Response to Request for Information (NOT-ES-15-002): Input on Making Data Usable - A Framework for Community-Based Data and Metadata Standards Efforts for NIH-relevant Research

Big Data presents an exciting opportunity to pursue large-scale analyses over collections of data in order to uncover valuable insights across a myriad of fields and disciplines. Yet, as more and more data is made available, researchers are finding it increasingly difficult to discover and reuse these data. One problem is that **data are insufficiently described** to understand what they are or how they were produced. A second issue is that **no single vocabulary provides all key metadata fields** required to support basic scientific use cases. A third issue is that **data catalogs and data repositories all use different metadata standards**, if they use any standard at all, and this thwarts efforts to easily search, aggregate, and exchange data and their descriptions. Therefore, we need a guide to indicate what are the essential metadata, and the manner in which we can express it.

Working under the auspices of the standards-setting World Wide Web Consortium (W3C), the **Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG)** has engaged a multi-stakeholder community to produce a **guideline for the description of datasets** (http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/) in order to meet key functional requirements. These functional requirements include identification, summarization, versioning, licensing, linking, querying, and provenance of datasets. This work is significant because it offers a clear guidance in the reuse of 14 standard and community-developed vocabularies (e.g. RDFS, Dublin Core, DCAT, PROV, VOID, CiTo, FOAF, PAV, etc) to describe datasets to meet the needs of data registries, data producers, and data consumers.

We offer our experience in the following solicited areas for information:

1. **Effective approaches, processes, and activities that could advance the community-based standards landscape**. We initiated our community activity by soliciting participation on our mailing list, which includes clinical, pharmaceutical, and biomedical researchers, bioinformaticians, developers, and semantic web enthusiasts. We held weekly teleconference calls to solicit expressions of interest, define use cases, identify requirements, prioritize metadata fields, survey existing vocabularies, develop consensus on using particular vocabularies and value sets, write a draft of the report, and revise the report based on community feedback. We used Google docs to enable collaborative editing of tables and documents. We took minutes of meetings and posted them on the mailing list. Our work is the product of significant consultation, consensus, and community feedback. It has created new opportunities for participants to discuss their experience and demonstrate its utility in addressing outstanding biomedical problems while addressing the needs of the community.
2. **Common challenges in CBS development.** A key aspect of standardization is that there exists people with functioning systems that share a common objective but differ in the details of their implementation. A major challenge for vested participants is to embrace standardization as a way to to achieve interoperability in a well described manner. We were fortunate in that HCLS participants were willing to adapt existing approaches in favour of a

shared specification so as to deliver longer term benefits to the community. We were also successful in engaging a broad set of stakeholders to contribute to our effort, but it is an outstanding challenge to connect to all relevant communities in a meaningful and effective manner. Yet, through our collective network, the HCLS dataset description guideline is being discussed as part of promoting standards in emerging data-focused organizations such as Force11, Elixir, Research Data Alliance, the Global Alliance, and CEDAR, a recently funded NIH BD2K center of excellence for expanded data annotation and retrieval.

3. **Considerations for evaluating progress and milestones to assess data standards development.** Ideally, standardization efforts would have experienced facilitators that understand the tasks and resources necessary to achieve the objectives. In our case, as an W3C Interest Group, we did not have the kind of standardization experience that W3C Working Groups normally have access to. While we had a rough sense of the desired outcome, we did not *a priori* define specific milestones and also could not reliably assess how much time would be required to achieve specific objectives. Thus, having a “standardization” starter kit with the list of activities, milestones and deliverables would probably be incredibly useful for individuals that want to undertake this effort. Moreover, we believe that the process could have been improved by having dedicated, knowledgeable staff to not only oversee the process, but also to aid in technical writing and revision. Our work largely hinged on having a core of 3-6 individuals devote some of their time (in an unpaid manner) to chairing the calls, writing the minutes, reaching out to inside and outside participants, and writing the documents. While our working group consists of talented individuals, the fact that they voluntarily contributed their time to push things forward also meant that they may be required to reprioritize their time over the course of the standardization effort. Having dedicated staff could ensure continuity among an evolving set of participants while ensuring conformance to expected milestones and timelines. Moreover, dedicated staff could play a pivotal role in outreach as well as maintaining the standard with new needs from connected communities.