Curating for Quality

Ensuring Data Quality to Enable New Science

Final Report: Invitational Workshop Sponsored by the National Science Foundation

September 10-11, 2012
Arlington, VA USA

http://datacuration.web.unc.edu

Gary Marchionini, Christopher A. Lee, and Heather Bowden, University of North Carolina at Chapel Hill
Michael Lesk, Rutgers University
October 19, 2012

Curating for Quality

Ensuring Data Quality to Enable New Science

Final Report: Invitational Workshop Sponsored by the National Science Foundation

September 10-11, 2012
Arlington, VA USA

http://datacuration.web.unc.edu

Gary Marchionini, Christopher A. Lee, and Heather Bowden, University of North Carolina at Chapel Hill
Michael Lesk, Rutgers University
October 19, 2012

Acknowledgements

This workshop was supported by a grant from the National Science Foundation (NSF III #1247471), Gary Marchionini & Cal Lee Principal Investigators

Thanks to Maria Zemankova, National Science Foundation Program Manager

Table of Contents

Executive Summary

Science is built on observations. If our observational data is bad, we are building a house on sand. Some of our data banks have quality measurements and maintenance, such as the National Climate Data Center and the National Center for Biotechnology Information; but others do not, and we do not even know which scientific data services have quality metrics or what they are.

Data quality is an assertion about data properties, typically assumed within a context defined by a collection that holds the data. The assertion is made by the creator of the data. The collection context includes both metadata that describe provenance and representation information, and procedures that are able to parse and manipulate the data. However data quality from the perspective of users is defined based on the data properties that are required for use within their scientific research. The user believes data is of high quality when assertions about compliance can be shown to their research requirements.

Digital data can accumulate rich contextual and derivative data as it is collected, analyzed, used, and reused, and planning for the management of this history requires new kinds of tools, techniques, standards, workflows, and attitudes. **As science and industry recognize the need for digital curation, scientists and information professionals recognize that access and use of data depend on trust in the accuracy and veracity of data**. In all data sets trust and reuse depend on accessible context and metadata that make explicit provenance, precision, and other traces of the datum and data life cycle. Poor data quality can be worse than missing data because it can waste resources and lead to faulty ideas and solutions, or at minimum challenges trust in the results and implications drawn from the data. Improvement in data quality can thus have significant benefits.

The National Science Foundation sponsored a workshop on September 10 and 11, 2012, in Arlington, Virginia on "Curating for Quality: Ensuring Data Quality to Enable New Science." Individuals from government, academic and industry settings gathered to discuss issues, strategies and priorities for ensuring quality in collections of data. This workshop aimed to define data quality research issues and potential solutions. The workshop objectives were organized into four clusters: (1) data quality criteria and contexts, (2) human and institutional factors, (3) tools for effective and painless curation, and (4) metrics for data quality.

Participants were invited to submit short position papers in advance of the event (see Appendix B for copies of submitted papers). The workshop began with personal introductions, followed by brief summaries of the position papers. This was followed by small group discussions of "pain points" and "promising directions" related to the main themes of the workshop. Participants then identified potential project ideas and voted on their top choices. Much of the second day was devoted to discussing the eight project ideas that received the most votes: investigate existing tools and assess best practices; measure costs of ten data curation projects; investigate how much is spent on indirect costs in funded projects; develop test corpora; develop solid tools for versioning; research on understanding, documenting, and preserving curation processes and workflows; identify generic terms for context information; and develop an end-to-end framework for actionable and enforceable data management plans. This report includes notes from those breakout discussions.

In addition to the contributed papers and breakout discussions, the workshop also yielded insights on several high-level themes. These include:

- There are many perspectives on quality: quality assessment will depend on whether the agent making the assessment is a data curator, curation professional, or end user (including algorithms);
- quality can be assessed based on technical, logical, semantic, or cultural criteria and issues; and

- quality be assessed at different granularities that include data item, data set, data collection, or disciplinary repository.

This implies that assessments of quality must carefully specify underlying assumptions and conditions under which the assessment was made. There is movement toward more nuanced models of data control and curation such as maturity levels (matrix models) that consider levels of stability and quality across different criteria and perspectives.

The workshop identified several key challenges that include:

- selection strategies—how to determine what is most valuable to preserve
- how much and which context to include—how to insure that data is interpretable and usable in the future, what metadata to include
- tools and techniques to support painless curation—creating and sharing tools and techniques that apply across disciplines
- cost and accountability models—how to balance selection, context decisions with cost constraints.

Introduction

Lots of information on the Internet may be wrong, including this statement. How do we know what is right? Our measures today are completely inadequate. Scientific data are accumulating at an impressive rate, not just in large archives such as the Virtual Observatory or GenBank, but also in many smaller and less formally maintained systems. How accurate are the data in those systems? How valuable is it to have high quality data? We do not really know today, and we are just beginning to develop processes to find out.

On September 10 and 11, 2012, attendees at a workshop about data quality asked what processes are needed to ensure data reliability and accuracy.

The value of data to the global economy has been well-documented (e.g., McKinsey Global Institute, 2011, World Economic Forum, 2011) and spawned calls for training professionals in data curation and stewardship, data analytics, and 'big data' management. The scientific challenges of digital data have been well-documented by special issues of leading journals such as *Science* (February 11, 2011) and *Nature* (September 4, 2008 Volume 455 Number 7209 pp1-136). In a 2009 editorial, *Nature* charged the scientific community, especially in the US with neglecting data sharing and preservation, suggesting that universities should provide as much attention to ensuring that students acquire data management skills as they do to the acquisition of statistical skills.

Science and scholarship in the 21st century depend on a variety of tools to aid all phases of knowledge generation, sharing, and use. Researchers use electronic devices, sensors, harvesters, and surveys to collect data; databases and spreadsheets to store and manage it; statistical software to perform analyses; text editors to write about results; and networks to transfer all of these elements of research to colleagues, publishers, and the public. Each of these tools creates and uses digital traces, which can themselves serve as part of the scholarly record (e.g., metadata, process control files, audit trails). The general purpose term 'research data' now encompasses the traces of collection, processing, transmission, and use of scholarly work. The impact of digital research data has been recognized on many fronts, two of which have garnered substantial attention in scholarly communities. First, it is argued that data are a primary asset of new research, that aggregation, mining, and reuse of data provide new avenues for scholarly investigation and contribute to what is termed e-Science (e.g., *The Fourth Paradigm: Data-Intensive Scientific Discovery*) or more broadly, e-research. Second, it is clear that there are challenges to managing and preserving electronic data that are essential for all fields to advance (e.g., National Academy Press: *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*). Because digital research data have become so important, funding agencies have begun requiring data management plans that encourage or require data preservation and data sharing; some publishers are requiring deposit of data before accepting papers based upon them; and universities and research laboratories are developing policies, registries, and repositories for research data and products.

All the above developments demonstrate the increasing importance of approaching data from a life cycle perspective rather than treating data merely as a means to conduct a specific study; and to consider this data life cycle within the context of scientific progress rather than as an independent phenomenon. Digital data can accumulate rich contextual and derivative data as it is collected, analyzed, used, and reused, and planning for the management of this history requires new kinds of tools, techniques, standards, workflows, and attitudes. We consider the processes associated with meeting these requirements to be *digital curation*. More specifically, Lee & Tibbo (2007) write: "Digital curation involves selection and appraisal by creators and archivists; evolving provision of intellectual access; redundant storage; data transformations;

and, for some materials, a commitment to long-term preservation. Digital curation is stewardship that provides for the reproducibility and the re-use of authentic digital data and other digital assets."

As science and industry recognize the need for digital curation, scientists and information professionals recognize that access and use of data depend on trust in the accuracy and veracity of data. In all data sets trust and reuse depend on accessible context and metadata that make explicit provenance, precision, and other traces of the datum and data life cycle. High quality data includes procedures that enable verification of quality assertions and procedures that enable parsing and transformations. Poor data quality can be worse than missing data because it can waste resources and lead to faulty ideas and solutions, or at minimum challenge trust in the results and implications drawn from the data. Improvement in data quality can thus have significant benefits.

As part of the data curation problem, we believe that it is urgent that data quality be specifically addressed as more and more systems are developed to preserve and share research data. It is imperative that data creators and curators are able to identify indicators of quality; develop and use tools and techniques that insure useful, usable, and accurate metadata discovery, data ingest, management, and sharing (e.g., painless curation); create and use best practices and open standards whenever possible; and provide auditable validations for data quality.

Workshop Organization and Execution

This workshop aimed to define data quality research issues and potential solutions. The workshop objectives were organized into four clusters:

1. Data Quality Criteria and Contexts. What are the characteristics of data quality? What threats to data quality arise at different stages of the data life cycle? What kinds of work processes affect data quality? What elements of the curatorial process most strongly affect data quality over time? How do data types and contexts influence data quality parameters? To address these questions, the workshop focused on the following goals:

    - identify sets of quality indicators (e.g., authority of source, reproducibility, precision of measure)
    - identify practices and potential standards or types of standards to represent these indicators (e.g., metadata scheme; ontologies)
    - consider how these indicators and representations vary across disciplines
    - consider threats to quality at phases of generation, analysis, storage and management, access, use and reuse, and preservation.

2. Human and Institutional Factors. What are the costs associated with different levels of data quality? What kinds of incentives and constraints influence efforts of different stakeholders? How does one estimate the continuum from critical to tolerable errors? How often does one need to validate data? To address these questions, the workshop focused on the following goals:

    - identify human and technical costs of insuring data quality
    - identify or develop risk models that allow curators to make return on investment (ROI) decisions about curatorial investments

3. Tools for Effective and Painless Curation. What kinds of tools and techniques exist or are required to insure that creators and curators address data quality? To address these questions, the workshop focused on the following goals:

    - identify extant or create recommendations for tools and techniques for selecting data sets for curation
    - identify extant or create recommendations for tools and techniques for automatic metadata generation, annotation (e.g., manual, automatic, crowd-sourced)
    - identify extant or create recommendations for management of data (e.g., ingest, audit, preserve)

4. Metrics. What are or should be the measures of data quality? How does one identify errors? How does one correct errors or mitigate their effects? To address these questions, the workshop focused on the following goals:

    - identify metrics for data quality (associated with criteria in cluster 1)
    - identify techniques for measuring data quality (e.g., appropriate ranges, sampling techniques, probabilities)
    - consider error correction techniques (e.g., interpolation, forensics)

The workshop began with introductions, an overview of the workshop and summaries of the position papers included in this report (see Appendix 3 for the workshop agenda). The workshop participants then broke out into four groups based on the topic areas of their position papers. During this first breakout session, the participants discussed prevalent "pain points" or challenging areas they perceive in data quality. The groups came back together and reported the highpoints of their individual discussions on key challenges (pain points). In a second breakout section, the same groups brainstormed about promising directions to address the research challenges. The promising directions were then summarized and discussed in a plenary session. Finally, projects were proposed based on promising directions and the entire set of possible projects were summarized and discussed. All participants voted on projects to discuss further. On the second day, these projects were discussed and developed. During the discussions, examples from specific data repositories and tools were used by participants to illustrate points.

Discussion Outcomes

Prevalent Pain Points

The following pain points emerged from the four separate group discussions:

- Domain specificity versus general solutions
- Not knowing future uses of data
- Managing access restrictions to sensitive data
- Cost trade-off for high quality data
- Maintaining or improving data quality for replication of research methods
- Knowing how much data to save
- Determining who selects what data gets saved
- Representing the "long tail" of data sets
- Tension between the popular and the important
- Understanding what "quality" means to whom
- Persistent identifiers
- Preserving not just the data, but the software and procedures used to collect it

- Creating generic data quality assessment criteria
- Adding quality assessment to data management plans
- Making data management plans actionable and enforceable
- Creating reliable and reproducible quality assessment metrics

Promising Paths Forward

The Pain Points discussion led naturally into the next group breakout session on potential paths forward in improving data quality management. The following ideas were shared during the breakout group reports:

- Build tools for basic checks and validation of assertions about data quality, for both common and domain-specific needs
- Review and re-appropriate existing tools and processes that have been developed in domain-specific spaces
- Conduct studies to determine where the actual problems lie to ensure that tools are built for real problems versus ones determined by conjecture
- Assess and determine where quality checks and management needs to happen in analysis workflows
- Build more web services that can be used with any repository
- Build tools that make it easy to add and/or extract metadata
- Find low-level common data quality checks
- Quantify what happens when you don't have quality data
- Conduct studies (interviews, focus groups, document analysis) to identify the dimensions of context
- Map data management plan guidelines to existing tools
- Conduct research to determine what percentage of indirect costs in funded projects go to preservation
- Collect success stories and from these identify useful metrics, useful behavior modification, examples of successful ROI, and novel research techniques
- Perform a real world evaluation of the usefulness of metrics (i.e., are data sets used more if they are of higher quality?)
- Explore the potential usefulness of crowd sourcing data quality
- Delineate where we need generalists and where we need domain specialists
- Develop recommendations for skill sets and course material recommendation for training
- Explore methods to determine usefulness of data quality tools

Potential Data Quality Projects

After further discussion of these pain points and paths forward, we asked the group to brainstorm ideas for research that could be conducted that would improve the overall state of data quality. The brainstorming session resulted in a list of twenty-eight projects that were put to vote by the participants. Votes were collected that night and the next morning using a Doodle Poll. From these results, eight projects were selected and groups were assigned to discuss each project in detail. The groups were given 45 minutes to discuss their projects and create a basic outline of a project proposal. This exercise was designed to explore, as a diverse group, real research paths that can be taken to better the state of data quality.

The top eight projects and the proposal outlines that emerged were:

- Investigate existing tools and assess best practices
- Measure costs of ten data curation projects
- Investigate how much is spent on data curation from indirect costs in funded projects
- Develop test corpora

- Develop solid tools for versioning
- Research on understanding, documenting, and preserving curation processes and workflows
- Identify generic terms for context information
- Develop an end-to-end framework for actionable and enforceable data management plans

We carried out two sessions in which participants divided into groups of five or six to discuss the proposed projects (four were discussed in the first session and four were discussed in the second session). The following are notes generated from those small-group discussions.

Investigate existing tools and assess best practices

Introduction:
We need not just bit preservation, but quality preservation

Methodology: Inventory, Classify, Discuss

1. Inventory: look for best practices
We will identify key institutions that do a quality job of quality data management, and interview key staff members.

We look at both tools that MEASURE quality and tools that IMPROVE quality.

We will ask what tools are used, and ask for each tool:
- What is the function of this tool?
- Who are the users?
- Which user performs which function?

2. Classify

- Tools vary by subject domain and by function.
- Their target population may be developers, curators, and/or researchers.
- They may be open source or proprietary.

We will list what kinds of quality improvements that can be made by automated tools, from simple format checking to consistency studies.

We will look both at tools that process data and those that process metadata. Metadata tools, in addition to auditing and preservation, may be involved in metadata extraction or creation.

Provenance tools, and other tools that manage data across time (logging, for example) are particularly important to track. More generally we need tools that operate temporally and enforce consistency and accuracy of data across time.

Policy tools that describe what kinds of operations are allowed or that implement policy changes or audit them, are also important. A particularly important policy question is personally identifiable data and rules of what fields must be concealed or even deleted after particular time lapses. Another policy question is international policy (e.g. which documents can be sent to what copyright regime).

We will look at what characteristics data must have to be processed by the various tools (for example, text files vs. numeric files vs. image files).

Some tools try to reduce not just inadvertent errors but maliciousness; these include spam and virus checking and are needed if public contributions to databases are allowed.

3. Discuss
We will produce a set of "use cases" where stories of tool use are explained and summarized, with key points for future users presented, and as much numerical data on costs and timings included.

4. Recommendations
We will discuss the best tools, what kind of costs and training are involved and what data they apply to, and suggest practices for use to improve data quality.

Measure costs of ten data curation projects

Task: Propose a project that looks at 10 data curation projects and determine cost

1) Methodology
    a. How to even begin to estimate the cost?
    b. Do we look at operational issues?
    c. Prospectively look at data curation projects
        i. How to sample? We want a diverse group of projects that represent different types of projects so we can gather all the variables of data curation processes and cost variables.
        ii. Do we want to do longitudinal study?
        iii. Is 10 too small of a sample size? Do we need to do a pilot study on 10 to inform a larger longitudinal study?
        iv. What variables to we want to focus on to determine the sample?
            1. Size of the project
            2. FTEs
            3. Datasets
            4. Services provided
            5. Metadata provided
        v. Should the pilot study serve as a guide in reporting methodology

    Questions asked by the team:
    - Do we want to decide whether we are tracking through the lifecycle?
    - Do we use Theoretical Sampling:
        - Funding, Discipline, Source, Scale, Individual Inst. Vs. Consortium

2) Second question: Should we care only about cost?
    a. Suggest the need to talk both with Financial Officers and employees involved to get the true story of cost.
    b. Can we get people to disclose costs?
        i. Maybe we can get NSF to write in reporting as a funding approval requirement/also some requirement written in that data mentoring is required for these projects
3) How to approach NSF with a proposal?
4) General questions
    a. Propose to spend 3 structured days with 10 projects where shadow the people involved
    b. Use feedback from structured visits to guide the methodology for longitudinal study and tools to capture the data we need

Questions this raised by the team?

- In one year, can we study lifecycle?
  - The pilot will determine this and then propose longitudinal follow up
- At this point we determined that we do indeed need a pilot

5) Focus change: Do we want to change our focus from just trying to quantify cost to studying methods of tracking cost
   a. Time Sampling Approach
   b. Online Time Management System
   c. Research Time Tracking Approaches

   Another approach is to identify 50 or so tasks and make list of what tasks constitute the data curation process

6) Final Approach

   1) First propose pilot study of structured visits to 10 current projects. Use this as background investigation to determine task list and other variables needed to study
   2) Create tools to address/track these identified tasks and develop methodology to track
   3) Think about profiles that emerged during this pilot study
   4) Develop larger longitudinal study based on the first three steps.

Investigate how indirect costs are used to support digital curation during and after projects

1. Canvas university overhead rates asking how much goes to the library and IT.
We would need someone familiar with university finance and international perspectives.

2. We will review library/IT budgets and attempt to map onto storage costs
We will ask whose responsibility it and whether it is a shared responsibility.
(Caveat: unit costs are likely to be high, e.g. empty Institutional Repositories)

Items to consider:

- Include the cost of ingest into research budget
- Overall Questions: what should be indirect / direct -- research specific vs. general
- Cost not dominated by storage, but by labor

3. What could/should be done: to maintain a catalog of library/data center output as metrics

Items to consider:

- Many repositories are discipline specific and don't show up in the university profile
- Universities have inaccurate information about what it produces; data collection is cumbersome
- Storage costs driven to 0 in this case, but all personnel costs hard to estimate (e.g., partial FTEs)

4. Make recommendations and establish guidelines for appropriate use of direct and indirect funds for data curation and related support and infrastructure services

Develop test corpora

1. Establish a clear purpose of and what types of tests may be performed on the corpora. Possible types of tests can include:

- Testing privacy protection techniques
- Citation analysis
- Significant properties
- File format identification
- Licensing identification
- Process mining
- Data annotation
- Anomaly and mistake detection
- Information extraction
- Classification tasks
- Scaling

Note: the corpora should include known problem files for anomaly detection, and should include a large number of files for scaling tests.

2. Search the landscape for existing test corpora

3. Build the test corpora and provide public access

4. Possibly hold competitions using the test corpora

Develop solid tools for versioning

Examines the problem of tracking version information of large binary files, but also looks into the bigger picture of large data sets within one team or project.

A tool will be developed for the management of large data sets through analysis.

- It will record series of steps and be able to replay/rewind, similar to Photoshop History . (See also: Google Refine)
- It will record conceptual transactions with annotations and will leave data in a useful state
- It will also record and display branching sequences of operations -- tree representations of alternative transformation paths to create data for different analytic purposes

Additional considerations:

- What to do for non-tech-savvy users?
- GUIs for this?
- Does it already exist?
- Format for describing change trees?

Research on understanding, documenting, and preserving processes and workflows

Introduction:
Institutions with lots of data need an organized, formal way to deal with it.

Methodology: Capture, Organize, Discuss

1. Capture: look for best practices
We will identify key institutions and understand what their workflow process is, interviewing key staff members.

We'll try out the formal workflow languages to see how applicable they are and where they are inadequate.

We need to report on which workflow designs do the best job at maintaining and improving data quality. The workflow process must also preserve provenance and an audit trail.

Workflows must extend to the steps taken by the researchers gathering the data and to the users, as well as the full-time curatorial staff. Workflows must enforce the creation or capture of the information required for curation, such as metadata, provenance, and temporal data.

2. Organize
Practices cover collection, storage, output, and re-use. We will organize practices for all of these. We care about temporal effects: how workflows deal with data over time.

Workflows are classified by domain, looking for similarities and divergences.

Workflows would also be classified by kind of institution (large public, university, private).

What are the gaps in scientific data workflow? Can we use commercial data management processes to help?

Researchers have needs to do particular analyses: we need to connect this to workflows to be able to assure researchers that they can do those analyses.

Workflows must be evaluated to determine whether important policies are maintained (eg data privacy or data validation). How do workflows ensure consistency and accuracy, and how do we know that they do so?

Some workflows, for example with crowdsourced data, must include defenses against malicious content (spam or viruses). All workflows must provide steps to deal with failures, bad data, and support archiving, rollback, and other data management processes.

Good workflows track steps for future auditing: this must be done clearly and easily.

3. Discuss
We will produce a set of "use cases" where workflows are described with their advantages, costs, and risks.

We are particularly interested in the possibility of providing a unified model which covers the best workflows but can be specialized to particular archives.

4. Recommendations
We will compare and contrast the best workflows for data quality assurance and recommend processes.

Identify generic terms for context information

Motivation: Users need to know about context in order to evaluate data. What types of contextual metadata are required?

Types of contextual information:
- Instrumentation (devices, survey instrument, scale of measurement)
- Administrative

- Descriptive
- Access Restrictions / Rights
- Environment in which data were collected
- Preservation - actions taken, decisions made

Study to investigate users who are outside of the original data domain to see what further contextual information they need to make meaningful use of the data.

Potential examples to explore:
- Polar bear researcher trying monitor snow and ice data sets to use in order to determine where the bears are
- K-12 classroom use of data sets
- Could focus on DataNet projects

Potential research methods:
- Interview people doing interdisciplinary research to see what issues they're confronting
- Experimentally test what types of contextual information actually help people perform tasks

Develop an end-to-end framework for actionable and enforceable data management plans

Design a software tool to help with data planning activities and implementation that is simple and easy to use. (Much like the TurboTax online GUI).

It will:
- Have an actual case study to inform design
- It will be a modular design with an open framework and discipline specific modules

The project team will establish a direct relationship with major funders to know what their requirements are.

Tasks of software:

Planning
- Captures requirements against framework. (see DCC tool like this that generates checklist.)
- Should be adaptable and will be designed in an iterative process

Implementation tasks
- Cross linking
- Standards for reporting to agencies
- Tracking citations
- Show sharable equipment
- Contain and display products of the project

Additional Notes:
- Carrots and sticks are useful to get people to use tools and planning
- Could also establish relationships with data repositories to allow easy depositions of datasets
- Might be possible to consider consolidated data planning services for smaller institutions

Conclusion

The project proposal presentations were concluded by final plenary discussion. We revisited where we were when we started the workshop and where we have found ourselves after the two-day journey. Throughout the entire process, we were able to establish a deeper understanding of the problems that face us in managing the quality of the data that is being collected across the globe and we were able to clear some paths to move forward in addressing these challenges. From all of this, we were able to distill these truths:

- We don't truly know what our data quality is today
- We need cooperative processes between creator, curator, and user
- Data curation should be as painless as possible

Major conclusions were:

Context

The chain from data capture through data curation to data users is too loose, and we need more and tighter interaction. Even defining "quality" without knowing the purpose of the data is difficult. Efficient capture of data including provenance and metadata is most easily done by working at the start of the process, not trying to retrofit quality in later. Later on, the aggregation of multiple databases often highlights errors that may have been overlooked in a single database, a problem aggravated by our lack of metrics for even separated areas.

Accounting

Few projects track their curation costs, and since many projects also do not measure the number and size of errors in their archive, we can not plan how much we should spend on quality assurance to achieve a given level of reliability. Nor do we yet have an understanding of how these costs will be covered, with research budgets, university administrative budgets, and library budgets all under pressure and competing for the same resources.

Technology

We lack toolkits for both quality management and workflow description. Different projects do not share expertise in essential activities such as auditing, provenance, and privacy policy. Tools are needed both for the actual data and for management of the metadata.

Selection

The explosion of sensor capacity is outrunning the increase in disk capacity; one estimate is that to save every bit in the world would, by 2018, require that the entire gross world product be spent on disks. We do not understand the tradeoff between more data and better data nor do we have a general model of tools to implement selection policies. It seems evident that observational data that cannot be replicated should be curated with higher priority than data that is replicable. Although no clear conclusions were made about who should make selection decisions, it seems reasonable that data creators should be most engaged with data elements and data curators with collections of data sets.

Specialization

Do different disciplines require different procedures? Mechanically collected sensor data has different errors than survey data, and databases involving people create privacy issues. Nevertheless there should be procedures that are shareable across domains. Can we distinguish areas, or even individual data items, which can be postponed until their importance to users can be better evaluated, from data items which must be captured at source if they are not to be gone forever?

Call to Action

The workshop project discussions raised many issues that demand research and development action. The following set seem most imperative and first steps to enhancing research data quality and use.

- Collect best practices, best tools, and best workflow from successful and well-managed archives. Explore the generalizations of these across domains and attempt to model the subject limitations of general processes. Press for increased automation of data curation including metadata creation.

- Document quality and its impact. If our data were half as accurate, what would we not know? What are visible and important results derived from well-maintained archives, such as our ability to document climate changes and to evaluate long-term impacts of pharmaceuticals or diet? Define metrics and estimate economic benefits from improved quality.

- Define policies we need to implement for selection, auditing, provenance tracking, temporal consistency, privacy, and visualization. How does quality relate to interoperability, when "good enough" for one purpose might not be good enough for another?

- What processes will most effectively and economically improve quality? Does more use create better quality, or is it the reverse? Now that NSF is creating a system of public data exchange, how can we manage it for best quality and best results?