would be helpful to make that data discoverable as long as it is properly characterized.

NIST recently developed a "data pyramid" to characterize different levels of data. The lowest level, i.e., raw data, not necessarily calibrated, is Working Data. Above that comes Derived Data, then Publishable Data and Published Data. Published Data are associated with peer-reviewed journal articles, so they are vetted at a reasonable level and considered to be useful, though are not 100% guaranteed.

The top levels of the pyramid are Resource Data (data upon which decisions are based), Reference Data, and Standard Reference Data. SRD are the most highly vetted and characterized data produced by NIST and they are protected by copyright, whereas Reference Data are typically of equal quality but do not fit under the specific terms of the Standard Reference Data Act of 1968.

I suggest that we can deal with data quality issues by adding one or two metadata elements to describe the degree of data assessment or confidence. We could use the NIST data pyramid terms, or come up with some arbitrary "data reliability scale" going from 0 (unknown) to 10 (SRD), though I'm not sure how one would discriminate between a "4" and a "5", say. We could agree on some coarser levels, such as Pre-Publication Data, Published Data, and Reference Data. And it could be perfectly acceptable from a policy perspective for an NMI to only publish SRD-level data descriptions if they so choose.

***Software***

The registry is intended to be a distributed system, with each NMI, or NMIs working through their RMOs, to collect and curate metadata about their data resources and then to aggregate the metadata into a searchable database. This aggregation step requires that we agree on a core metadata schema and on a protocol for metadata harvesting.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been in use for 15 years and provides a simple model for updating and synchronizing metadata collections.[2] There are also a number of readily available software implementations. OAI-PMH uses XML for message transport, though other formats such as JSON[3] can be embedded for actually encoding the metadata fields.

JSON-LD[4] is quite widely used today as a mechanism for describing and aggregating distributed resource descriptions. We could choose one of these options, or perhaps support both protocols if there is sufficient diversity of expertise within the NMIs.

***Telecons and Meetings***

We should have two or perhaps three telecons between now and our planned face-to-face meeting at BIPM on April 15.

---

[2] https://www.openarchives.org/pmh/
[3] http://www.json.org/
[4] http://json-ld.org/

Telecon 1: Metadata (and perhaps Policy)

Discuss and agree on an initial, core set of metadata elements. We can adjust and update if we find that we are missing some important components, or if we have elements that no one is using.

Telecon 2: Policy (if not covered in Telecon 1)

Discuss what metadata elements are needed to describe data quality. If we agree that only SRD-level data should be in the registry, we need not define the metadata elements.

Telecon 3: Software

Discuss various approach to metadata harvesting and tools available for metadata population and curation.

Face-to-face meeting: 15 April 2016, BIPM, Paris

Since a number of participants will be traveling some distance it might make sense to start our technical discussions a day earlier. We could have some informal work sessions and report out to the group as a whole on the 15th. At the end of the face-to-face meeting we should have reached agreement on our initial metadata schema, what software we will use for harvesting and curation, and how we will go about a pilot implementation to demonstrate in October.