Data Sharing and Metadata Curation: Obstacles and Strategies

Future strategies for managing scientific data and metadata for basic and applied research

May 29, 2013
National Science Foundation, Room I-1235
4201 Wilson Blvd.
Arlington, VA 22230
9AM – 4:15 PM

Introduction:

This workshop was organized on behalf of the Big Data Senior Steering Group (BDSSG), an interagency body chartered by the White House Office of Science and Technology Policy (OSTP) and facilitated through the National Coordination Office for Information Technology and Networking R&D (NCO-NITRD). The goal was to have focused discussions on future strategies in data and metadata for basic and applied research; specifically, (a) how to better enable, encourage, and realize sharing of data, both across disciplinary divides and between "micro-silos" within research domains, and (b) how to acquire, manage, and curate metadata in order to ensure usability and comprehensibility of data over time and between disciplines.

Our intent was to bring together representatives from distinct data-intensive research domains who have been contributing to community-based solutions to data challenges. Key individuals from communities such as the Materials Genome Initiative, Space Weather, Global Climate, Environmental Health, DataOne, iRODS, and the Research Data Alliance participated.

The workshop was comprised of three main sessions. The first session featured five presentations from practitioners who focus on a particular domain; the second session included three groups that focus on trans disciplinary data; and the third session was an open discussion focused around several questions that had been distributed to the participants in advance.

Executive Summary

Attendees were asked to consider the following five questions.

1. What metadata, and what kinds of metadata management, are needed to enable re-use of data, both across domains and across silos within domains?
2. How can we incentivize researchers and providers to curate their data, organize it with useful metadata, and make it publicly available?
3. Maximum impact of data occurs when analytics make use of all available relevant data; how can analytics developers be challenged to make this standard practice?

4. What are the data ownership and personal identifiable information issues (obstacles/solutions) that can be addressed in this context?

5. What are the top two data/metadata problems you would like to solve?

There were at least four areas of agreement from our discussions:

- Active data stewardship/curation adds value and is needed at some level. However, cost is a major issue. There is no funding model to support the resources needed, and no way to assess the value of data management compared to, e.g. new research grants.
- Exclusively top-down solutions are not desired; but the correct balance between grass-roots vs. "middle-out" initiatives is unclear.
- There is a need for easy-to-use tools for metadata creation, improvement, and workflows that incorporate good data practices.
- Funding agencies can provide incentives for researchers to share data. For example, applicants could receive credit for making their data more readily accessible through the use of community best practices for sharing; funded researchers could be required to use their research field's metadata standards.

Discussion and Presentation Summary:

The following summary is intended for the workshop participants and the members of the BDSSG and is not meant to be a complete review of the subject. It is comprised of summary notes and links to the presenter's slides and video.

Practitioners' Perspectives:

Moderated by Robert Chadduck, NSF

- **DataOne** (DataNet Observational Network for the Earth) - *Rebecca Koskela, University of New Mexico* slides; video
    - *Purpose:* to promote data discovery in earth/environmental sciences.
    - *Method:* Three major nodes (CA, NM, ORNL) fed by member nodes. Becoming a member node gives more visibility to your data.
    - *Provide:* investigator toolkit; process to align diverse metadata; index metadata for the search API; Tools such as DataUP, OneShare, and Dryad to help researchers improve data practices, create metadata, help with uploading, repositories and DOIs.
    - *Working on:* Semantic mediation, provenance, and automated annotation
    - *Observations:* Of the scientists they work with:
        - 80% want to share data, but only 6% share all of their data
        - have almost no metadata standards
        - most use Excel spreadsheets

- **DFC/iRODS** (DataNet Federation Consortium/innovative rules oriented data system)- *Reagan Moore and Mary Whitton, RENCI* slides; video

- *Purpose:* Provide a federated collaboration environment that supports reproducible data-driven research.
- *Provide:* Mechanisms to enable interoperability and allow domains and services to interact. Not just metadata but the procedures used to create the data product: procedures for data acquisition, data management, automation of the application of domain knowledge; policies for data control. iRODS policy-based data management.
- *Working on:* Encapsulation of domain knowledge for accessing domain repositories, analyzing domain data sets, and managing domain data products. Application of virtualization mechanisms that manage metadata properties and the processes to derive the metadata.

- **NIST/ITL/MML** (NIST Information Technology Laboratory and Material Measurement Lab)- *Mary Brady, Ram Sriram, NIST ITL, and Jim Warren, Carelyn Campbell, NIST MML* slides; video
  - *Purpose:* To facilitate the Material Genome Project by enabling data exchange, ensuring data quality, and establishing new methods and metrologies.
  - *Provide:* Developed repositories and other necessary infrastructure. Currently moderated submission but working toward more automation. Standing up office of data and informatics at NIST, developing universal identifiers and ontologies for materials development.

- **NCN/nanoHUB** (Network for Computational Nanotechnology)- *Gerhard Klimeck, Purdue* slides; video
  - *Purpose:* Resource for the use of the Nanotechnology Community of Researchers
  - *Provide:* Simulation tools, collaboration tools; resources to teach and learn such as nanoHUB-U, courses, seminars, and tools to share and publish tools and research.
  - *Observations:*
    - Perceived myths:
      - You can't use research codes for education, you must write your own,
      - Building user interfaces is too hard, you must rewrite for the web,
      - There are no incentives to share and no end-to-end science cloud possible.
    - Observations regarding these myths: Large development collaborations that serve large number of users = predictable success.
    - Criteria for the success of a science gateway:
      - Outstanding science,
      - Commitment to dissemination,
      - Technology for dissemination,
      - Tech transfer process (i.e. people),
      - Understand the stakeholders,
      - Open assessment, and
      - A business model. Consider the iPad…may not be as capable as a typical desktop, but it's much more useable.

- **BMIR** (Stanford Center for Biomedical Informatics Research) –*Mark Musen, Stanford* slides; video
  - *Purpose:* Research to improve the exchange of health information
  - *Current State:*
    - Research data:

        - BioSharing Initiative – tried to provide a path through all the data and metadata policies, standards, databases.
        - BioDBcore – uniform description of public biological databases
        - Minimal Information About a Microarray Experiment Initiative – grass roots standard that is now adopted by some organizations. Is leading to a markup language and ontology. Many different "minimal information" checklists under the MIBBI portal, all grass-roots efforts
    - Clinical data:
        - HL7 Organization's Reference Information Model has had limited adoption (too complex).
- *Observations:*
    - Development of meaningful use criteria is a necessary first step (also the conclusion of a PCAST report).
    - There must be progress made toward a robust exchange of health information.
    - We need a universal exchange language and an IT infrastructure to support it, but this is not what the vendors involved in HL7 are envisioning.
    - Many metadata solutions have been met with outright hostility.

- **Open Discussion – Practitioners' Perspectives**
    - *"What keeps you up at night?"*
        - Lack of strong incentives,
        - Attribution to individuals is an important incentive,
        - There are concerns about sharing standards, ensuring quality, requiring open source, that need to be addressed,
        - Data sharing is not in the workflow and needs to be,
        - Metadata collection and generation is not scalable to meet the needs,
        - Tools/solutions must be pragmatic and integrated into the workflow,
        - Can we do an "overarching ontology"? Perhaps the best we can do is start with small ontologies as a foothold into crossing domains,
        - Administrative metadata is standard, but descriptive and provenance is not standard across domains,
        - Uses for metadata include provenance and curation, description, and state information
        - The context in which the data was acquired is crucial for getting out of the silo.

Trans-disciplinary Community Perspectives:

Moderated by Alan Hall, NOAA, and Jon Petters, AAAS Fellow at DOE

- **RDA** (Research Data Alliance)- *Fran Berman, Rensselaer Polytechnic* slides; video
    - *Purpose:* to build social, organizational, technical infrastructure to reduce barriers to data sharing and accelerate development of coordinated global infrastructure.
    - *Method*:
        - Working groups work for 12-18 months to build targeted pieces of infrastructure. Interest groups include agricultural data, big data analytics, legal, etc. WG examples are: Persistent Identifier Information Types, Data Type Registries, Data Foundation and Terminology, Practical Policy (latter 2 pending).

        - 45 countries participating, over 700 participants, US =31% of participants. Academics = 61%, private sector 21%. US leadership considered strategically important.

- **EarthCube**- *Joel Cutcher-Gershenfeld, University of Illinois* slides; video
    - *Purpose:* To study institutional context and the impact of "culture" in the context of a community such as EarthCube.
    - *Method:*
        - Reaching out and hearing the voice of the customer. This was a key piece in the development of EarthCube. Results of stakeholder survey = Everybody thinks using disparate data and tools is important but very hard to do.
        - Power is not in pushing on the drivers but in lowering the barriers, ex: if you lower a barrier (like lack of credit) then you create an incentive (credit).
        - Engage cross-cultural conversations; Culture eats strategy for breakfast.
        - Be wary of building a great system that no one uses.

- **Interagency Perspectives**- *Ted Haberman, HDF Group* slides; video
    - *Observations:*
        - We are in an "era of ferment" where there are risks, uncertainty, waste, lack of interoperability. Will reach a point where there is a selection of options, focusing of energy, and a convergence of communities.
        - A lot of the "tools" we talk about are basically portals. Instead, tools for creators of metadata are needed-- rubrics for evaluation, views, connections, development, evolution, and documentation.
        - There is a need for a community support environment (e.g. a Documentation Consortium).

- **Open Discussion- Community**
    - *"What are the successes and worries?"*
        - Success: There is energy around RDA. Worries: Stewardship and economics of stewardship. How to store it and pay for it is a multi-sector issue.
        - Success: Groups are becoming more intentional and focused. Worries: "It takes internal alignment to get lateral alignment." There are risks that you can build it but nobody will come.
        - Success: Lots of resources going into things. Worries: Resources are not being used to contribute to the big picture (ferment) e.g. lots of documents (words being produced), but converting them into actions is a worry.
        - It is hard to do something new and still maintain credibility in your "home domain".
    - *"How do we focus on building cultural infrastructure and removing barriers?*
        - Similar to academic cultures in universities, you have to incentivize the desired behaviors in a mainstream way. Data sharing is dis-incentivized by competition.
        - Behavior changes culture but only very slowly.
        - Find the "positive deviants", people who are already doing things "right" and share their successes.
    - *"We need to keep data that is valuable, and discard that which is irrelevant. Who is responsible for paying for and doing this kind of curation?"*

- Reference to Sustainable Economics for a Digital Planet and other documents on Fran Berman's website. We have to understand that every dataset has multiple stakeholders. Who takes it? Who keeps it? Who manages it?
    - It is time to develop reasonable policies; e.g. what's the value of the data and how hard would it be to regenerate it? The same infrastructure is not necessary for all types of data.
    - Some datasets are interesting only in the context of a paper.
    - In some cases, the cost of regeneration could be huge compared to the cost of curation.
- There are organizations in the EU that don't exist in the US. There needs to be outreach to the science communities to let them know that a larger community is forming.
- Identify the leaders, fund them, and make exemplars of them.

Open Discussion: Barriers and Opportunities:

Moderated by Peter Lyster, NIH, and Mark Suskin

- *"What does "metadata" mean?"*
    - Metadata:
        - Is what you need to ensure that the person you will never meet will not reach incorrect conclusions by using your data.
        - Is contextual and implies active curation.
        - Is used but not talked about: In smart laboratories this information is standardized for at least certain kinds of experiments. But the idea that the metadata you save is supposed to plug into a global infrastructure isn't talked about.
        - Is expensive: nanoHUB is spending 60% of its budget on content stewardship i.e. curation. It requires a PhD level person who can interact with colleagues.
        - Has few but expensive experts: who is it that wants to learn best practices in data curation? E.g. Kirk Borne's efforts at GMU, teaching data practices to their astronomy students.
        - Has few built-in incentives: Curation doesn't lead to scientific publication.
- *What does it mean to succeed in any of these areas? Can we lay out clear-cut desired outcomes? What should the Government do and not do?*
    - Start with what you have already invested in and build on it.
    - High level abstractions that make searches simple is a lesson from business data management. For example "faceted" classification has been adopted and used for the Earth System Grid Portal. It could be developed and used for research funded by multiple agencies (see Habermann's presentation).
    - Develop a grand challenge to develop a cost model. Are we willing to sacrifice research grants for better curation of datasets?
    - Create an environment where there is room for an entrepreneur to do something new.
- *In response to the question: What are the top 2 metadata problems?*
    - The Lack of:
        - Success stories to demonstrate return on investment.
        - Modeling from Government agencies who:

        - Do not collaborate or use international metadata standards effectively.
        - Do not provide clarity for what will be supported by the public sector and how, and what should be supported elsewhere.
    - Agreement on:
        - Scientific ontologies that allow categorization at the right level of abstraction to facilitate the creation of metadata tools.
        - Templates and standards so everybody knows what they are supposed to deliver.
        - Persistent identifiers for everything (to build trust).
        - Semantic standards.
        - Consistent IP rights across data in the US e.g. credit, provenance, citation.
    - Innovative Tools:
        - That IMPROVE metadata e.g. reduce uncertainties that cause errors
        - For data documentation, that points to available standards.
        - That fit data curation into the workflow
    - A Community to:
        - Share best practices and soft knowledge e.g. workflow libraries
        - Help establish the identity of existing collections.
        - Find stable homes for valuable data.
        - Teach data literacy e.g. educating researchers in preserving and sharing data
        - Provide a platform for scholarly communication that isn't publishing a paper but communicating inside a network, "adding data points rather than producing inaccessible works."
    - Incentives that:
        - Treat data as a 1st class publication if it's fully integrated with context.
        - Change the value perception of metadata for the PI.
        - Grade scientists on how well they cite their data.
        - Find no-cost policy changes that enable credit to accrue to data creators.

Summary and Wrap-up:

Tom Statler, NSF **slides**

With Big Data comes big risk: risk of reaching incorrect conclusions (through misunderstanding, misuse, or abuse of data), risk of data investment losing value, risk of data becoming unusable. Metadata is the essential information needed to minimize these risks. Metadata curation is managing these risks and accepting them where appropriate. Standards and practices developed for domain-specific needs are just starting to interact. The hazard of a top-down unification of standards across domains is that it can appear to lower barriers while being doomed to internal fracturing; the sociological problem may be as hard as the technological one.

Johns Hopkin University's Grant Reviewer's Guide was offered as just one example of policy guidance that can be cost effective even in the short term: http://dmp.data.jhu.edu/resources/grant-reviewers-guide/ .

The workshop concluded with a challenge to the participants, "What are YOU going to do?"

- "Talk with my program officer about highlighting dataset developments."
- "DOE Office of Science is giving guidance to PIs and POS about using data management as part of evaluation. For attribution and citation, it's harder but on the list."
- "Helping to organize meetings and sessions. All of this stuff for data also applies to software."
- "NCO will vigorously support the BDSSG and domain group."
- "Get outside of one's own portfolio."
- "My WG will deliver a prototype of a digital object registry."
- "Bring discussions back to NIH, to three workshops."

Participant List

Bell, Randy, DOE/NNSA
Berman, Fran, RPI/RDA
Biven, Laura, DOE/SC
Brady, Mary, NIST
Campbell, Carelyn, NIST/MML
Cutcher-Gershenfeld, Joel, UI
de Groot-Lief, Christina, NOAA
Eigen, Ana, DOT
Gupta, Amarnath, UCSD
Habermann, Ted, The HDF Group
Hagan, Don, CENDI/NTIS
Hamilton, Carol, RTI
Klimeck, Gerhard, Purdue
Kolker, Eugene, SCRI
Koskela, Rebecca, DataONE
Kuznetsova, Maria, NASA
Larkin, Jennie, NIH
Lewis, Suzanna, BBOP
Maddox, Marlo, NASA
Martone, Maryann, UCSD
McDermott, Michael, USGS
Moore, Reagan, RENCI
Musen, Mark, Stanford
Rindflesch, Tom, NIH/NLM
Scott, John Henry, NIST
Shoshani, Arie, LBL
Smith, Barry, NCOR
Szalay, Alex, JHU
Tilmes, Curt, NASA/USGCRP
Vieglais, Dave, UK /DataONE
Viereck, Rodney, NOAA
Ward, Charles, AFRL
Warren, Jim, NIST/MML
Whitton, Mary, RENCI
Wilbanks, John, NCO/NITRD

Organizing Committee:

Bristol, Sky, USGS
Chadduck, Bob, NSF
Dearry, Allen, NIH
Grumbling, Emily, NSF
Hall, Alan, NOAA
Lee, Tsengdar, NASA
Lyster, Peter, NIH
Pantula, Sastry, NSF
Petters, Jonathan, DOE/SC
Preuss, Don, NIH
Statler, Tom, NSF
Suskin, Mark, NSF

Glossary of Acronyms:

AFRL: Air Force Research Laboratory
BBOP: Berkeley Bioinformatics Open-source Projects
CENDI/NTIS: An interagency working group of senior scientific and technical information (STI) managers/ National Technical Information Service
DataONE: Data Observation Network for Earth
DOE/NNSA: Department of Energy/National Nuclear Security Administration
DOE/SC: Department of Energy/Department of Science
DOT: Department of Transportation
HDF: Hierarchical Data Format
JHU: Johns Hopkins University
LBL: Lawrence Berkeley National Laboratory
NASA: National Aeronautic and Space Administration
NASA/USGCRP: National Aeronautic and Space Administration/United States Global Change Research Program
NCOR: National Center for Ontological Research
NCO/NITRD: National Coordination Office/Networking and Information Technology Research and Development
NIH: National Institutes of Health
NIH/NIGMS: National Institutes of Health/National Institute of General Medical Science
NIH/NLM: National Institutes of Health/National Library of Medicine
NIST: National Institute of Standards and Technology
NIST/MML: National Institute of Standards and Technology/Material Measurement Laboratory
NOAA: National Oceanic and Atmospheric Administration
NSF: National Science Foundation
RENCI: Renaissance Computing Institute
RPI/RDA: Rensselaur Polytechnic Institute/Research Data Alliance
RTI: Research Triangle Institute
UCSD: University of California San Diego
UI: University of Illinois
UK: University of Kansas
USGS: United States Geological Survey