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?

Appendix 1. Position Papers

Position Papers: 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 curatiorial process most strongly affect data quality over time? How do data types and contexts influence data quality parameters? To address these questions, the workshop will:

- 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.

Mitigating Threats to Data Quality Throughout the Curation Lifecycle[1]
[Draft 10/1/2012]

Micah Altman
<http://micahaltman.com>
Director of Research -- Libraries, Massachusetts Institute of Technology

Introduction: Measurable "Data Quality" is Field-Specific

What is 'good' data? The answer typically varies from field to field, and from application to application. Nevertheless, several fields have, independently, developed frameworks that aim to identify (and, in some cases, measure) the attributes that simultaneously are *independent of the specific subject area, domain of measure, parameterization, and specific semantic content* of information and that e*nhance the value of and/or fitness for use of* that information.

The specific term "data quality" is most commonly used in the discipline of Management and Information Sciences (MIS), where it is has become defined generally as *fitness for use* [Madnick, et. al 2009] – while, in other fields, terms such as "value of information", "information content", "reliability" and "validity" are used to describe analogous assessments of data (see Table 1, in the following section). These general frameworks for assessing the quality of data vary widely in theoretical foundations and in practical application. "*Data Quality" is an overloaded term*: In scholarly practice, the term "quality", as applied to data, is most often used in a general, inexact and ambiguous way:

Existing Information Quality Frameworks are Diverse

Developing a discipline-independent definition of data quality that is useful, consistent, and reasonably comprehensive is challenging. Within MIS (and to a lesser extent, within Computer Science) there have been numerous attempts to define "data quality" generally. As Wand and Wang [1996] note, these attempts have been based on diverse methodologies including intuitive understanding, industrial experience, surveys of practice and use, and literature reviews. Wand and Wang further note that within this literature, there is no general agreement on the dimensions or attributes of data quality.

In response to this general lack of agreement within MIS, Wand and Wang develop an ontological model of data quality, in which they conceptualize quality as a mapping between an information system and true states of the world. This ontological model yields the intrinsic dimensions of completeness, unambiguity, meaningfulness, and correctness.

[1] This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

Price and Shanks [2005], however, note continuing disagreement in this area, and develop a separate semiotic approach to extend and reconcile Wand and Wang's proposed dimensions. This approach separates "data quality" into syntactic, semantic, and pragmatic categories -- which are characterized (respectively) by conformance to metadata, correspondence to external phenomena, and data use-value.

Both Wand & Wang [1996] and Price & Shanks [2005] are well worth studying, but these models have not been widely applied, and are formulated at a high level of abstraction that resists direct measurement. Moreover, other disciplines have developed alternate frameworks for the generic evaluation of information, such as: information theory in applied mathematics; decision theory in economics; measurement theory and classical test theory in psychometric; and estimation theory in statistics. These frameworks, although not labeled "data quality", overlap significantly with data quality concepts: Each of these theories aims to measure the attributes and/or values of information that are independent of the specific subject, measures, and semantics of that information. **Table 1** summarizes these approaches -- and also demonstrates the divergence in the specific data quality attributes used within each framework.

Framework	**Field**	**Evaluative Categories & Quality Dimensions**
Mathematical Information Theory [For a summary, see Weshler & Ho 2011; Burgin 2003]	Applied Mathematics, Computer science	1. Shannon Entropy – mathematical measure of unpredictability 2. Kolmogorov Complexity -- measure of information complexity/compressibility 3. Fisher Information (or simply "information") -- measure of the amount of information that a random variable carries about an underlying parameter value
Measurement theory, generalization theory, test theory [For a summary, see Raykov and Marcoulides 2010]	Psychometrics; Education Research; Social Science	1. Scales of measurement -- measurement scale types (nominal, ordinal, interval, ratio) determine permissible statistics, and admissible transformations 2. Reliability – consistency of repeated measure under consistent observable conditions; precision a) inter-rater, b) test-retest, c) inter-method, d) internal consistency 3. Validity – correspondence between measure as implemented and concept it purports to measure; accuracy a. internal validity, b) predictive validity, c) construct validity 4. Measurement error – bias and variance introduced by measurements
Statistical Inference [For a summary from a Bayesian perspective see Gelman, *et* al. 2004]	Statistics	1) Random error 2) Sampling variance 2) Measurement/observational error 4) Missing data/non-response
Value of Information [see A.J. Repo 1989 for a summary]	Economics	1. Equilibrium Analysis -- expected effect of information on predictive equilibrium of economic market 2. Statistical Decision Theory – expected difference in outcome states resulting from obtaining additional information.[2] a. decision system (actions, states, signals, outcomes), c) prior probability distributions over states, d) function mapping data (signals) to posterior distribution over outcomes 3. Multidimensional Value – descriptive/heuristically chosen attributes a. uncertainty, b) diffusion (affecting scarcity), c) applicability, d) content, e) decision relevance
Ontological Analysis of "Data	MIS	1. Intrinsic data quality: properties of mapping

[2] Statistical decision theory, in essence, applies classical or Bayesian statistical theory to a decision tree or model of the decision problem to evaluate the potential change in optimal outcome yielded by additional information.

Quality” [Wand & Wang 1996]		between data and real world states a. Completeness: all states of worlds are represented ; b. Unambiguity: 1-to-1 mapping between states of world and model state c. Meaningfulness: no model state that does not correspond to potential state of world d. Correctness: model state maps to correct state of world 2. 2. Internal View: design & operation a. Data Related: (i) accuracy, (ii) reliability, (iii) timeliness, (iv) completeness, (v) currency, (vi) consistency, (vii) precision b. System-related: (i) reliability 3. External view: use, value a. Data-Related: (i) timeliness, (ii) relevance, (iii) content, (iv) importance, (v) sufficiency, (vi) useableness, (vii) usefulness, (viii) clarity, (ix) conciseness, (x) freedom from bias, (xi) informativeness, (xii) level of detail, (xiii) quantitativeness, (xiv) scope, (xv) intepretability, (vxi) understandability b. System-related: (i) timeliness, (ii) flexibility, (iii) format, (iv) efficiency
Semiotic Analysis of “Data Quality” [Price & Shanks 2005]	MIS	1. Syntactic: conformance to metadata 2. Semantic: correspondence to external phenomenon a) Completeness, b) (Un)ambiguity. c) Correctness, d) Non-redundancy, e) Meaningfulness 3. Pragmatic: use value a) Perceived rule conformance, b) Perceived reliability, c) Perceived completeness , d) Understandability, e) Accessibility, f) Security, g) Flexibility of presentation, h) Suitability of presentation, i) Relevance, j) Value
Literature Review/Practitioner Survey [see Lee, Kahn, Strong and Wand, 2002 for a summary[3], also see Knight and Burn 2005, for an alternative survey of frameworks that substantially overlaps]	MIS	1. Intrinsic: quality of information in its own right accuracy, believability, completeness, consistency, correctness, credibility, factualness, freedom from bias, objectivity, reliability, reputation, unambiguity, validity 2. Contextual: quality within task context accuracy, appropriate amount, completeness, consistency, correctness, currency (general, source currency, data warehouse currency, cycle time), essentialness, level of detail (general, attribute), quantity, reliability, timeliness, usage, validity, value-added 3. Representational: quality of representation within information system ability to represent null

[3] Also note that this Lee, *et* al. includes a summary and comparison the framework developed by Wand & Strong, which is cited as Wang &Strong [1996] several other contributions to this workshop.

		values, arrangement, appropriate representation, comparability, compatibility, concision, consistency (general, semantic, representational, structural), definitional clarity, efficiency (storage), format flexibility, homogeneity, identifiability, interpretability, lack of confusion, meaning, meaningfulness, metadata characteristics, naturalness, portability, precision (format, domain), presentation, readability, reasonableness, redundancy, semantics, syntax, understandability, uniqueness, version contro 4. Accessibility: quality of access within information system accessibility, assistance, availability (system, transaction), ease of use (general, operational, h/w, s/w), locatability, flexibility, reliability (delivery), obtainability, privacy, privileges, quantiativeness, usableness, robustness, security

Table 1: Data Quality Frameworks

Within specific sub-fields, data quality is sometimes understood to comprise a set of highly specific and measurable attributes. However, these attributes are generally so closely-tailored to the particular needs of a specific sub-field that they defy ready generalization to other fields. The definition of quality in public opinion survey research is illustrative.

The central concern of the sub-field of public opinion making inferences about characteristics of a population of individuals based upon measurements applied to a sample from that population. (This statistical inference problem is shared with a number of other fields of research, although the specific populations of interest and measurements applied differ.) Within both the theory and practice of public opinion survey research, there is a particularly broad consensus, developed over the last three decades, concerning the sources of error and threats to 'quality' related to survey data and subsequent inference.[4] [Generally, see Groves 1989; Biemer & Lyberg 2003; Weisberg 2005] These sources of error may be divided into three categories: sampling error, non-observation error, and errors of observation: Sampling error is the statistical uncertainty that results from estimating population characteristics based on a sample, due to sample-to-sample variation. Errors of non-observation comprise a set of errors that reduce the correspondence between the sample obtained and the intended population of inference, including non-response (both at the measure/item and unit/subject level), non-compliance, and coverage error. Errors of observation, or measurement errors, are introduced by imperfect measurement processes, which yield observed measurements that are not perfectly precise, valid, and reliable summaries of the underlying quantity of interest. There is general agreement in the field of survey research that high-quality surveys minimize total survey error (which is

[4] Much of this framework can be viewed as an application of *statistical population inference* and *psychometric generalization theory* to the domain of public opinion measurements. See Table 1 below.

most often operationalized as mean square error of estimates) subject to cost, technical and legal constraints.

Despite a wide theoretical consensus, however, systematic measurement and reporting of non-sampling error remains quite rare in the practice of public opinion surveys. And total survey error is, with the exception of the largest government surveys, at best assessed informally, and more typically ignored altogether. This is in large part because reliably quantifying the various sources of non-sampling error is quite challenging in practice.

Data Quality Frameworks Used in this Workshop are Diverse

The position papers submitted for this workshop illustrate the diversity of approaches to characterizing data quality. As **Table 2** illustrates, most of the responses are placed within one of the frameworks above, but there no single framework that accommodates all of the various positions. The positions papers also underscore the difficulty of systematic measurement and evaluation of quality – few propose objective or quantifiable measurements, and those proposed measurements that are readily quantifiable are limited to small subsets of the data-quality attributes that are potentially relevant.

Position Paper	**Explicitly Defined Quality Frameworks**	**Additional Implicit Frameworks**
Duerr	Measurement error; Use-Value (citing Palmer, et al 2012)	Use-value for designated community (related to fitness for use; semiotic – pragmatic value; ontological – use value)
Fiore	Use-Value (citing Wang & Strong 1996)	Ontological (Intrinsic; Representation; and Use—Accessibility)
Lesk	Use-Value -- Replication	Fitness for use – Accessibility; Pragmatic (Replicability; Provenance; Open Data)
McDonough	Use-value -- sensemaking	Fitness for use – Accessibility (locatability, persistence); Contextual;
Mayernik		Fitness for use – Pragmatic
Johnston		Fitness for use – Intrinsic data quality; Accessibility
Moore	Reliability of creator asserted properties. Fitness for use.	Fitness for use – Syntactic ; Pragmatic
Young, et. al		Fitness for use – Internal View/Representational
Ashley	Use-Value (citing Wang & Strong 1996)	Fitness for use – Intrinsic, Conceptual, Representational, Accessibility
Conway	Measurement error relative to expected use	Fitness for use
Giarlo	Trust, Authenticity, Understandability, Usability, Integrity (citing Knight & Burn, 2005)	Fitness for use – Representational (understandability); Accessibility (usability); Intrinsic (Integrity); Pragmatic (Authenticity); Pragmatic (Trust)

Table 2: Data Quality Frameworks Corresponding to Submitted Position Papers

Mitigating Threats to Quality Through the Curation Lifecycle

The data quality criteria implied by the candidate frameworks are neither easily harmonized, nor readily quantified. A generalized systematic approach to evaluating data quality seems unlikely to emerge soon. Fortunately, developing an effective approach to digital curation that respects data quality does not require a comprehensive definition of data quality. Instead, we can appropriately address "data quality" in curation by limiting our consideration to a narrower applied question:

> *Which aspects of data quality are (potentially) affected by (each stage of) digital curation activity?*

Digital curation is fundamentally concerned with maintaining information assets and access to them over a medium-to-long term. Narrower conceptualizations of digital curation typically focus on the the following activities:

- storage and disposal
- format (preservation) transformations
- access (including discovery and reuse)

Broader conceptualizations of digital curation also include ingest; (re)appraisal and (re)selection; disposal; and creation of derivative works. These activities interact with data quality in distinct ways.

Appraisal & selection. It is difficult to imagine an effective appraisal and selection process that entirely ignored data quality. In formal terms, the goal of appraisal and selection is to maximize the expected future value of access to a collection, by and for designated target communities, subject to resource (e.g. cost) and feasibility constraints. For this set of curation activities, the general characterization of data quality by the field of MIS, as "fitness for use", seems apt. Notwithstanding, the practical evaluation of data quality for appraisal and selection purposes is necessarily tied closely to a particular set of actual and potential uses and users, and hence to the attributes that facilitate those uses. Thus it appears that a a general definition of data quality can provide little guidance in this area.

Furthermore, it should be noted that the appropriate selection of objects is i not primarily a function of the quality of those objects, in many cases. This is the case regardless of how one measures quality of those objects (and regardless of the cost of their curation), for two reasons: The value of a collection is determined not only by the properties (including quality) of the individual objects but of the properties of the entire set of objects in the collection (such as completeness). Second, the future value of individual objects and of the collection itself is uncertain, and thus an implication of modern portfolio theory [see for a summary, Elton, Gruber 1997] is that any optimal selection strategy will diversification against risks to future value -- and not simply selection of the individually "best" objects.

Short term storage and disposal. Short-term storage can be characterized in terms of maintaining invariant properties for a representation of an information object. Typically the properties maintained by storage systems include its "bit sequence", and a minimal set of associated metadata elements, such as creation time and creator. These properties are maintained in the storage system through fixity computations, replication of information, auditing/detection of corruption, and repair. As long as a storage system maintains these standard invariants, the effect of the storage system on data quality properties should be neutral, and the choice of a particular short-term storage system can be separated from general management of data quality.

Long-term (preservation) storage and properties. Long-term preservation can be usefully viewed more broadly, as Moore [Moore 2008] points out, as communication of information from the past to the future. Future consumer of the information must be able to understand it, and to validate statements about the relationship of the information received to past communicative actions. Specifically, the field of preservation is traditionally concerned with validating properties related to information authenticity, organization (respect des fonds), chain of custody, and integrity. [Moore 2008, pg. 69]

It is tempting to label these properties “preservation quality” attributes, as these preservation attributes are likely to be associated with a higher expected value for future use (or, to put it another way, with increased “fitness for use”), and the effect of these properties is largely independent of the specific semantic content of the digital information objects. Note however that evaluating these attributes is complicated by the fact that, in practice, these preservation quality attributes are often largely determined by the properties of the preservation system acting on the objects, rather than on the properties of the digital objects themselves at the time of selection, appraisal, and ingest. Notwithstanding, efforts to assess and manage data quality for curation should give substantial weight to these preservation quality attributes.

Ingest; format transformation; preservation actions; access; and derivative works. Even absent a comprehensive definition of data quality, it is clear that these stages of curation can threaten potential data quality properties (if not enhance these properties). In particular, the ability of the target community to make use of an information object or collection, regardless of its semantic content of the object, could readily be affected by reduction of accuracy (or loss of fidelity, or introduction of random noise); failure to maintain the relevant semantics (abstract information content) of the object; or failure to record sufficient context (e.g. documentation & metadata) during these curatorial phases. This implies that an important consideration in maintaining the qualit of the object during curation is to maintain the semantic content of the object, or to directly measure and control the loss of semantic information.

For example, format obsolescence is a well-known threat to preservation -- communication with the future will fail if the future receiver can no longer understand how to interpret the sequence of bits making up the message. Several general approaches have been proposed to mitigate the threat of semantic information loss from format obsolescence [Lee et. al 2002] , including encapsulation, standardization (such as through universal formats or universal virtual machines), emulation, and migration. Of these approaches, format migration is the most commonly used -- and format migration is also generally regarded as the most practical of these, at least for the time being. However, format migration raises questions about the quality of the translation: How do we confirm that it means approximately the same thing as the original? How do we measure the closeness of approximation? And how does the community accessing the information confirm that they have understood it correctly (in other words, that their internal representation the information retains its semantic properties)?

Format migration typically rely primarily on manual processes to verify the fidelity of the final product. (This is true of digitization practice as well.) Current good practice involves manually verifying the integrity of the conversion process. This is labor intensive, Even when only a sample of material is checked,, and many errors will go undetected. So, despite their popularity and practical value, both format transformation and digitization introduce the potential for silent information loss, loss of fidelity, or corruption during the transformation. Because of the complexity of file formats and the imperfect nature of software, errors during transformation are relatively common. This lack of systematic quality assurance threatens both the evidence base for research, and our continued access to the nation’s cultural heritage.

One potential solution to maintaining quality through ingest, transformation, and access is to formally identifying and characterize the properties of cultural and scholarly objects that are relevant to use by the target community; and to develop “semantic fingerprint” algorithms to provide automated, quantifiable, measurements of semantic fidelity in the representation of an object. Semantic fingerprints are 'smart', and make use of the *meaningful characteristics* of the object to create a fingerprint.

A wide variety of semantic fingerprint algorithms have been developed in the commercial sector to aid in resource discovery, or in the application of digital rights management to user-submitted content. For example, acoustic fingerprints capture key perceptual qualities of a piece of music -- enabling music search services like "SoundHound" to help one find tens of millions of songs just by humming a fragment; while video fingerprinting algorithms enable Youtube to identify copyright protected content that was contributed independently of the righst holder (possibly in a different format, in fragmentary form, and with substantially reduced fidelity). Semantic fingerprints are also used to ensure the integrity of scientific data in some digital library systems. [Altman 2007; Altman & King 2007] These methods are powerful, efficient, and flexible within their domains-- but they have not necessarily designed,or calibrated to capture the properties of objects relevant to digita curation. Concerted research and development efforts in this area has the potential to develop applied fingerprint techniques that will mitigate risks to quality during digital curation.

Conclusions

To summarize, it is unlikely that rigorously measurable and discipline-independent measures of data quality will soon emerge. Nonetheless, we can identify threats that affect a wide-spectrum of potential quality properties, and systematically address these within curation activities. Some potential approaches include the following:

- Incorporate portfolio diversification in selection and appraisal.

- Support validation of preservation quality attributes such as authenticity, integrity, organization, and chain of custody throughout long-term preservation and use -- from ingest through delivery and creation of derivative works.
- Apply semantic fingerprints for quality evaluation during ingest, format migration and delivery.

These approaches are independent of the content subject area, the domain of measure, and the particular semantics content of objects and collections -- so they are broadly applicable. By mitigating these broad-spectrum threats to quality, we can improve the overall quality of curated collections, and their expected value to target communities.

References

M. Altman, G King, 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data, 1082–9873. In DLib Magazine 13 (3/4)

M. Altman 2007. A Fingerprint Method for Scientific Data Verification, 311-316. In *Proceedings of the International Conference on Systems Computing Sciences and Software Engineering* 2007.

P.P. Biemer, 2003. L.E. Lyberg, Introduction to Survey Quality, John Wiley & Sons.

M. Burgin, 2003, “Information Theory: a Multifaceted Model of Information”, *Entropy* 5: 146-160.

K.H. Lee, O. Slattery, R. Lu, X. Tang, V. McCray, 2002, “The State of the Art and Practice in Digital Preservation”, *Journal of Research of the National Institute of Standards and Technology* 107: 93-106.

E.J. Elton, M.J. Gruber, 1997, “Modern Portfolio theory, 1950 to date”, Journal of Banking & Finance, 21(11-12): 1743-1759.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, 2004. *Bayesian Data Analysis*, Chapman & Hall.

R.M. Groves, 1989. *Survey Error and Survey Costs.* John Wiley and Sons.

Y.W. Lee, D.M. String, B.K. Kayn, R.Y. Wang, 2002. “AIMQ: a methodology of information quality assessment”, *Information & Management* 40:133-146

S. Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. Informing Science, 8:159–172, 2005.

S.E. Madnick, R.Y. Wang, Y.W. Lee, H. Zhu, 2009. “Overview and Framework for Data and Information Quality Research”, ACM Journal of Data and Information Quality 1(2) 1-22

Palmer, C. L., Weber, N. M., and M. H. Cragin. 2011. The Analytic Potential of Scientific Data: Understanding Re-use Value". Proceedings of the American Society for Information Science and Technology 48: 1-10

T. Raykov, G.A. Marcoulides, 2010. *Introduction to Pscychometric Theory*, Routledge Academic.

R. Moore, 2008. "Towards a Theory of Digital Preservation", *The International Journal of Digital Curation* 1(3): 63-75.

A. J. Repo, 1989, The Value of Information: Approaches in Economics, Accounting and Management Science, *Journal of the American Society for Information Science* 40(2):68-85

R. Price, G. Shanks, 2005. "A Semiotic Information Quality Framework: development and comparative analysis", *Journal of Information Technology* **20:** 88-102

B Stvilia, 2007, *Measuring Information Quality*, University of Illinois at Urbana-Champaign, Ph.D. Thesis.

Y. Wand, R. Y. Wang, 1996, "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the Acm 39(11)* 86-95.

Wang, R.Y., and Strong, D.M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12 (4): pp. 5–33.

H. Weschsler, S-S. Ho, 2011, "Intelligent Evidence-Based Management for Data Collection and Decision Making Using Algorithmic Randomness and Active Learning", *Intelligent Information Management* 3: 142-159.

H.F. Weisburg, 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, University of Chicago Press.

NOAA’s National Climatic Data Center’s Maturity Model for Climate Data Records

John J. Bates
NOAA’s National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001
(828)271-4378
John.J.Bates@noaa.gov

Jeffrey L. Privette
NOAA’s National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001
(828)271-4331
Jeff.Privette@noaa.gov

Alan D. Hall
NOAA’s National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001
(828)271-4071
Alan.Hall@noaa.gov

ABSTRACT

In this paper, we describe the NOAA’s National Climatic Data Center’s (NCDC) Maturity Model for assessing a Climate Data Record (CDR). We will describe the model and process for transitioning a CDR from research to operations and methods for maintaining quality of the data.

1. INTRODUCTION

A Climate Data Record (CDR) is a time series of measurements of sufficient length, consistency, and continuity to determine climate variability and change. Observational data sets and methods that are used to study these phenomena evolve over time. Here we consider how the scientific community can provide the needed objective information on data sets and methods that are required by those wanting to use the data for a specific application or by future generations.

2. BACKGROUND

Objective information is needed in the management of large and complex systems, an issue the engineering community has wrestled with for several decades. In the 1990s, NASA formally adopted the concept of Technical Readiness Level (TRL) to formally capture the progression of steps in turning a basic research concept into a fully operational product. In this maturity model, TRL 1 is when basic research has taken the first steps toward application and TRL 9 is when a technology has been fully proven to work consistently for the purpose designed and is operational (Banke, 2010).

Similarly, the software industry has widely adopted the Capability Maturity Model Integration (CMMI) as a maturity model for effective development of reproducible software processes. The CMMI model has five steps from level 1; denoting processes are unpredictable, poorly controlled and reactive, to level 5, focusing on deliberate process optimization/improvement (Humphrey, 1988).

These maturity models from the systems engineering communities provide the basis for quantifying the maturity for CDRs.

3. CDR MATURITY MODEL

3.1 The need for a maturity model

The need for a maturity model for climate records first arose in discussions between NASA and NOAA when considering how NASA Earth Observing System (EOS) climate instruments might transition to NOAA for long-term sustained observations. For this research to operations transition, we took the traditional approach, using the so-called dynamic linear model, for the transition of basic research through applied research and development and into operations. The most widely known application of this dynamic linear model is within the US Department of Defense's categories for research and development (Stokes, 1997).

As the scientific and socio-economic impacts of climate variability and change have increased over the past decades, the need to ensure transparency as well as full and open access to all data sets and all processes used to create such products has greatly increased. A framework for assessing the completeness and maturity of CDRs is required to support the broad range of potential users both now and in the future.

The proposed CDR maturity matrix thus combines best practices from the scientific community, preservation description information from the archive community, and software best practices from the engineering community.

3.2 The model

There are six thematic areas for assessment, based on discussion and feedback from many scientists over the past several years. These include: software readiness, metadata, documentation, product validation, public access, and utility. Each of these thematic assessment areas is expanded into six levels of completeness, or maturity. These maturity levels capture the business practices that have arisen over the past two decades in fielding climate observing systems, particularly satellite observing systems. Maturity levels 1 and 2 are associated with the analysis of data records from new instruments or a new analysis of historic observations or proxy observations which is designated research. This may be seen as an analogue to the initial commissioning of an observing system. Although products at this stage of development may show interesting results, there is insufficient maturity of the product for it to be used in decision making. Initial operational capability (IOC) is achieved in maturity levels 3 and 4. At these levels, the product has achieved sufficient maturity in both the science and applications that it may tentatively be used in decision making. Finally, full operational capability (FOC) is achieved only after the product has demonstrated all aspects of maturity are complete.

3.3 Verifiable Standards

Quantifiable standards can, or will, exist for each thematic area and each maturity level. For example, peer reviewed publications are required in three separate areas to address product documentation, validation, and utility. Particular attention is also paid to software maturity and access. This includes requiring the code to be managed and reproducible, metadata with provenance tracking and meeting ISO standards, and all code publicly accessible. Uncertainty of the product must be documented, assessed by multiple teams and positive value demonstrated. Each of these steps must be independently verifiable.

3.4 Assessment of the Maturity Model

Scientists in several academic institutions and governmental agencies have performed a self-assessment using this and earlier drafts of the maturity matrix. Their feedback has been incorporated into the current version 4. The first formal assessment of CDRs occurred in April of 2011 by the World Climate Research Programme (WCRP) Observation and Assimilation Panel (WOAP). Eight satellite observation products, covering the atmosphere, ocean, and land surface, were evaluated. This assessment was the first to apply an independent standard to CDRs across disciplines. It used the maturity matrix as part of a more comprehensive discussion applying the Global Climate Observing System (GCOS) guidelines for climate data record preparation. Perhaps one of the most important outcomes was simply having interdisciplinary discussions of CDR maturity against a common standard. Most of the feedback on the maturity matrix concerned interpretation of the terms used. These discussions have moved forward the adoption of more precise language and standards and templates for the elements of the matrix. This maturity matrix model may serve in the future as a requirement for use of data sets in international assessments or in other societal and public policy applications similar to certification programs that engineering professions conduct.

4. CONCLUSIONS

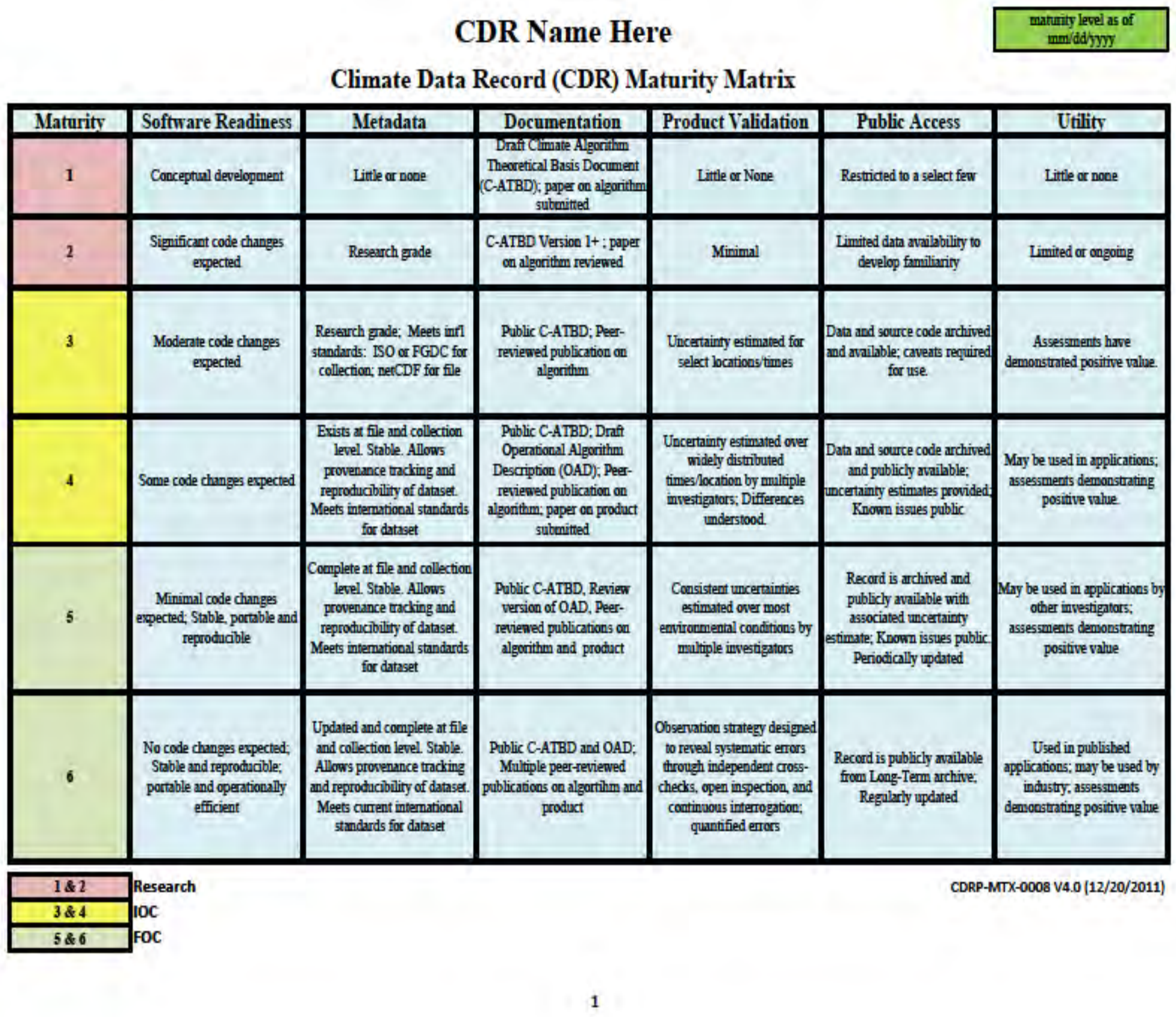
The demand for climate information and its broad application across many socio-economic sectors requires that geophysicists adopt more rigorous standards that are more typically found in the systems engineering community. In this article, a maturity matrix for climate records has been proposed. This notion is similar to that of the NASA technical readiness levels and the software industries capability maturity model. Adoption of such a standard by the climate community would help ensure transparency and traceability of climate records and facilitate their use across both natural and social science disciplines. It would also help spur a re-examination of the traditional research to operations paradigm, as advocated by Stokes (1997).

5. REFERECNES

[1] Banke, Jim (20 August 2010)."Technology Readiness Levels Demystified". NASA.

[2] Humphrey, Watts (March 1988). *Characterizing the software process: a maturity framework*, *IEEE Software* **5** (2): 73–79.doi:10.1109/52.2014

[3] Donald E. Stokes, *Pasteur's Quadrant – Basic Science and Technological Innovation*, Brookings Institution Press, 1997.



CDR Name Here

maturity level as of mm/dd/yyyy

Climate Data Record (CDR) Maturity Matrix

Maturity	Software Readiness	Metadata	Documentation	Product Validation	Public Access	Utility
1	Conceptual development	Little or none	Draft Climate Algorithm Theoretical Basis Document (C-ATBD); paper on algorithm submitted	Little or None	Restricted to a select few	Little or none
2	Significant code changes expected	Research grade	C-ATBD Version 1+ ; paper on algorithm reviewed	Minimal	Limited data availability to develop familiarity	Limited or ongoing
3	Moderate code changes expected	Research grade; Meets int'l standards: ISO or FGDC for collection; netCDF for file	Public C-ATBD; Peer-reviewed publication on algorithm	Uncertainty estimated for select locations/times	Data and source code archived and available; caveats required for use.	Assessments have demonstrated positive value.
4	Some code changes expected	Exists at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset	Public C-ATBD; Draft Operational Algorithm Description (OAD); Peer-reviewed publication on algorithm; paper on product submitted	Uncertainty estimated over widely distributed times/location by multiple investigators; Differences understood.	Data and source code archived and publicly available; uncertainty estimates provided; Known issues public	May be used in applications; assessments demonstrating positive value.
5	Minimal code changes expected; Stable, portable and reproducible	Complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets international standards for dataset	Public C-ATBD, Review version of OAD, Peer-reviewed publications on algorithm and product	Consistent uncertainties estimated over most environmental conditions by multiple investigators	Record is archived and publicly available with associated uncertainty estimate; Known issues public. Periodically updated	May be used in applications by other investigators; assessments demonstrating positive value
6	No code changes expected; Stable and reproducible; portable and operationally efficient	Updated and complete at file and collection level. Stable. Allows provenance tracking and reproducibility of dataset. Meets current international standards for dataset	Public C-ATBD and OAD; Multiple peer-reviewed publications on algorithm and product	Observation strategy designed to reveal systematic errors through independent cross-checks, open inspection, and continuous interrogation; quantified errors	Record is publicly available from Long-Term archive; Regularly updated	Used in published applications; may be used by industry; assessments demonstrating positive value

1 & 2	Research
3 & 4	IOC
5 & 6	FOC

CDRP-MTX-0008 V4.0 (12/20/2011)

1

Figure 1. Climate Data Record Maturity Matrix.

Data Quality: On the Value of Data

Ruth Duerr
National Snow and Ice Data Center
University of Colorado
Boulder, CO 80309
011-1 (303) 735-0136
rduerr@nsidc.org

ABSTRACT

In this paper, an attempt is made to define “high quality” data for the purposes of curation.

Categories and Subject Descriptors

H.1.1 Systems and Information Theory: Value of Information

General Terms

Measurement, Documentation, Performance, Economics, Reliability, Security, Human Factors, Standardization, Verification.

Keywords

Designated communities, data quality.

1. INTRODUCTION

It is taken as a given by many that it is not possible for repositories to curate all of the data generated today and into the future[1]. If that assumption is true, the implication is that data curators must carefully weigh the pros and cons of accepting any particular data submission and if accepting the submission to equally carefully consider the level of support or services to be provided these data. In other words, data must be of high quality in order to be worthy of initial curation and must continue to be high quality to be worthy of ongoing maintenance. The question then becomes what constitutes high quality data and how can that state, the state of being “high quality,” be maintained over time.

2. DEFINING HIGH QUALITY DATA

2.1 Measurement Quality

When the science community talks about data quality, they are typically talking about indicators such as measurement error (the sensor readings are good to +/- 1 degree), presence of contaminating factors (it was cloudy so the ground could not be observed), accuracy of auxiliary information (e.g., pointing accuracy, temporal fidelity) or other such objective measurements. This generally works reasonably well to characterize the accuracy (or not) of the raw data, but the situation becomes much more complicated when describing the quality of derived products. For example, it typically is not obvious how to get from "the instrument accuracy is 1%" to the accuracy of the derived snow cover is 93%, since a wide variety of other non-instrument related factors get in the way (i.e., reality gets in the way). Factors such as whether a instrument, such as the Moderate Resolution Imaging Spectroradiometer (MODIS) observing the ground in the visible wavelengths, was over the boreal forest where the ground may be 100% snow covered, but the evergreen canopy is typically at least partly snow free such that the snow cover is under observed. That is where the science community, in this example the remote sensing community, comes in with the notion of "ground truth", "validation campaigns", "calibration methods", and other such external factors that allow one to take data with some purported statistical accuracy based off the characteristics of the instrument itself and perform some other comparison with "reality" to come up with statements like "the overall absolute accuracy of … the snow cover is ~93%, but varies by land cover type and snow condition"[1].

The question is whether these kinds of characteristics, termed measurement quality for the purposes of this paper, are necessary and sufficient for data to be deemed “high quality” for curation purposes. The answer is clearly not. Science is rife with examples where “one man’s noise is another man’s signal” [2]. Two examples spring immediately to mind. The first involved ozone data gathered by spacecraft during the 1970’s which was ignored by the science community due to the extremely low values observed, values that at the time were considered to be erroneous; but which later turned out to reflect actual thinning of the ozone layer. The second involves the use of multi-path noise in GPS receivers, noise which leads to positioning errors in the GPS data, but which can be used to measure the soil moisture which is important for water cycle studies [3].

If measurement quality is not necessary and sufficient for data to be high quality for curation purposes, the next question is whether it is necessary at all. Here again the answer is clearly not, at least not in absolute terms only in relative terms. Whether a measurement is characterized to 1% or 1/10000th % is not nearly as important as how it compares to the state of the art for measurements of that type (e.g., this is the first measurement of its type) and whether or not the measurement is carried out in a regime rarely achieved (e.g., in a region or temporal period otherwise unobserved). What is important to the science community as well as data repositories is that the measurement quality is both known and documented. Documentation of measurement quality is one aspect of what Weaver et al terms science and preservation maturity [4] and what Palmer et al terms preservation readiness [5].

[1] Note that the validity of that assumption will not be addressed here.

2.2 Quality in the Data Center

While perhaps not generally accepted throughout all domains, the Reference Model for an Open Archival Information System (OAIS) [6] is broadly accepted within Earth and Space science data centers as the recommended best practice for archives that are endeavoring to preserve information expressed as data for the long tem. The OAIS reference model, also known as ISO Standard - ISO 14721:2012, was developed under the auspices of the Consultative Committee for Space Data Systems a long standing, multi-national forum for the development of communication and data system standards, and is updated every 5 years based on community input. The OAIS reference model notes that to preserve information for the long-term it is necessary to not just preserve the data itself but also to preserve enough Representation Information so that the data is understandable, as well as enough additional Preservation Description Information - information about the Provenance, Context, Reference, Fixity, and Access Rights - to identify, preserve and understand the environment of the original data and its Representation Information.

Core to the model is the concept of a designated community, the community that should understand the Representation Information without requiring assistance. Where the designated community is narrow, perhaps focused on a particular sub-discipline, Representation Information is expected to be highly specialized, full of technical jargon with substantial community-specific tacit knowledge assumed, not explicitly spelled out. As the community broadens for whatever reason (e.g., the passage of time or changes in community interest) the character of the Representation Information must change to accommodate the less specialized knowledge base of the new designated community. Changes that broaden the community are generally assumed to require additional information; the information needed to document the tacit knowledge that can no longer be assumed and to replace the technical jargon with general terminology and explanations suited to the broader audience.

While quite explicit in the general types of information that must be maintained in order for data to be preserved, the OAIS Reference Model is just that, a reference model, and does not provide detailed specification of the content that must be gathered and maintained to create an entire Information Package (i.e., the data that is the target of preservation along with its Representation and Preservation Description Information). In recognition of this, some domains have begun to fill this gap by explicitly defining the content of a complete Information Package for their domain. For example, in the Earth sciences the Stewardship Committee of the Federation of Earth Science Information Partners (ESIP) has begun to work on a Preservation and Context Content Standard (PCCS) for Earth Science data [7]. The draft specification, which included input from primarily NOAA and NASA ESIP members, has already been adopted by NASA, which re-formulated the content for use by its missions [8]. Simultaneously, the European Space Agency has developed Long Term Data Preservation Common Guidelines for use in their archives [9]. A comparison of the NASA and ESA guidelines has revealed that there is significant overlap and often one-to-one correspondence in the content, a circumstance that bodes well, given that the long-term intent of all three communities, NASA, ESA and the ESIP Federation is to evolve these efforts to produce an IEEE or an ISO standard on preservation content [10].

While this author is not intimately familiar with the genesis of the ESA guidelines, the ESIP and NASA guidelines are derived from the results of a 1998 workshop sponsored by the US Global Change Research Program (USGCRP) to discuss the Global Change Science Requirements for Long-Term Archiving [11]. The workshop considered several use cases where long-term analyses were helped or hindered by the preservation of metadata and documentation beyond the original data. The workshop concluded by publishing a short list of the documentation that is required if a data set is to meaningfully contribute to long-term change studies. This list is included below, as it may be useful to consider what the content of equivalent lists for other types of study should contain:

1. "Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g. spectral response, noise characteristics, etc.);
2. Instrument/sensor calibration data and method;
3. Processing algorithms and their scientific basis, including complete description of any sampling or mapping algorithm used in the creation of the product (e.g., contained in peer reviewed papers, in some cases supplemented by thematic information introducing the data set or product to scientists unfamiliar with it);
4. Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product;
5. Processing history including versions of processing source code corresponding to version of the data set or derived product held in the archive;
6. Quality assessment information;
7. Validation record, including identification of validation data sets;
8. Data structure and format, with definition of all parameters and fields;
9. In the case of earth-based data, station location and any changes in location, instrumentation, controlling agency, surrounding land use and other factors which could influence the long-term record;
10. A bibliography of pertinent Technical Notes and articles, including refereed publications reporting on research using the data set;
11. Information received back from users of the data set or product [11]."

Is having the equivalent of all of this information documented and available what makes data "high quality" for curation purposes? Well it certainly is part of the answer; but three things should be noted. First, the information discussed in this section includes documentation of the measurement quality as defined in Section 2.1 above. And again, these components correspond well and more completely with Weaver's concept of scientific and preservation maturity [4] and Palmer's preservation readiness [5]. Second, these requirements can be quite costly to implement. Even NASA, an organization that spends millions to ensure that its data is well managed and accessible, does not require that all of their missions provide all of this information as they recognize that they don't truly have the resources to do so. In addition, it is not obvious that every data set needs or deserves this level of support. At the very least a cost vs. benefit appraisal would be needed, as is mentioned in both the NOAA and USGS procedures for accepting data [12], [13]. Third, central to these definitions is the concept of the "designated community," the idea that the potential community of users for the data can be large or small, uniform or diverse, and that the amount of detail and contextual information needed is actually a function of the organization of that community. In other words, who your users are, what and how diverse their needs are, and what uses they are going to make of the data are key and part of the definition of what makes data "high quality" for curation purposes.

2.3 Quality and Designated Community

According to the previous sections, there is no way of answering the question "what is the quality of that data" without also answering the question "to what purposes can the data be put by what user communities" (i.e., what is its designated community). By definition that makes the question of what is "quality data" a sociotechnical issue as users and their intentions are central.

But if fitness for purpose must be assessed in any discussion about data quality, then a wide variety of considerations become important and the range of potential considerations for any particular data set, type of data set and potential use may be large. One example of such a consideration is timeliness. If environmental data is not available within hours or in some cases minutes of the measurement the data are useless for weather forecast purposes. As another example, Parsons and Duerr [14] discussed the use case of a biologist tracking polar bear migration coming to the National Snow and Ice Data Center's (NSIDC's) website in search of sea ice data. NSIDC at that time, had 38 sea ice concentration data sets derived from passive microwave remote sensing data not one of which was relevant to this user who really needed data with high spatial and temporal resolution over the locations of their study. In other words, the considerations were spatial location and geospatial and temporal resolution. NSIDC's primary data set, the consistently processed nearly 40-year record of sea ice extent so useful for long-term trend assessments was exactly the wrong data set for this user's purpose. One particularly worrisome consideration is how the continual growth in knowledge and understanding over time by the science community and perhaps humanity in general will affect the types of data that will be useful in the future. There are plenty of examples in the literature, including the examples mentioned earlier in this paper, where data initially deemed valueless turned out to be extremely important.

At the other end of the spectrum the question "is there any data with quality so bad that there isn't some use for it?" can be asked. Upon reflection the answer must be no. If nothing else, examples of bad practice are very useful for training purposes. The question repositories must ask themselves becomes what is the range of potential uses for these data and are those uses so limited and valueless that there is no point wasting resources on the data. Without unlimited resources the answer can only be a qualified yes, with the caveat that the state of knowledge in the future may make any decision to eliminate the data today erroneous. This then becomes an issue of balancing risk vs. reward.

3. DISCUSSION AND CONCLUSIONS

The previous sections have discussed what makes data "high quality" for curation purposes. The answer seems to be data that has a potential designated community that is sizable enough to justify the expense of providing the level of Representation and Preservation Descriptive information needed by that community is "high quality". In some sense, this is equivalent to saying that the data is suitable for supporting the range of uses needed by its user community, both now and in the future. If these statements are warranted, then a number of research questions come to mind. They include:

1. How does the needed Representation and Preservation Description Information vary from discipline to discipline and as a function of the designated community?
2. How can repositories assess a priori the potential user community and range of uses to which a data set may be put?
3. Is it possible to group use cases by type such that the relevant considerations are uniform across all use cases of a given type?
4. What are the considerations relevant to a particular use case or use case type?
5. How does the increasing state of knowledge of the scientific community and humanity as a whole affect the range of uses and potential community for a data set as a function of time? Can that be predicted with any accuracy so that responsible decisions can be made now that won't be regretted in the future?
6. How can discipline-specific repositories, assess the potential for their data holdings to meet the needs of communities outside their core discipline expertise?

4. ACKNOWLEDGMENTS

My thanks go to the Data Conservancy and the members of the ESIP Federation for the many discussions that informed my thoughts on this topic.

5. REFERENCES

[1] Hall, D. K. and Riggs, G. A. (2007), *Accuracy assessment of the MODIS snow products*. Hydrol. Process. 21: 1534–1547. DOI= http://dx.doi.org/10.1002/hyp.6715.

[2] Ng, E. 1990, in the article "Lost on Earth: Wealth of Data Found in Space" by Sandra Blakeslee, The New York Times, March 20, 1990.

[3] Larson, R. M., E. E. Small, E. D. Gutmann, A. L. Bilich, J. J. Braun, and V. U. Zavorotny (2008), Use of GPS receivers as a soil moisture network for water cycle studies, *Geophys. Res. Lett.,* 35 L24405, DOI= http://dx.doi.org/10.1029/2008GL036013.

[4] Weaver, R. L. S., Meier, W. M., and R. E. Duerr, 2008. *Maintaining data records: Practical decisions required for data set prioritization, preservation, and access*. Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium. Volume 3, 617-619. DOI= http://dx.doi.org/10.1109/IGARSS.2008.4779423.

[5] Palmer, C. L., Weber, N. M., and M. H. Cragin. 2011. *The Analytic Potential of Scientific Data: Understanding Re-use Value"*. Proceedings of the American Society for Information Science and Technology 48: 1-10 DOI= http://dx.doi.org/10.1002/meet.2011.14504801174.

[6] CCSDS, 2012, Reference Model for an Open Archival Information System (OAIS), Recommended Practice, Issue 2, CCSDS 650.0-M-2. CCSDS Secretariat. Washington, DC. http://public.ccsds.org/publications/archive/650x0m2.pdf

[7] ESIP Federation, 2011, "Provenance Context Content – 2011-06-08", http://wiki.esipfed.org/index.php/File:Provenance_Context_Content_2011-06-08.xls

[8] NASA, 2011, “NASA ES Data Preservation Content Spec (423-SPEC-001, Nov 2011)”, http://earthdata.nasa.gov/sites/default/files/field/document/NASA_ESD_Preservation_Spec.pdf

[9] Albani, M., R. Guarino, and R. Leone, Long Term Data Preservation Preserved Data Set Composition, Issue 4, Revision 0, May 4, 2011, ESA, http://earth.esa.int/gscb/ltdp/LTDP_PDSC_3.pdf.

[10] Ramapriyan, H. K., Moses, J., and R. Duerr. (in press). *Preservation of Data for Earth System Science – Towards a Content Standard,* Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium.

[11] USGCRP, 1999, Global Change Science Requirements for Long-Term Archiving, Report of the Workshop, October 28-30, 1998, National Center for Atmospheric Research, Boulder, CO. Sponsored by NASA and NOAA, through the USGCRP Program Office. http://wiki.esipfed.org/images/4/40/USGCRP_Long-Term_Archiving.pdf.

[12] NOAA, 2008, NOAA Procedure for Scientific Records Appraisal and Archive Approval: Guide for Data Managers. https://www.nosc.noaa.gov/EDMC/documents/NOAA_Procedure_document_final_12-16-1.pdf

[13] USGS, ???, USGS EROS Operational Procedure: Acceptance of Data Collections by the USGS. EROS-DMD-01.

Data Quality at Web Scale: Examining Context and Privacy

Andrew T. Fiore
Facebook, Inc.
1601 Willow Rd.
Menlo Park, Calif.
atfiore@fb.com

ABSTRACT

Large-scale data provides a powerful tool to scientists and organizations seeking to understand complex processes. However, issues of data quality are often overlooked. Thorough checks of correctness may be prohibitively costly in terms of time and resources. Similarly, supporting the context of analysis matters even more with an unwieldy volume of data. An aggregation or metric suitable for one purpose may be unsuitable, or effectively low quality, for another, yet transforming among data representations or units of analysis may be expensive and storing numerous versions or copies unfeasible. Furthermore, as the importance of online social-behavioral data grows, privacy must be considered another aspect of data quality, in terms of both safeguarding the privacy of people represented in the data and understanding how those safeguards affect the utility of the data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining, scientific databases, statistical databases.*

General Terms

Management, Measurement, Documentation, Standardization, Legal Aspects

Keywords

Data warehousing, big data, computational social science, privacy

1. INTRODUCTION

In the era of big data, analytic methods are parallelized for distributed computing; patterns are uncovered; truth and perhaps causation are sought amongst tiny *p*-values (cf. Anderson 2007). The volume of the data is taken almost as proof of its quality, as though the sheer amount of detail renders conclusions drawn from it unimpeachable.

Although the more fantastical claims about big data have been effectively debunked (e.g., Timmer 2008, Rieder 2008), scientists and engineers developing new inferential techniques to operate at scale, they may still overlook fundamental concerns. Are the data complete and correct? Are they unbiased? Are they aggregated and stored appropriately for the question at hand? Who has made decisions about what should be represented in what way? For what purpose?

As the volume of data in the sciences and in organizations grows, the quality of the data becomes both increasingly important and increasingly hard to verify. In the following sections, I will argue for (a) analytic context and (b) privacy protections and their impact on analytic flexibility as two of the most relevant dimensions of quality, particularly as computer-mediated social-behavioral data sets grow in number, volume, and importance.

2. CONTEXTUAL DATA QUALITY

What are the characteristics of high-quality as compared to low-quality data? The answer may seem straightforward at first, but not all aspects of data quality are universal. Some depend on the task at hand.

Corrupted data is almost certainly undesirable in any circumstance, a clear sign of low quality. Missing data would seem to be as well. But missing data may be bad to varying degrees depending on the pattern of absence and the task at hand. For example, randomly missing rows of data may be tantamount to a random sample, which is unproblematic for a variety of inferential tasks. (Of course, in this case it is vital to know that observations are randomly and not systematically missing, which may not always be apparent.) On the other hand, for network data, random samples are not useful, and randomly missing observations constitute a threat to quality.

These scenarios are perhaps far-fetched — if there were some failure in a system or process, having missing data with known and desirable properties is unlikely. Yet one can imagine intentionally discarding some data if space constraints made it necessary to limit the amount of data retained. Depending on the purpose of the data, one might wish to retain partitioned chunks of data in some cases (e.g., network analysis) and random subsets in others. If longitudinal analysis is the goal, retaining all of the most recent data is strictly inferior to retaining a sample of data over time.

Context and purpose matter in any discussion of data quality; sometimes what is high quality for one purpose is deficient for another. Wang and Strong (1996) provide a descriptive framework from the point of view of data consumers, derived from an inductive-deductive multi-stage survey, that captures the distinction between what they call the "Intrinsic" and "Contextual" data quality dimensions.

Intrinsic quality comprises accuracy in an objective sense — that is, whether the data constitute unbiased, uncorrupted measurements — but also, in Wang and Strong's framework, the more subjective credibility and reputation of the data product or process. Contextual data quality, on the other hand, refers to the suitability of data for a particular purpose in terms of its relevance, timeliness, volume, and completeness (Wang and Strong 1996).

The framework comprises two other dimensions as well. "Representational" quality is the degree to which data are consistently and concisely represented as well as readily interpretable. One might argue that interpretability is also a contextual quality, as data that are easily interpretable for some purposes may be confusing for another. The final type of quality, "accessibility," includes both ease of access and appropriate access controls or security (Wang and Strong 1996; Wang 1998).

2.1 Quality of Web Data

When it comes to web data, low-level records such as individual web server requests may be high quality from the perspective of the systems engineer — they provide information about the behavior of a server process — but they are low quality from the perspective of the scientist studying the human behavior represented by the web requests. The original semantics of HTTP had some meaning (e.g., GET vs. POST requests) but those are subsumed today by other technical considerations.

On modern web sites, asynchronous requests from client-side programs load small pieces of data rather than requesting whole pages or monolithic state changes that correspond to single records in web server logs. Web requests were never particularly rich records of behavior, but today the semantics has diverged even more from the high-level intent of the user. As a result, web request data are not high quality for the researcher whose unit of interest is the user, not the technical infrastructure. For these data consumers, aggregation is an important contextual aspect of quality.

To analyze individual differences, data aggregated at the level of users makes sense. To analyze change trajectories, aggregation at the level of salient high-level behaviors, such as sending a message, within users may prove most useful. The number of possible aggregations is large, and for sufficiently voluminous data sets, it is not possible to pre-compute and store all of them. One could make the same argument for other types of operations on the data, such as numerical transformations or sampling. Thus, in a practical sense, data quality is inextricably intertwined with the purpose that the data consumer has in mind.

3. ASSESSING DATA QUALITY

In terms of the four dimensions identified by Wang and Strong, only intrinsic quality need be explicitly assessed. Assuming the overall schema of the data is known and correct, contextual, representational, and accessibility quality can likely be inferred without examining the data themselves. But intrinsic quality checks demand a more thorough approach.

3.1 Computational challenges

With small data sets — for the sake of discussion, small enough to fit on the hard drive of a typical PC — passing over the entirety to verify intrinsic data quality or to transform the data into a different form is a tractable task. When data volume demands distributed systems, sufficient storage to maintain more than one copy (as with an original and a transformed version) may be impractical, and the cost in terms of time and resources to make repeated passes over the data may be prohibitive.

In such cases, samples or spot checks may be sufficient in some circumstances for evaluating intrinsic data quality. Summary statistics or roll-ups that deviate systematically from known historical rates or patterns can serve as red flags that prompt further investigation, as such metrics can of course move for legitimate reasons as well.

Whether sampling approaches suffice depends, again, on the application domain. High-reliability contexts may demand validation of every observation simply for the sake of certainty. Even in other contexts, some types of analysis and estimation tasks could suffer if all the data are not included, e.g., analyses of rare events or techniques susceptible to the influence of outliers. Intuitively, sampling under these circumstances introduces more randomness into the analysis than would sampling when the quantities of interest are normally distributed.

Checking intrinsic quality has benefits beyond the confidence that data are not corrupt. Removing observations with missing or invalid values reduces overall volume and key skew, which can improve the speed of lookups and joins. [Data preparation cite]

3.2 Organizational challenges

Another obstacle to assessing data quality is the need to join data sets across potentially heterogeneous systems in order to provide a full assessment. Large data stores may be compartmentalized by source or purpose, yet combining them may be necessary for effective validation. Maintaining current and consistent metadata across disparate systems, often controlled by different parts of an organization, is itself a challenge. Enforcing the creation and updating of documentation as an organizational policy or norm is burdensome and likely to lag behind any changes to the data store.

Yet data consumers need up-to-date schema documentation to make the best use of data. The documentation is itself an aspect of data quality. Automated systems for identifying data distributions, potential foreign keys, and join paths can reduce this burden somewhat (Dasu et al. 2002). In a dynamic data environment, with the types, volumes, or locations of data changing rapidly, an automated approach offers benefits. In particular, even if data curators do not document a new table, a system like that proposed by Dasu and colleagues could infer data types and probability distributions, information which could facilitate detection of matching columns in other data stores against which the new data might usefully be joined. This matching may also allow metadata imputation, depending on the certainty of the match.

4. PRIVACY DATA QUALITY

For social-behavioral data, an additional dimension of interest not included in Wang and Strong's (1996) general-purpose model is privacy. We can consider "privacy quality" to be similar to Wang and Strong's "accessibility quality" in that both dimensions encompass two sides of the same coin. Accessibility quality refers not only to how readily accessible the data are but also to how well secured they are against unauthorized access. Similarly, we can take privacy quality to encompass both the degree of privacy protection afforded the people represented in the data and also the extent to which privacy protections make the data harder to use or less useful.

Clearly, excluding or obfuscating explicit personally identifiable information is important to privacy quality. Yet as numerous instances from organizations like AOL and Netflix have shown, simply removing obvious PII is insufficient, as other day may facilitate reidentification, especially in conjunction with external data sets that contain partially overlapping information. (In the case of the Netflix data that was partially reidentified, merging the data with IMDB was a key step.) Guarding against reidentifiability entails a trade-off between protecting privacy and preserving the usefulness of the data. If all potentially unique combinations of characteristics are excluded, the data cannot be linked back to individuals, but this requirement so restricts the utility of the data for many reasonable classes of question as to make it practically useless. In other words, it seems inevitable that we exchange aspects of contextual quality (e.g., relevance and completeness) for the protective aspect of privacy quality.

4.1 Heuristic and provable approaches

The U.S. Census Bureau takes a more nuanced approach than simple redaction. Instead of excluding data that could lead to reidentification, it systematically perturbs such data, sometimes swapping nearby values of a variable within the data set or even generating synthetic values based on an observed distribution (e.g., Muralidhar and Sarathy 2006, Nissim 2008). These perturbations are designed to preserve summary statistics and the ability to conduct common analyses without risk of bias. Extremely high and low values can also be grouped to prevent unique identification of outliers (U.S. Census Bureau 2006). These techniques can greatly reduce the risk of identity disclosure and perhaps eliminate it if used aggressively with limited combinations of types of data. However, this combination of techniques appears to be evaluated heuristically, not formally.

A formal approach is offered by "differential privacy." Dwork (2006) provides an overview of this approach while pointing out that a general guarantee of non-identifiability is impossible in the presence of "auxiliary information," or external data sets. However, differential privacy does allow us to characterize the risk to privacy as a function of both the data and the questions asked of it. A malicious actor could ask a carefully crafted series of questions to elicit combinations of information that together pose a threat to privacy; as a result, a database that gives perfectly correct responses to every query poses a threat to privacy quality (Dwork 2007). The administrator can determine the maximum acceptable probability of such a disclosure and prohibit further queries as this limit approaches.

Interestingly, the validity of differential privacy estimates of risk requires that the cumulative exposure of data via queries be construed globally. That is, any attacked could work in conjunction with other users of the database, so the differential privacy approach must assume that any or all potential data consumers could be collaborating, pooling the results of their queries, to potentially uniquely identify one or more people represented in the data. This implies that the ability to query a given database will eventually be "used up"; with use over time, it will reach the risk threshold set by the administrator in terms of the volume and nature of data in the hands of data consumers. At this point, the administrator must either accept the heightened risk or disable access to the data.

Certain types of data pose additional challenges. Network connection data, even if sampled in clusters and anonymized in some of the ways described above, may be reidentifiable due to the graph structure alone. Perturbing the graph structure to preclude this may alter structural metrics in ways that introduce bias or noise into analyses. Effective heuristics to accomplish this, as Census researchers apply to numeric variables, do not yet exist.

5. CONCLUSION

Data quality at large scale presents unique challenges. Direct assessment of intrinsic quality is costly, particularly when sampling is not a viable option. The context of analysis is important to understand, as different selections, transformations, and aggregations of data are needed for different purposes. When data volumes are large, generating such contextualized metrics on the fly is often not viable, leading to low-quality data for specific purposes. Finally, for data about humans, privacy should be considered both part of data quality and a source of tension with contextual utility. Developing techniques to ease that tension so that high-quality data can be generated without compromising privacy is an important direction for future research.

6. REFERENCES

[1] Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine* 16.07.

[2] Dasu, T., Johnson, T., Muthukrishnan, S., and Shkapenyuk, V. (2002). Mining Database Structure; Or, How to Build a Data Quality Browser. In Proc. *ACM SIGMOD 2002*.

[3] Dwork, C. Differential Privacy. (2006). In (Bugliesi, M., et al., eds.) *Lecture Notes in Computer Science: Automata, Languages and Programming*. Berlin: Springer.

[4] Dwork, C. Ask a Better Question, Get a Better Answer: A New Approach to Private Data Analysis. (2007). In (Schwentick, T., and Suciu, D., eds.) *Lecture Notes in Computer Science 4353: Database Theory*, *ICDT 2007.*

[5] Muralidhar, K., and Sarathy, R. (2006). Data Shuffling—A New Masking Approach for Numerical Data. *Management Science* 52 (5): pp. 658–670.

[6] Nissim, K. (2008). Private Data Analysis via Output Perturbation: A Rigorous Approach to Constructing Sanitizers and Privacy Preserving Algorithms. In (Aggarwal, C.C., and Yu, P.S., eds.) *Privacy-Preserving Data Mining*. New York: Springer.

[7] Rieder, B. (2008). Statistics vs. science (and why this is rather political). In *The Politics of Systems*. Available: http://thepoliticsofsystems.net/2008/06/statistics-vs-science-and-why-this-is-rather-political/

[8] Timmer, J. (2008). Why the cloud cannot obscure the scientific method. In *Ars Technica*. Available: http://arstechnica.com/uncategorized/2008/06/why-the-cloud-cannot-obscure-the-scientific-method/

[9] U.S. Census Bureau. (2006). Advanced Query Technical Design: DADS Project. Version 3.8, December 7, 2006. Available: http://www.census.gov/procur/www/dads2/j-11%20att%20d.aq%20technical%20design.pdf

[10] Wang, R.Y. (1998). A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41 (2): pp. 58–65.

[11] Wang, R.Y., and Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12 (4): pp. 5–33.

Scientific Data Quality: Openness, Provenance, and Replication

Michael Lesk
Rutgers University
New Brunswick, NJ 08901
+1-732-932-7500 x8230
lesk@acm.org

ABSTRACT

An archive of scientific data has to start with experimental observations, and those must be reliable for the archive to have quality content. Recently an attempt to replicate more than 50 key studies in cancer research found that only a few could be reproduced. Especially in a world of automated conclusions based on databases, scientific progress depends on reliable data. To make repositories useful, we will need public availability of research data, tracking of data provenance, and rewards for data collection and replication.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific databases

General Terms

Measurement

Keywords

Data quality, open data, replication, provenance.

1. INTRODUCTION

In a frightening report in *Nature,* Begley and Ellis [1] recently detailed attempts to replicate 53 important oncology research experiments, of which the results of only 6 could be reproduced. The implication is that vast amounts of effort is being wasted in attempts to develop cancer treatments based on incorrect papers, effort that might have been used to explore other potential treatments. Prinz [2] and Booth [3] have raised similar doubts about many other medical studies. A few years ago Ioannidis [4] gave an explanation of why we might expect most papers to be wrong, based on the way we do experimental design and the number of researchers. Begley notes that many of the papers he could not replicate have hundreds of citations and entire research areas have been built around them. Large chunks of recent medical research, it seems, are no better than fan fiction.

In addition to financial costs there are human costs. The incorrect research suggesting that childhood vaccines cause autism [5] has resulted in a resurgence of whooping cough, particularly in the wealthiest and usually healthiest areas. In 2010 Marin County, California, had 350 cases of whooping cough, attributed to the fears of vaccination created among parents by bad medical research [6]. California as a whole had 9,000 cases, the largest number since the 1940s, and including ten children who died [7].

Problems with replication are not unique to biomedicine. McCullough [8] reported that most of the research in the *American Economic Review* did not contain enough public data even to attempt replication, and Evanschitsky and Armstrong [9] cast doubt on research papers in forecasting. Bartlett [10] discusses experimental psychology and suggests that psychological researchers should be worried that replication attempts will show their field to be full of doubtful results.

The traditional scientific view is that quality and reliability depend on reproducibility. However, the current publications process discourages doing the same experiments over again. Yong [11] discusses a case where three different investigators failed to replicate a study claiming that "pre-cognition" exists but could not find journals willing to publish their result.

So if we want high quality scientific data archives, what should we do? We need better measurements, and measurements we can believe. Steps to be taken include (a) continuing to encourage open data, (b) better data curation, including tracking provenance, and (c) encouraging replication of more studies and measurement of reliability in science.

2. OPEN DATA

The more obscure a paper, the more difficult it is for anybody to replicate it. There is an increasing movement to make the data underlying every paper generally available. This permits meta-analysis, and for others to build upon the work. Open access to publications is growing: the US National Institutes of Health have required this from their grantees since 2009, and the UK government has joined in this demand [12]. We are now also looking for the actual numerical observations from research to become public. As an example of the value of access to original data, we might find drug side effects more rapidly if the full results from the original clinical trials were available. The US National Science Foundation now requires that data from experiments be made available. NIH also has such a requirement although the systematic enforcement of it lags. We also need agree metadata and format standards; their absence hampers automated exploitation of the data. Some areas do have them, such as radiology or seismology.

If the details of research are not publicly available, it is often impossible to replicate it and validate it. Victoria Stodden has spoken eloquently of the need to see research results in order to check them [13]. In addition to papers and data, she observes that in computational areas, we need to see code in order to know

Conference'10, Month 1–2, 2010, City, State, Country.

whether the project has been done correctly. The benefits from replication are societal benefits, but there are also benefits to the original authors from open data. Piwowar [14] notes that papers with open data are cited more than papers without data, and in a world where academic promotion relies on bean-counting such as h-index and impact factor, citations are extremely valuable.

On the other side, there are both individual and societal costs to open data. The authors who have collected the data fear that providing open data will be a substantial amount of work. Perhaps more seriously, the original experimenters fear that other researchers who can see their data will have insights the original authors might have had with private use of their data for a few more months. In terms of society, there are already arguments of the form "we are providing free data that country X can see, but country X does not do the same."

In some cases the choice between open and closed data seems purely historical. Why do astrophysicists provide their data while high energy physicists do not? As best I can tell, this traces to decisions made decades ago. Certainly fields that are open, such as protein chemistry, seem to make progress at a rapid rate and are not planning to change. The idea of "open data" is spreading; fairly recently, for example, China began providing seismographic data to the international consortium that collects such information.

The most serious problem with open data may well be academic credit. When the same person who gathered data also interpreted it (Galileo dropped his own objects off the Leaning Tower of Pisa, he didn't send a PhD student), there was no problem of allocating credit. Today, however, it is common to give more credit to the person who interprets the data than to the person who collected it. As a result, data gatherers fear making data public. This is something that could be adjusted by the community with journals for data collection, tenure decisions recognizing the importance of providing the data used for insights, and other reward choices such as prizes.

3. PROVENANCE

Many databases are derivative. Protein structures are curated by a variety of organizations such as the Protein Data Bank, UniProt, TrEMBL, and so on. There are sequence databases and 3-D structure databases. Often raw data are annotated or labeled as they move from one data archive to another. Data are also re-arranged, since there are archives that focus on some context for proteins, e.g. the database of HIV-1 proteins. And there are databases that combine protein and genomic data, adding as well clinical and other data. Computational biologists are always combining and comparing multiple databases. An amusing paper compares their techniques to the word games of Lewis Carroll [15]; a more serious paper by the same author describes many ways in which databases are combined and exploited [16].

When one database is used to create another, we need to know when a change in the first database implies a change to be passed forward to the second database (or when a change in the second is actually a correction to be propagated back to the first). Peter Buneman [17] has discussed the problems of tracking provenance between databases. Realistically, if we don't know where data comes from, we can't reliably assess it or depend on it. When we find a mistake in some data element, we need to know what other data elements depended on it (or have been contradicted).

Tracking data provenance is not quite the same job as either running experiments or interpreting the data. It is a part of data curation, and it's probably not a task for an individual researcher, since it needs to be done by an organization of some permanence and with the specialized skills needed for data handling. This raises a larger problem addressed elsewhere of how such a profession of data curators will be educated and supported [18]. We must not view data management as something inferior to data interpretation and modeling.

4. REPLICATION

Perhaps the most serious issue we face is the need to replicate more experiments. As long as it is possible to have a successful academic career based on data fabrication [19] we will run the risk of fraud. And as long as journals do not wish to publish replications, we'll see academics avoiding doing such work. Attempting to replicate also involves personal risk, of course; suggesting that somebody else's work is wrong may well provoke antagonism, even if the mistakes are not deliberate. Most mistakes are not fraud: they may just represent random chance. If 20 people evaluate something, one of them may well get a result which is significant at the 0.05 level and also be ignorant of the other experimenters. And, of course, experimental methods improve over time, and earlier results may be corrected by work done with better equipment or improved techniques.

However, the advice to researchers continues to be that checking previous work is not all that important. Price writes [20] that replication "is usually a waste of time and resources that would be better spent on original research that will further your career." So long as this attitude persists, we will not see much checking.

One could argue that checking unimportant results is indeed not worth much effort. From the standpoint of scientific progress, papers that are unread have no effect, either positive or negative. Note that somebody planning deliberate fraud is well advised to fake results that will not attract enough attention to be checked. Although this research might seem harmless, there is a huge opportunity cost as research funding and research positions are frittered away on insignificant results. Begley & Ellis [1] were too polite to say so, but the company that attempted the replications, Amgen, has a yearly research budget of $3B, and if 90% of any significant part of that is being wasted following up mistakes, it is a huge waste both economically and of investigator time.

How much replication is needed? Many European transport authorities use a "proof-of-payment" policy: passengers buy tickets ahead, and do not present them on boarding, but spot checks are made on the vehicles. The bus or train operator needs to check enough tickets to feel that the typical passenger will perceive that it is not worth trying to cheat. Part of this is a high penalty for those caught without tickets, and part is a sufficiently frequent inspection. A typical inspection rate is perhaps 0.5-2% of passengers [21]. Clearly, the cost of replicating 1 percent of scientific studies is a tolerable cost in the context of overall research funding, especially if focused on the most significant studies. The penalty, certainly for deliberate cheating, is already extremely severe in academia. The problem is to persuade somebody to do the replications, and the funding agencies ought to be able to manage that. Or, in the same way that tenure committees expect service on administrative committees to be part of any evaluation, they could expect replication of at least a few important experiments.

What is more complex is to reach a balance in which we routinely measure what fraction of papers are actually reliable, and adjust the funding of replications in order to achieve a desired level of reliability. We would like to do this without getting into a high level of contentiousness; it will serve no scientific purpose to be seeing more libel suits over research evaluations [22,23,24].

5. CONCLUSIONS

To build good repositories of scientific data, we need to balance the cost of data curation against the cost of bad content. We expect to see increasing data mining and automated knowledge creation, for which we need reliable data. If your medical treatment is going to be suggested by algorithms that search for similar cases and successful treatment, we want that medical research data to be dependably accurate. We need to measure scientific data quality by experimental replication, track the dependencies among data files, and provide as much openness as practical. We should use sampling methods to estimate data quality and then manage research to improve it. Most important, academic reward systems, as managed by university departments and granting agencies, should tilt towards encouragement of high quality data gathering, data curation, and data validation, and away from claims made without adequate support or checking.

6. REFERENCES

[1] Begley, C. Glenn and Ellis, Lee M. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483, 531–533 (29 March 2012

[2] Prinz, Florian, Schlange, Thomas & Asadullah, Khusru. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Rev. Drug Discov*. 10, 712.

[3] Booth, B. 2011. Academic bias and biotech failures. See http://lifescivc.com/2011/03/academic-bias-biotech-failures/#0_undefined,0_.

[4] Ioannidis, J. P. A. 2005. Why Most Published Research Findings Are False. *PLoS Med* 2, e124. http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124.

[5] Harris, Gardiner. 2010. Journal Retracts 1998 Paper Linking Autism to Vaccines. *The New York Times,* Feb. 3, 2010, page A9. http://www.nytimes.com/2010/02/03/health/research/03lancet.html.

[6] Lupkin, Sidney. 2011. Marin County's Efforts Against Whooping Cough Pay Off. *The New York Times*, page A27A, September 18, 2011. http://www.nytimes.com/2011/09/18/us/marin-countys-efforts-against-whooping-cough-pay-off.html.

[7] California Department of Public Health. 2011. Pertussis (Whooping Cough). http://www.cdph.ca.gov/HealthInfo/discond/Pages/Pertussis.aspx.

[8] McCullough, B.D. 2007. Got Replicability? *The Journal of Money, Banking and Credit*. In *Econ Journal Watch* 4(3): 326-337.

[9] Evanschitzky, Heiner, and Armstrong, J. Scott. 2010. Replications of forecasting research. *International Journal of Forecasting*, Volume 26, Issue 1, January–March 2010, Pages 4-8.

[10] Bartlett, T. 2012. Is Psychology About to Come Undone? *Chronicle of Higher Education*, April 17, http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045.

[11] Yong, Ed. 2012. Replication studies: bad copy. *Nature* 485, 298–300 (17 May 2012).

[12] Sample, Ian. 2012. Free access to British scientific research within two years. *The Guardian*, Sunday 15 July. http://www.guardian.co.uk/science/2012/jul/15/free-access-british-scientific-research.

[13] Stodden, Victoria, and Arbesman, Samuel. 2012. Scientists, Share Secrets or Lose Funding. *Bloomberg View*, January 9, 2012. http://www.bloomberg.com/news/2012-01-10/scientists-share-secrets-or-lose-funding-stodden-and-arbesman.html.

[14] Piwowar Heather, Day Roger, & Fridsma, Douglas. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308.

[15] Searls, David B. 2001. From Jabberwocky to genome: Lewis Carroll and computational biology. *J Comput Biol.* 8(3):339-48.

[16] Searls, David B. 2005. Data integration: challenges for drug discovery. *Nature Reviews Drug Discovery* 4, 45-58 (January 2005).

[17] Buneman, Peter, Cheney, James, Tan, Wang-Chiew and Vansummeren, Stijn. 2008. Curated databases. *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS '08). pp. 1-12.

[18] Lesk, Michael. 2012. Curators of the future. Invited for *Library and Information Technology* special issue.

[19] Vogel, Gretchen. 2011. Dutch 'Lord of the Data' Forged Dozens of Studies. *ScienceInsider*. http://news.sciencemag.org/scienceinsider/2011/10/report-dutch-lord-of-the-data-fo.html.

[20] Price, Michael. 2011. Career Advice: To Replicate or Not To Replicate? *Science Careers* (Dec. 2, 2011*)*. http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2011_12_02/caredit.a1100133.

[21] ASSTRA. 2010. Fare Evasion in Public Transport. http://www.apta.com/mc/fctt/previous/2010fare/Presentations/ASSTRA-Fare-Evasion-in-Public-Transport.pdf.

[22] Clark, Liat. 2012. Nature journal wins libel case, highlights holes in UK defamation law. *Wired.* <http://www.wired.co.uk/news/archive/2012-07/06/nature-wins-libel-case>.

[23] Boseley, Sarah. 2010. Simon Singh libel case dropped. *The Guardian*, 15 April. http://www.guardian.co.uk/science/2010/apr/15/simon-singh-libel-case-dropped.

[24] Frankel, Alison. 2012. Libel suit over autism-vaccine link lands in the heart of Texas. *Thomson Reuters News & Insight/Westlaw.* http://newsandinsight.thomsonreuters.com/Legal/News/2012/01_-_January/Libel_suit_over_autism-vaccine_link_lands_in_the_heart_of_Texas/.

Start Making Sense: Quality, Context & Meaning

Jerome McDonough

University of Illinois at Urbana-Champaign
Graduate School of Library & Information Science
501 E. Daniel Street, MC-493, Champaign, IL 61820
jmcdonou@illinois.edu

1. Introduction

Klein, Moon & Hoffman [1] defined sense-making as "a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively." Given such a definition, the entire research enterprise, collectively and individually, can be understood as a form of sense-making activity. Regardless of discipline, research is intended to help us make sense of our world and by so doing help us better negotiate our course through life.

The past several years have seen a rapid increase in the creation of research data and an attendant increase in concern regarding that data's continued availability and usefulness. The rise of data mining and data-driven science has meant that loss of data may mean not only losing documentation of previous scientific discoveries, but also forfeiting opportunities for further research and discovery. Data curation thus aims not only at preserving existing knowledge, but preserving the possibility of new discoveries.

I would argue that any discussion of data quality should take as its foundation sense-making theory as set forth by researchers such as Dervin [2] and Weick [3]. The job of the curator is to help insure that researchers can continue to 'make sense' of data, where 'make sense' carries a dual meaning: both enabling researchers to apprehend the meaning of a particular set of data, and to create new meaning based on that information. Quality data is data which is in optimal condition to allow researchers to 'make sense' of it, for both meanings of sense-making. If we accept that the goal of data curation is to assist in researchers' sense-making activities, then we must ask what specific aspects of data are likely to contribute to easing its apprehension and its application in the future.

2. Facets of Data Quality

Our dual meanings of sense-making give us a starting point for identifying the various aspects of data that determine its quality for particular uses. With respect to scholars' apprehension of data, we can say that quality data exhibits the following characteristics:

Quality data is accessible. Good data is data a scholar can actually use, which implies that both technical and social impediments to its use have been minimized or eliminated. Too often we tend to think of technical aspects of access in terms of whether we can deliver a copy of information to those who want it. In an era of big data, however, access will not be solely a matter of delivering information, but providing scholars with the ability to work with data where it resides. This implies the existence of open technological infrastructures that allow for data to be examined and processed *in situ*. Quality data is data that provides scholars with not only online access, but also the necessary computational access.

Access also has a social dimension, a fact that grant agencies such as the NSF and NIH have recognized and acted upon over the last decade. Those who create data must be willing to share it and do so on terms that enable its use by others. The use of the Creative Commons public domain license for data sets by journals like *GigaScience* (and its associated data repository *GigaDB*[1]) provides an interesting model for handling the social aspects of data accessibility.

Quality data is locatable. Being able to access data is of no moment if you're not aware that the data exists. Just as it is too easy to think of access as the ability to deliver a copy of an item, it is too easy to think of the problem of locating data as a matter of creating a precise metadata record to enable search and retrieval. While descriptive metadata certainly plays a role in insuring the discoverability of data, it should be remembered that different communities of practice may benefit from the same data set, and a metadata record which suffices for search for one community may not be at all adequate for another. Portability/usability of description across disciplinary boundaries is a potential issue for the on-going discoverability and use of data.

The ability to locate data is also implicated in processes surrounding journal publication and preservation repositories. Those involved in Web archiving can testify to the dramatically short half-life of URLs on the Web. Scholars may be as likely (perhaps more likely) to learn of the existence of potentially valuable data from scholarly publications as they are from searching data repositories. Quality data is data with *persistent* integration with the larger scholarly corpus that enables its identification and retrieval.

Quality data is contextualized. Assuming a scholar can identify and locate (as well as access) a potentially useful data set, they need to be able to understand it well enough to determine its appropriateness for their own purposes. That requires adequate documentation of the data set's creation, use, handling and history. This information includes, but is broader than, what we typically think of as provenance. It includes not only typical provenance information such as the identities of the data's creators, reason for its creation, processes applied to the raw data before its release, chain of custody, etc., but also information about the larger research project which drove the creation of the data, its relationship to other data sets, its subsequent use for other projects not originally envisioned by its creators, identification of scholarly literature drawing upon the data, and more. A scholar examining a potentially useful data set needs the information that

[1] See *GigaDB*'s terms of use at http://gigadb.org/terms-of-use/.

will allow them to determine not merely what the data contains but what role the data has played in research over time, what has made it suitable (or unsuitable) for particular uses, and whether it has been superseded by more recent research activity.

Quality data is stored in formats that allow it to persist. The digital preservation community has done significant work in the last decade in identifying sustainability factors for file formats (see [4] and [5]). These issues of sustainability are of particular significance in the realm of research data. A slight shift in the color space of a digitized book occurring during a migration of page image data to a new format might be regrettable but would typically not be seen as an existential threat to the data. A similar shift to a video stream received from the Mars rover *Curiosity* could have a disastrous impact on research uses of the data. Precision matters in research data, and all practical measures which can be taken to insure that the information contained within the data remains unchanged should be, including showing great caution with respect to formats chosen to originally contain the data and new formats to which it might be migrated.

Quality data is stored in formats that allow it to be re-used. As noted in the discussion of access, good data is data a scholar can actually get at, where 'get at' may have a variety of meanings. While we want data to be stored in formats that persist, we also wish it to be in formats that simplify to the greatest extent possible scholars' ability to apply the data in new situations. Obviously these issues of persistence and usability may be in tension, particularly given the emphasis on insuring data's continued integrity.

Quality data is suitable for the task in hand. There is a strong tendency in discussions of data quality to invoke factors such as precision, accuracy or currency, to assume that quality is something intrinsic to the data. It is debatable whether a view of quality as intrinsic is ever appropriate, but it is certainly not an appropriate response when considering data's long-term curation and re-use. As an example, Wikipedia is hardly something that most scholars would point to as the epitome of accuracy; yet that in itself has made it a research subject, and thus a unique and valuable piece of research data [6, 7]. For these scholars, the accuracy of Wikipedia is the subject of research, not a pre-condition for its use. A 'corrected' Wikipedia would be of no use to their research. Quality is a relative determination rendered with respect to a particular task or purpose and is not something for which a curator can make *a priori* judgments. Duranti's views regarding the appropriate role of appraisal in archives and the attribution of value seem relevant here; an archivist's role is to preserve evidence, not determine the truth.

3. A Research Agenda for Data Quality

What does a sense-making view of data curation activity suggest in terms of a research agenda for the field of LIS? The quality facets outlined above suggest several areas where data curation may encounter difficulties that further research might help alleviate. In the area of contextualizing data sets to enable their ready apprehension, it would be of some value to the data curators of the world to have a clearer idea of where they might most productively focus their efforts. In particular, more comprehensive examination of the nature of scientific communities of practice and the identification of regular areas of interdisciplinary exchange/overlap could help curators both in knowing what communities are likely to need to try to interact with each others data, what their typical methodological and epistemological approaches to data are and how that might influence their appropriation of data from outside their immediate community. This in turn might assist in prioritizing curators' efforts in creating new contextualizing descriptions of data.

As a related issue, both creating contextualizing descriptions of data and more traditional forms of description for retrieval are time-consuming and expensive processes. Far more study is needed of ways to automate the production of descriptions, and where automation is not feasible, to speed and simplify their manual production. As noted earlier, we also cannot assume that descriptions intended to assist retrieval for one community may necessarily work as well for another. Research has been done on mapping between subject vocabularies so that individuals familiar with one vocabulary may employ it as an entry point for databases employing somewhat different vocabularies [8, 9, 10], but much of this research has focused on traditional library materials and traditional forms of description. Further work is needed on mechanisms to allow scholars to easily identify useful data that may exist outside their typical range of disciplinary experience.

Ancillary to these issues of creating descriptions of data are problems regarding the standardization of vocabularies and the design of vocabularies that are employable across disciplinary domains. Much of the work that has been done in the design of controlled vocabularies has had the unspoken assumption that a vocabulary was intended for the use of a particular community of practice and should address the entirety of their vocabulary needs. We should at least contemplate the possibility that this approach is responsible for many of the problems that we have with cross-domain retrieval, and consider whether alternative approaches to vocabulary creation, ones which anticipate their use by multiple communities of practice rather than a single one, might simplify the job of curators and those looking for research data.

Somewhat more broadly, we need more research into how the preservation of different types of information can/should mutually reinforce each other. The European Commission's Information Society and Media Directorate-General organized a meeting in 2011 to discuss the European Union's funding priorities for research into digital preservation [12]. One discussion point that emerged from that workshop was that there needed to be a shift from focusing on the preservation of data to the preservation of *knowledge*. Data sets are not information islands; they are part of a larger framework of documents that include project reports, journal articles, press interviews, standards documents, and a host of other related pieces of information. Both because of the need to insure that data sets are appropriately contextualized and the need to insure that they are more completely discoverable (that is, persistently integrated with the larger document sphere), we need to achieve a better understanding of how data curation as an activity integrates with the larger problem of knowledge preservation.

As previously noted, there may be at least some tension between a desire to store data in formats which render it fit for long-term access and a desire to store data in formats which render it fit for immediate use. Some work has already been done on risks associated with file formats, but it would be useful to see more such work done focusing on formats more likely to be used by the scientific community for data storage; much of the work to date has focused on formats employed by the digital library community, and while there is some overlap in the use of formats, obviously scientific data is stored in a variety of formats unlikely to be seen within many digital libraries at this point.

A related issue that should be examined is determining what problems existing scientific communities have in appropriating and using data from outside their own communities. Some research efforts (I'm thinking in particular of the Long-Term Ecological Research Network) have attempted to proactively define a larger, interdisciplinary community and establish common practices and standards to simplify sharing data among its members. While that has proved a fruitful approach, it probably will not prove practical in all instances, and trying to better understand what impedes groups from using data today would obviously help guide curation efforts. A more in-depth examination of the problems (and solutions) already in play with respect to sharing of research data across disciplinary boundaries is essential to improving curation.

4. Conclusion

If we review the above facets and potential research topics related to curation and data quality, I believe an over-arching theme emerges: while the digital preservation and digital curation research communities have made tremendous progress in the past decades, it may be time for us to consider the words of T. S. Eliot:

> *Where is the knowledge we have lost in information?*[13]

The ultimate goal of curation is more properly the preservation of knowledge than the mere preservation of information. However, knowledge is something that resides in the mind and not on the page (or the disc). Properly speaking, we cannot preserve knowledge in the long-term; we can only preserve people's building to acquire knowledge.

However, if we recast the job of curation as one of preserving the ability to acquire knowledge rather than simply preserving information, as a job of helping people make sense of the world, it throws a somewhat different light on both the job of the curator and the job of the researcher looking to support curators' ability to preserve knowledge. We need to widen our research focus so that we are no longer looking at the curation of research data in a vacuum, but examining its complex and messy relationships with other forms of information, with various communities of knowledge and practice, and with the institutional and policy structures affecting them all. We need to place data in its larger context, because only when it is so situated will people be able to fully exploit it to make sense of our world.

5. References

[1] Klein, G., Moon, B. and Hoffman, R. R. 2006. Making sense of sensemaking 1: Alternative perspectives. IEEE Intelligent Systems 21(4) (July/August 2006), 70-73.

[2] Dervin, B. 1998. Sense-making theory and practice: An overview of user interests in knowledge seeking and use. Journal of Knowledge Management 2(2), 36-46.

[3] Weick, K. E. 1995. Sensemaking in Organizations. Thousand Oaks, CA: Sage Publications.

[4] Library of Congress. 2011. Sustainability of Digital Formats: Planning for Library of Congress Collections. Washington, DC: Library of Congress. Accessed Aug. 19, 2011 at http://www.digitalpreservation.gov/formats/index.shtml

[5] Brown, A. 2003. Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation. DPGN-01. Surrey, UK: The National Archives.

[6] Clauson, K. A., Polen, H. H, Kamel Boulos, M. N and Dzenowagis, J. H. 2008. Drug information: Scope, completelness, and accuracy of drug information in Wikipedia. *The Annals of Pharmacotherapy* 42(12), 1814-1821.

[7] Rector, L. H. 2008. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review* 36(1), 7-22.

[8] Duranti, L. 1994. The concept of appraisal and archival theory. *The American Archivist* 57(2), 328-344.

[9] Buckland, M., Chen, A., Chen, H.-M., Kim, Y., Lam, B., Larson, R., Norgard, B., Purat, J., and Gey, F. 1999. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine* 5(1), Jan. 1999.

[10] Chan, L. M., and Zeng, M. L. 2002. Ensuring interoperability among subject vocabularies and knowledge organization schemes: a methodological analysis. *68th IFLA Council and General Conference, Aug. 18-24, 2002, Glasgow*.

[11] Buckland, M., and Shaw, R. 2008. 4W vocabulary mapping across diverse reference genres. In Arsenault, C. and Tennis, J. T. (Eds.) *Culture and Identity in Knowledge Organization: Proceedings of the Tenth International ISKO Conference, 5-8 August 2008, Montréal, Canada*, 150-157.

[12] Billenness, C. S. G. 2011. *The Future of the Past – Shaping New Visions for EU-Research in Digital Preservation: Report on the Proceedings of the Workshop, Luxembourg, 4-5 May, 2011*. Cultural Heritage and Technology Enhanced Learning, European Commission Information Society and Media Directorate-General.

[13] Eliot, T.S. 1980. *The Complete Poems and Plays 1909-1950.* Orlando, FL: Harcourt Brace & Co.

A Plan For Curating “Obsolete Data or Resources”

Michael L. Nelson
Old Dominion University
Norfolk, VA USA
mln@cs.odu.edu

ABSTRACT

Our cultural discourse is increasingly carried in the web. With the initial emergence of the web many years ago, there was a period where conventional mediums (e.g., music, movies, books, scholarly publications) were primary and the web was a supplementary channel. This has now changed, where the web is often the primary channel, and other publishing mechanisms, if present at all, supplement the web. Unfortunately, the technology for publishing information on the web always outstrips our technology for preservation. My concern is less that we will lose data of known importance (e.g., scientific data, census data), but rather that we will lose data that we do not yet know is important. In this paper I review some of the issues and, where appropriate, proposed solutions for increasing the archivability of the web.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]

Keywords

Curation, Web Archiving, Memento

1. WHO WANTS “OBSOLETE DATA”?

Perhaps the largest problem facing web archiving is that it remains at the fringes of the larger web community. The most illustrative anecdote pertains to a web archiving paper we submitted to the 2010 WWW conference. One of the reviews stated:

> Is there (sic) any statistics to show that many or a good number of Web users would like to get obsolete data or resources?

This is just one reviewer, but the terminology used (“obsolete data or resources”) succinctly captures the problem: web archiving is not widely seen as a priority or even as in scope for a conference such as WWW. Another common related misconception we have encountered is that the Internet Archive has every copy of everything ever published on the web, so preservation is a solved problem. Despite the heroic efforts of the Internet Archive, the reality is more grim: only 16% of the resources indexed by search engines are archived at least once in a public web archive [1].

While there are many specific challenges with regards to quality criteria, tools, and metrics, the common thread goes back to the fact that we, the web archiving community, have failed to articulate clear, compelling use cases and demonstrate immediate value for web preservation. For too long web preservation has been dominated by threats of future penalties, such as hoary stories about file obsolescence that have not come true[1]. The lack of a compelling use case for archives has relegated preservation to an insurance-selling idiom, where uptake is unenthusiastic at best.

2. I BLAME THOMPSON AND RITCHIE

The web has a poor notion of time, and it is getting worse instead of better. An early design document for the Web addressed the problem of generic vs. specific resources [2]. That document identified three dimensions of genericity: time, language (e.g., English vs. French), and representation (e.g., GIF vs. JPEG). The latter two dimensions were the basis for HTTP content negotiation as originally defined in HTTP/1.1 [5]. Content negotiation allowed, for example, GIF and JPEG resources to have unique URIs (i.e., specific resources), but to be joined together with a third, generic resource with its own URI. When a client dereferences this generic URI, the appropriate specific resource is selected based the client’s preferences for representations. Content negotiation works similarly for language, but content negotiation in the dimension of time was not part of the original HTTP core technologies (the Memento project added content negotiation in the dimension of time in 2009 [11]). One result of not having time as part of the core technologies is that the web community’s concept and expectations regarding time have not become fully mature.

I believe the reason for this underdeveloped notion of time can be traced to the tight historical integration of HTTP and Unix, specifically the Unix filesystem. Metadata about files in the Unix filesystem is stored in “inodes”, and the original description of the Unix filesystem defined three notions of time to be stored in an inode: file creation, last use, and last modification [8]. However, at some early point the storage of the file creation time in the inode was replaced with the last modification time of the inode itself. The result was that we could know the last modification and access times of a file, but the creation time, a crucial part of establishing prove-

[1] David Rosenthal has a series of convincing blog posts on this topic, see: http://blog.dshr.org/2010/09/reinforcing-my-point.html

nance, was lost (most URIs contain semantics, and creation time can be critical in establishing priority). Although web resources and Unix files are logically separate, in practice they were tightly integrated during the formative years of the web, and so the HTTP time semantics were limited by what could be provided by the Unix inode. For example, here is an HTTP response about a JPEG file:

```
% curl -I http://cdn.loc.gov/images/img-head/logo-loc.png
HTTP/1.1 200 OK
Date: Sun, 19 Aug 2012 13:30:06 GMT
Server: Apache
Last-Modified: Fri, 03 Aug 2012 03:54:26 GMT
Content-Length: 1447
Connection: close
Content-Type: image/png
```

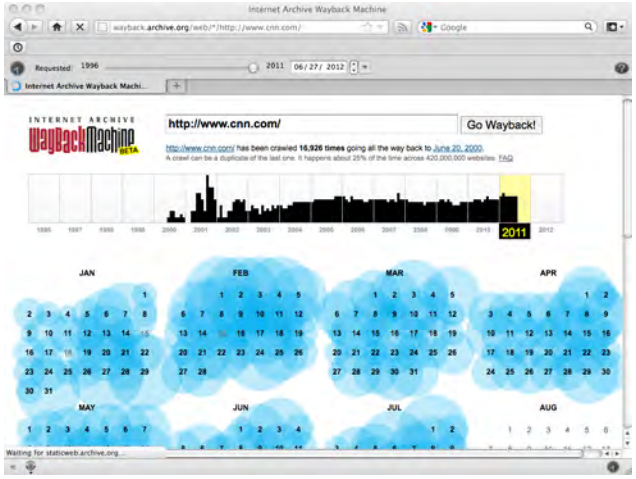
In the above example, the server is expressing the response was sent on August 19th, but the JPEG file itself was last modified on August 3rd. Notable by its absence is the creation time: via the inode limitations, we cannot know when this file was created. It might have been created on August 3rd or it might have been created at an earlier time, and being unable to establish even this basic level of metadata is a severe limitation for archiving and provenance. Unfortunately, even the limited semantics of last modified are becoming less frequent as more resources are dynamically generated. The example below is in response for a dynamically generated home page:

```
% curl -I http://www.digitalpreservation.gov/
HTTP/1.1 200 OK
Date: Sun, 19 Aug 2012 13:30:33 GMT
Server: Apache
X-Powered-By: PHP/5.2.8
Connection: close
Content-Type: text/html
```

In the above example, there is the data of the response (August 19th), but last modified times for dynamically generated representations are not defined. Dynamically generated resources make possible the web as we know it today, but the net result is even fewer time semantics are present in HTTP responses. Evolving publishing technologies such as personalization, Ajax, Flash, and streams[2] will only serve to make it more difficult to ascribe a creation time to any particular web page.

3. W{H}ITHER ARCHIVES?

I maintain that the entire web community has a poor notion of time and are trapped in the "perpetual now". Because the lack of capability has shaped our expectations, we never object when prior versions of web pages are unavailable. We tolerate temporal inconsistency in our browsing, even 404 errors, in part because we do not know enough to expect better. Remember "lost in hypertext" [4, 3]? That has been solved in part through better navigation tools and design practices, but also in part due to increased familiarity with the hypertext navigation metaphor. Now imagine if a temporal dimension was added for each page – there would be much confusion, but eventually tools, practices, and user awareness would prevail.



http://www.cnn.com/ has been crawled **16,926 times** going all the way back to June 20, 2000.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ

Figure 1: All available versions of cnn.com at the Internet Archive. This page is not reachable from cnn.com.

3.1 Archives Are Not Destinations

The most fundamental problem is that we have designed web archives as if they are destinations in themselves. The motif of "go to the library/archive and spend an afternoon in the stacks" has been replicated in our web archives. Figure 1 shows the list of archived pages (or "mementos") for `cnn.com` at the Internet Archive. If you want to browse the past versions of this news site, you go to the archive and perform a browsing session within the archive, and then return to the live web once you are done with your journey to the past.

In our experience, most web users do not know about the Internet Archive or how to access it. The Memento project has demonstrated a framework for tighter integration of the past (i.e., archived) web and the current web, but the tools exist as add-ons for both servers and clients and have yet to reach mainstream acceptance, which will only arrive when the archiving community can demonstrate a "killer app" that will cause users to demand the functionality.

3.2 Web Archiving Is Not Social

I am not sure what an archiving killer app would look like, but there is a good chance it will be social. People like to share links with each other via Twitter, Facebook, Pinterest, et al. However, with the exception of Pinterest (which makes copies of "pinned" images) this sharing is done by-reference and not by-value, exposing it to the same link rot problems of common web pages (for example, we found 10% of the shared links about the Egyptian Revolution were lost after one year [9]). I am constantly surprised at the tasks that people are willing to undertake if there is a social or gaming component (i.e., "games with a purpose"), yet I am unaware of any such activity with a web preservation component. Diigo (`diigo.com`) is a site that provides social bookmarking services (similar to Delicious) with an archiving component, but enthusiasm for social bookmarking seems to be less than it once was.

[2] For example, see Anil Dash's call to "Stop Publishing Web Pages" in favor of streams: `http://dashes.com/anil/2012/08/stop-publishing-web-pages.html`

A web archiving application that could leverage the collection development of Pinterest and the collaborative editing of Wikipedia and other wikis would be a welcome development. Archive-It (`archive-it.org`) is nearly such an application, but it is targeted for archiving and librarian professionals, not as a general purpose social application. Perhaps the legal challenges[3] of creating such collections would prevent the development of such an application, but I would observe that early legal challenges about the mechanics of HTTP and "making copies" were eventually overcome.

3.3 Watchdog Archiving and Trust

Perhaps a social web archiving activity that will grow to take on a larger role is that of distributed, citizen watchdogs of public figures and politicians. For example, a supporter of blogger Andrew Breitbart brought down Congressman Anthony Weiner by zealously following and archiving Weiner's twitter feed[4]. Most tweets are of arguably limited historical value, but this particular tweet and the fact that it could not be fully redacted turned out to have significant political and cultural implications.

In another example, consultant and commentator Richard Grenell deleted over 800 tweets after he was elevated to a senior position in the Romney campaign in 2012[5]. Presumably Grenell's lesser status at the time did not warrant a corresponding campaign to monitor and archive Grenell's twitter feed like there was with Weiner's twitter feed. Grenell's tweets most likely do not exist outside of Twitter's own archives (and those they share with the Library of Congress).

And what if someone did come forward with a correspondingly damning tweet from Grenell, how could we verify it? Aside from Weiner's ultimate confession, was his tweet ever verified by an independent third party? And if so, how would we trust such a third party – where would the chain of trust terminate? Could he not find a technologically savvy staffer to fabricate evidence that contradicted Breitbart's evidence (which is especially easy given the low level of provenance regarding third-party archives)? It is easy to envision a market for a trusted, tamper-proof archive for tweets and other social media so a person can *deny* that they ever released an offending tweet?

Our current approach to web archiving involves implicitly trusting the Internet Archive and other public web archives as incorruptible. Eventually the magnitude of scandals associated with web content will grow to the point where less scrupulous web archives will be offered as proof. A combination of trusted archives and citizen activism might form the basis for the first killer app for web archiving. Instead of canvassing a neighborhood, volunteers can canvass/archive web pages.

4. WISH LIST

This section contains a personal wish list of features that would make archiving web pages much easier.

4.1 Machine-Readable Time Semantics

We have moved beyond the limitations of the Unix filesystem and its inode, so we should increase the time semantics in our HTTP transactions. Unfortunately, this is not the case. In the example below, when dereferencing the URI of a specific tweet, twitter.com shows a last modified time that matches the date the response was generated (this is true for all responses, not just this one). More importantly, Twitter has a concept of time similar to "Memento-Datetime", which captures the time a page was first observed on the web (see [7] for a discussion of how this differs from "Last-Modified"). Although this date (June 27, 2012 in this example) is displayed in the HTML page and is accessible to authenticated users via the Twitter API, the correct date semantics are not presented, and the incorrect value for the last modified time is presented instead. This phenomenon is not unique to Twitter, but Twitter makes for a good example due to its well-known nature.

```
% curl -I http://twitter.com/machawk1/status/218015444496416768
HTTP/1.1 200 OK
Date: Mon, 20 Aug 2012 00:41:38 GMT
Content-Length: 85440
Last-Modified: Mon, 20 Aug 2012 00:41:38 GMT
Content-Type: text/html; charset=utf-8
Server: tfe
```

4.2 APIs for Archives

Talk to anyone who has built applications using archived web data and they will have crawled and "page scraped" the archives at some point. Page scraping puts an undue burden on the archive itself, is error prone, and doesn't facilitate inter-archive interaction. The Memento project defines a simple, inter-archive HTTP access mechanism, but this is not enough. The Internet Archive's Wayback Machine software supports a simple API for file upload and searching, but this API is not evolved like APIs for services like Google, Twitter, and Facebook. If we want archives to be used in the current web programming idiom, we have to go beyond the "afternoon in the stacks" model (see section 3.1) and provide fully-featured APIs.

4.3 Impedance Matching

The Internet Archive does not have full-text search on the main Wayback Machine. While this is a limitation, it is probably not as big a limitation as many think, in part because it is not clear what we would do with full-text search at this scale if we had it (cf. the discussion in section 3). The kinds of questions that scholars wish to answer using web archives are of the form "what role did the Tea Party play in the 2010 mid-term elections?" The kind of access we can offer right now is "this is what `cnn.com` looked like November 1, 2010." Adding full-text searching, while useful in some cases, would not immediately help address the kinds of questions that scholars want to ask. An example of the kind of advanced analysis that needs to be performed on web archives is entity tracking experiments of the LAWA project [10], in which entities (e.g., people, companies) can be tracked through time and different URIs.

5. CONCLUSIONS

I expect data of known value to be successfully curated and available well into the future. I am more concerned with our cultural record, with which we have made a Faustian bargain of increased volume and ease of access (i.e., the

[3] A discussion of which is beyond the scope of this paper; for a primer see `http://1.usa.gov/QgaUZO`

[4] See `http://en.wikipedia.org/wiki/Anthony_Weiner_sexting_scandal`

[5] See: `http://huff.to/I6dpQo`

web) at the expense of permanence and provenance (i.e., paper). We are stuck in the perpetual now and due to the initial limitations of the Unix inode, the notion of varying temporal access to web pages is so unexpected that even web researchers need to be convinced of the utility.

One problem is the limited design motif for web archives: destinations that are wholly unconnected from their live web counterparts. The related problem is that we, as a community, have failed to envision and deliver a "killer app" for web archiving. Perhaps it is in a watchdog role over public figures and institutions. Or perhaps the emerging field of personal digital preservation[6] will energize the field and increase what are often laissez-faire user expectations regarding archiving [6].

I would like to see a more careful approach to specifying temporal semantics in common web services like Twitter. Similarly, I expect web archives to offer richer APIs for accessing their content, and to eventually offer the higher-level services, like entity tracking, that will assist scholars in using the ~~obsolete data or resources~~ archives.

6. ACKNOWLEDGMENTS

This work sponsored in part by the Library of Congress, NSF IIS-0643784 and IIS-1009392.

7. REFERENCES

[1] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? In *Proceeding of the 11th annual international ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, 2011.

[2] T. Berners-Lee. Web architecture: Generic resources. http://www.w3.org/DesignIssues/Generic.html, 1996.

[3] J. Conklin. Hypertext: A survey and introduction. *IEEE Computer*, 20(9):17–41, 1987.

[4] W. Elm and D. Woods. Getting lost: A case study in interface design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 29, pages 927–929, 1985.

[5] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. Hypertex Transfer Protocol – HTTP/1.1, Internet RFC-2068, 1997.

[6] C. Marshall, F. McCown, and M. L. Nelson. Evaluating personal archiving strategies for Internet-based in formation. In *Proceedings of IS&T Archiving 2007*, pages 151–156, May 2007.

[7] M. L. Nelson. Memento-Datetime is not Last-Modified. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html`, 2011.

[8] D. Ritchie and K. Thompson. The UNIX time-sharing system. *Communications of the ACM*, 17(7):365–375, 1974.

[9] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: How much social media content has been lost? In *TPDL*, 2012.

[10] M. Spaniol and G. Weikum. Tracking entities in web archives: the LAWA project. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, 2012.

[11] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.

[6] See for example: `http://www.personalarchiving.com/`

Position Papers: 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 will:

- 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

The Economics of Data Integrity

Ricky Erway
OCLC Research
777 Mariners Island Blvd. #550
San Mateo, CA 94404
01-650-287-2125
erwayr@oclc.org

Brian Lavoie
OCLC Research
6565 Kilgour Place
Dublin, Ohio 43017
01-614-764-4399
lavoie@oclc.org

ABSTRACT

This brief paper is in response to a call for papers for the UNC/NSF Curating for Quality workshop. We describe the aspects of costs and sustainability of data curation as they pertain to data integrity.

Categories and Subject Descriptors

E.5 [**Data**]: General

General Terms

Economics

Keywords

Sustainability. Data curation.

INTRODUCTION

It is difficult to consider sustainability of data quality without defining what data quality entails.

To data creators and to those who reuse the data, data quality may refer to ensuring accuracy of the data and supplementing the data with rich description, ancillary materials, context, or other enhancements that facilitate leveraging value from the data. These aspects are addressed in other workshop papers.

This paper addresses data quality in the traditional library or archives sense: ensuring data quality consists of making sure the data is uncorrupted, is what it purports to be, and that it persists and is accessible into the future. This includes technical aspects (are the processes adequate to preserve the data?), social aspects (can we trust that the processes will be followed reliably?), and economic aspects (is there adequate ongoing funding to preserve the data into the future?). In a library context, these issues generally fall within the scope of long-term digital preservation. This paper advocates for achieving sustainability through economies of scale made possible by collaboration.

The final report of the Blue Ribbon Task Force on Sustainable Preservation and Access identifies five conditions for sustainable digital preservation[1]:

- recognition of the benefits of preservation by decision makers;
- a process for selecting digital materials with long-term value;
- incentives for decision makers to preserve in the public interest;
- appropriate organization and governance of digital preservation activities; and
- mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.

The focus of this paper is on this fifth condition, mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.

"… ALLOCATION OF RESOURCES…"

In allocating resources to digital preservation, we are essentially creating mechanisms to "transfer funding and other resources from those who benefit from and are willing to pay for digital preservation, to those who are willing to provide preservation services."[2] Sometimes these stakeholders groups are one and the same; in more complicated situations, they are distinct. In either case, the fundamental condition is clear: there must be recognition that allocating resources to ensure the long-term future of quality digital assets is a desirable, indeed necessary, activity. Without this recognition, the goal of maintaining long-term access to quality data is not achievable.

Allocating resources to digital preservation involves some key trade-offs that should be recognized upfront. One is that investing significant resources in curating high quality datasets could detract funds from producing new datasets via new research. Most funders are interested primarily in new research, creating a significant obstacle to preserving existing high quality datasets over the long term. The data management plans required by NSF and other grant-giving entities help to balance the two needs. One must also weigh the cost of preservation against the costs of replacing the data—and monitor when that balance tips in either direction. For example, for data produced through computer simulation models, it may be less costly to store the algorithms that produced the dataset than the data itself, thus preserving the option to re-create the data at a future time. Another possible trade-off is between sustainability and access. Generating a flow of funds to support long-term preservation may require charging a fee for access, which inevitably limits the scope for potential reuse to those able and willing to pay. Providing dark archive service is less expensive, but it has been shown that access has a positive effect on strengthening the incentive to preserve and also provides a monitoring function to alert the curator to changes to the data or to the need to migrate data to another format.

Another trade-off is the one between risk and reward. There are a variety of risks impacting the long-term sustainability of research data curation, from unexpected data loss to uncertainty about the value to future users. In allocating resources to the preservation of high-quality data, we are making a "bet" on realizing some future value from that investment (new scholarship, replicability of scientific findings, etc.). Only time will tell if efforts to preserve a particular dataset prove to be a wise allocation of resources. Decision-makers need to be prepared to revisit their preservation decisions frequently over time, and adjust their resource allocations accordingly.

Management of the trade-offs associated with preservation decision-making should be informed by the range of risks undertaken by service providers that could potentially impact the preservation process over time. Many times funding is available to establish a data archiving service, but is not available for ongoing operations. Likewise, funding may be included in a grant proposal to cover the costs of preparing a particular dataset for deposit in the archive, but not for its long-term care. Another challenge for service providers is that technology and preservation practices change over time, requiring acquisition of new hardware and software and reworking of processes. One of the best reasons to commit to preservation of a particular dataset is when the data is not reproducible. In these cases, the ramifications of failure are high.

CALCULATING COSTS

Before addressing where funding for data curation might come from, we need to have a sense of what the actual costs are. However each situation is unique and there are no set answers. For example, economies of scale can have a dramatic effect on the per-unit cost of preservation.

The Keeping Research Data Safe (KRDS) framework[3] provides a way to calculate the costs of data curation for a specific situation. Costs can be broken down by the activities in which they are incurred; these costs can then be adjusted to reflect the particular conditions of a given digital archiving scenario (service adjustments), and appropriately distributed over time (economic adjustments). For example, preservation activities can differ in how long the data is to be maintained; the type of online, near-line, and offline storage used; security requirements of the data; frequency of refreshment; and the number of versions, editions, or copies to be kept. Service limitations can control costs, e.g., limiting the file formats accepted or insisting on a standard IPR statement. Moreover, costs can spread over time via inflation and depreciation.

Though the results are very individual, case studies using the KRDS cost framework have turned up some indicative findings. In general:

- The costs involved in the generation of the data during research, far outweigh the costs of archiving the data.
- Archiving costs are highest up front and become less significant over time. This is true for the archive overall (set-up costs are far higher than operational costs) and for each dataset (the ingest process is more costly than the ongoing maintenance costs).
- Use of off-the-shelf software and hardware solutions brings costs down significantly.
- Initial capital costs of storage media and systems are less than a third of the overall costs of ownership.
- Staff costs exceed those of any other component. In academic institutions, staff costs range from 50% - 90% of total costs. The degree to which processes are automated can have a significant impact.
- The number of depositors can affect costs. One or more middlemen, aggregating submissions and ingesting them in a standard manner, will mitigate the high costs associated with a large number of depositors.
- Changes in workload can have substantial effects on unit costs. In one case, when workload increased 600%, costs increased only 325%.
- Timing can be a factor. For example, addressing data migration early on is much cheaper than attempting to migrate from an already obsolete format.
- In some cases, the costs of deaccessioning a dataset exceed those of continuing to maintain it.

"… ONGOING AND EFFICIENT …"

Ensuring economically sustainable datasets (and their associated services) goes far beyond simply allocating resources. It also involves using those resources efficiently and leveraging collaboration to achieve economies of scale. Furthermore, preservation is an ongoing process, so the flow of funds to preservation activities must also be ongoing if long-term preservation objectives are to be achieved.

The requirement that the allocation of resources to digital preservation needs to be ongoing over time seems obvious, yet it is too frequently neglected in practice. It is easy to find examples of long-term preservation projects that are funded through short-term, one-off grants. When the funding runs out, the preservation activity must scramble to find another grant or other resources to keep the project running for a while longer; alternatively, the project simply ends. An example of such a situation is the UK Arts and Humanities Data Service, which had been funded by JISC .[4]

Just as preservation activities require a long-term view of the maintenance of, and access to, data assets, so too do they require a long-term view of their funding. Mechanisms that secure a reliable, ongoing flow of resources are the optimal way to fund long-term activities such as digital preservation.

Efficient use of available resources is another necessary aspect of sustainable digital preservation. Economies of scale argue for collaboration, leveraging fixed costs over a larger number of deposits. Data curators from the library, archive, and information technology sectors can ingest and preserve datasets. But ensuring the *quality* of data requires specific subject area expertise, due to varying needs of the disciplines. This is perhaps an even more significant opportunity for economies of scale. If every repository had to have a wide range of subject experts, the costs would be prohibitive.

In some university settings, data is curated in the department of origin. Here the benefit is the proximity of the researchers and others who understand and might use the data. In this case, however, the technical curation skills may be lacking.

A collaborative approach allows for a pool of subject specialists that serve a wide range of depositors and draw on a pool of people with experience in various technical aspects of data curation. An example is the Interuniversity Consortium for Political and Social

Research (ICPSR),[5] which hosts research data files in the social sciences on behalf of 700 institutions. Staff with specialties in various fields work closely with researchers to prepare data for submission and ensure data integrity, while staff with technical skills are tasked with the preservation component.

An informal network of subject-based data repositories allows subject specialization at each repository. A related example is the array of disciplinary repositories for research preprints and published articles. Aggregating dataset deposits for a particular discipline not only allows for specialized help for ingest and ensuring data quality, but it also allows for aggregation of users. A single set of functionality and support services can meet the needs of researchers in that particular field.

Specializing in a narrow discipline can encourage compartmentalization. A benefit to the ICPSR approach is that it encourages cross-disciplinary research.

In some countries a national approach is taken, as with the DANS service in the Netherlands.[6] In the US, it is less likely that we would have a single national service, but a national *network* of data archives would help with discovery of relevant datasets for reuse and would facilitate multidisciplinary research. Additionally, a central infrastructure that provides support for locally-curated datasets might help in disciplines that aren't as well-funded as some of the big sciences.

No matter what approach is taken, preservation planners need to be cognizant of opportunities to lower the per-unit cost of preservation by spreading costs over higher volumes of preservation activity. Digital preservation is a shared problem, and shared problems often lend themselves to shared solutions through collaboration.

SUSTAINABLE BUSINESS MODELS

There are a number of ways to supplement start-up funding and to provide ongoing financial support. The report, *Lasting Impact: Sustainability of Disciplinary Repositories*,[7] identifies several different business models for disciplinary document repositories:

- Institutional support
- Use-based institutional contributions
- Support via consortium dues
- Distributed network of volunteers
- Federal government funding
- Decentralized arrangement
- Commercial "freemium" service (basic access is free; value-added services for fee)

The report notes that, in most cases, a combination of funding sources is used. These funding models can equally apply to data repositories. ICPSR, for example, is supported by member fees, use fees, and grants. [A more thorough listing of 155 repositories and their funding models is available from DataCite.[8]]

RECOMMENDATIONS

Because data curation involves many invested providers and beneficiaries—and many of those involved have both roles—there is much potential for addressing the challenges. The following are recommendations for optimizing for sustainability.

- Have a discipline-specific entity in between the researchers and the repository to help with setting policy regarding aspects such as selection criteria, retention periods, and transfers of stewardship.
- Use aggregators to work with depositors to normalize their submissions prior to ingest.
- Due to high degrees of change and uncertainty, agreements between content providers and repositories should include options to review, renew, refine, or terminate.
- Funders should consider providing ongoing support for trustworthy data archives and encourage automation developments that will decrease the number of manual processes.
- Institutions, funders, and publishers should impose and enforce meaningful mandates.
- The academy should recognize datasets as first-class scientific contributions in academic credentials to provide a personal incentive for researchers to prepare and submit their data for archiving.
- Because so much is in flux, we include this final counsel from the Blue Ribbon Task Force: "Hedging against uncertainties, postponing decisions when possible, recognizing that benefits, demand, and users will change, anticipating better information over time—these are the habits of mind that mark responsible digital stewardship and will help husband scarce resources while creating enough flexibility for bold moves and rescue of endangered assets when that becomes necessary."[9]

CONCLUSION

All academic institutions have or will have a need for some sort of data curation, but it is unrealistic to think that every institution will establish local data curation capacity. Due to the need for specialization in each subject, the need for a range of curation skills, the risks undertaken, and the economies of scale, it is unwise to attempt to replicate a broad range of data curation services, infrastructure, and expertise at every institution. Institutions so inclined might specialize in a particular field and offer services to all researchers in that discipline. Scholarly societies, government agencies, and commercial entities might take on similar roles. Consolidated solutions, where systems, infrastructure, and expertise can be spread over higher volumes of curation activity, offer lower per-unit costs. From the access perspective, specialized data repositories can focus on the needs particular to those who may want to reuse those datasets. Taking it up one more level, having broad discipline coverage, like ICPSR, or aggregated access to datasets at specialized repositories, facilitates interdisciplinary research. When these services are raised to the network level, expertise, economies, and benefits are shared.

REFERENCES

[1] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. February 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information; Final Report of the Blue Ribbon Task Force on Sustainable Preservation and Access.* p. 12 http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

[2] Lavoie, Brian F. 2012. Sustainable research data. In *Managing Research Data*, G. Pryor, Ed. Facet Publishing, London. p. 70.

[3] Beagrie, Neil., Julia Chruszcz, and Brian Lavoie. April 2008. *Keeping Research Data Safe: a cost model and guidance for UK universities, Final Report.* http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

[4] Arts and Humanities Data Service http://www.ahds.ac.uk/.

[5] Inter-university Consortium for Political and Social Research www.icpsr.umich.edu/.

[6] Data Archiving and Networked Services http://www.dans.knaw.nl/en.

[7] Erway, Ricky. 2012. Lasting Impact: Sustainability of Disciplinary Repositories. OCLC Research, Dublin, Ohio. http://www.oclc.org/research/publications/library/2012/2012-03.pdf.

[8] DataCite http://datacite.org/repolist.

[9] BRTF p. 47

Quality Control and Peer Review of Data Sets: How do Data Archiving Processes Map to Data Publication Requirements?

Matthew S. Mayernik
National Center for Atmospheric
Research (NCAR)
University Corporation for Atmospheric
Research (UCAR)
Boulder, CO
mayernik@ucar.edu

ABSTRACT

Establishing the quality of data sets is a multi-faceted task that encompasses many automated and manual processes. Traditionally, research quality has been assessed by peer review of textual publications, such as journal articles, conference proceedings, and books. This paper discuss the question of whether the peer review process is appropriate for assessing and ensuring the quality of data sets.

General Terms

Management, Documentation, Human Factors.

Keywords

Peer review, data management, data citation, data quality

1. INTRODUCTION

Data sets exist within scientific research and knowledge networks as both technical and non-technical entities. Establishing the quality of data sets is a multi-faceted task that encompasses many automated and manual processes. Data sets have always been essential for science research, but are becoming more visible as first-class scholarly objects at national, international, and local levels. Many initiatives are establishing procedures to publish and curate data sets, as well as to promote professional rewards for researchers that collect, create, manage, and preserve data sets [see for example 2,4,7,8]. Traditionally, research quality has been assessed by peer review of textual publications, such as journal articles, conference proceedings, and books. Citation indices then provide standard measures of productivity used to reward individuals for their peer-reviewed work. Whether a similar peer review process is appropriate for assessing and ensuring the quality of data sets remains as an open question.

2. PEER REVIEW AS THE GOLD STANDARD?

The peer review system is currently under stress due to the exploding number of journals, conferences, and grant applications [5]. In addition, self-publication tools on the internet, such as blogs and wikis, are allowing many scholars to disseminate their research results and products much faster and more directly than the traditional peer review-based publication system allows. Well-established scholars, in particular, might be less reliant on peer reviewed publication venues when releasing research results [3]. From a sociological perspective, peer review supports scholarly communication and knowledge production in various ways, but is also a heavily normative process, with many assumptions and expectations that may not be met in reality [1].

Adding research data into the publication and peer review queues will only increase the stress on the scholarly communication system. Do data sets have to be peer reviewed in the same way as other traditional scholarly products? Data quality control processes are widely studied and implemented within scientific research settings, and within research data archives. Examining those well-established processes might shed light on how peer review processes for data might be reconceptualized and reconfigured.

3. MAPPING PEER REVIEW TO DATA ARCHIVING PROCESSES

How does the traditional process of peer review apply to data sets? The National Center for Atmospheric Research (NCAR) is a federally-funded research and development center. Within NCAR, a number of data management teams manage, archive, and provide public access to data sets of various kinds. NCAR data management teams perform various kinds of quality assessment and review of data sets prior to making them publicly available. How do notions of peer review relate to the types of data review already in place within data archives like those at NCAR? Certain data set characteristics and management/archiving processes challenge the traditional peer review processes. For example, who is qualified to review data sets? Within NCAR data management teams, scientists and software engineers work together to conduct quality control checks. But what formal and informal documentation would be necessary to allow someone outside of their research and data management teams to review a data set?

Another prominent challenge to any data review process is that data sets are often not published as singular items. Within NCAR data management and archiving processes, data sets are often updated, corrected, and augmented, both before and after they are officially posted on a public web site. From a peer review perspective, what data set review can be done pre-publication, and what must be done post-publication? Data users regularly find problems with data sets that data management teams' quality control processes do not find. This is not due to negligence on the part of data management teams, but is instead due to the fact that some data quality problems can only be found through intensive use.

Finally, what components of the data sets review processes can be automated, and what components will always require human expertise and evaluation? Automation can simplify the manual efforts required to perform routinized tasks, such as data reformatting and integrity checks [6], but some data quality decisions cannot be formalized. Data quality is often a situation-specific assessment, based on the data that are available, the research questions being asked, and the research processes being used to investigate those questions.

4. INSTITUTIONAL EMBEDDING

Ultimately, the push for peer review of data sets is tied to the ways that researchers are rewarded for their professional activities and products. Peer review is a cornerstone of promotion and tenure systems within universities and research organizations. Non-peer review products are often completely left out of scholarly output assessments. Thus, data publication and citation initiatives will be slow to grow until some equivalent of peer review is established for data sets. If scholars receive no direct rewards for producing and archiving high quality data sets, data management and archiving will always be a secondary task. Gaining a better understanding of what peer review might mean for data archives and data sets will help to identify ways that the traditional peer review process might be supplemented or changed to enable quality control assessments that are acceptable at institutional levels.

5. ACKNOWLEDGMENTS

Thanks to the UCAR/NCAR data citation working group, in particular Steve Worley, Mike Daniels, Steve Williams, and Gary Strand.

6. REFERENCES

[1] Bornmann, L. 2008. *Scientific Peer Review:* An Analysis of the Peer Review Process from the Perspective of Sociology of Science Theories. *Human Architecture: Journal of the Sociology of Self-Knowledge*, 2, 23-38. http://www.okcir.comwww.okcir.com/Articles%20VI%202/LutzBornmann-FM.pdf

[2] Callaghan, S., Lowry, R., & Walton, D. 2012. Data Citation and Publication by NERC's Environmental Data Centres. *Ariadne*, 68. http://www.ariadne.ac.uk/issue68/callaghan-et-al

[3] Ellison, G. 2011. Is peer review in decline? *Economic Inquiry*, 49, 3, 635-657. http://dx.doi.org/10.1111/j.1465-7295.2010.00261.x

[4] Killeen, T. (2012). *Dear Colleague Letter - Data Citation*. NSF 12-058, March 29, 2012. http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click

[5] Miller, G. & Couzin, J. 2007. Peer Review Under Stress. *Science,* 316, 5823, 358-359. http://dx.doi.org/10.1126/science.316.5823.358

[6] Ni, K., Ramanathan, N., Chehade, M.N., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M., & Srivastava, M. 2009. Sensor network data fault types. *ACM Transactions on Sensor Networks,* 5, 3, 1-29. http://doi.acm.org/10.1145/1525856.1525863

[7] Starr, J. & Gastl, A. 2011. isCitedBy: A Metadata Scheme for DataCite. *D-Lib Magazine*, 17, 1/2. http://www.dlib.org/dlib/january11/starr/01starr.html

[8] Thomson Reuters. 2012. *Thomson Reuters Unveils Data Citation Index for Discovering Global Data Sets.* Thomson Reuters Press Release, June 22, 2012. http://thomsonreuters.com/content/press_room/science/686112

Position Papers: 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 will:

- 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)

Position Paper on Tools for Effective and Painless Data Curation

Leslie Johnston
Library of Congress
101 Independence Ave, S.E.
Washington, DC 20540-1340
202-707-2801
lesliej@loc.gov

ABSTRACT

This paper describes issues and tools potentially used in the selection, metadata extraction, management and preservation of data.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – Collection, Dissemination, Standards

General Terms

Metadata, Automation, Datasets.

Keywords

Datasets, Data Selection, Data Curation, Digital Curation, Metadata, Metadata Extraction, Digital Preservation

1. INTRODUCTION

Data quality describes data that is authentic, complete, well-managed, discoverable, appropriately available, and usable. What kinds of tools and techniques exist or are required to ensure that creators and curators address data quality in the collections we steward? To address these questions, this position paper will:

- Identify any extant tools and techniques for selecting data sets for curation and make recommendations where there are gaps.
- Identify extant tools and techniques for automatic metadata generation and annotation.
- Make recommendations for the management of data.

The goals for these tools are not just support for the general concept of data quality, but the need for ease of management (a reduction of effort in ingest as well as making data easier to sustain and preserve) and ease of discovery for use and reuse of the data being stewarded.

2. THE SELECTION OF DATA FOR CURATION

Data selection should not follow different curatorial collection development criteria than analog collections. All collections should be selected on the basis of fit for the scope of collections of the institution, which can include topical coverage and documenting the output of the organization, its faculty, and staff.

Selection for data quality by humans must be similar to the processes used in selecting, for example, serial titles: the fit for the organization's collection development scope and quality indicators, such as the reputation of the authors and the research institutions they represent.

As to automating the selection of data against data quality metrics, this may occur in either the selection or the ingest digital lifecycle stages. The data should be compared against an appropriate format-based and/or discipline-based profile to determine both the authenticity and completeness of the data and the institution's ability to steward the data. This assessment may be human, used a documented framework, or automated, as would take place during a formalized ingest workflow.

The important factors in automating such selection assessment are in validating against local sustainability factors, including:

- Completeness of the data upon transfer
- File formats to create and use data in common use in the relevant community
- File formats that are considered sustainable and preservable by the stewarding organization
- Data files that pass validation
- Accompanied by metadata or metadata can be easily created

3. METADATA GENERATION AND ANNOTATION

Metadata extraction is the process of automatically pulling (extracting) metadata from a resource's content. Resource content is mined to produce structured ("labeled") metadata for object representation.

The key to identifying potential tools is the understanding that the goal is not just the curation and management of data, but the discovery and reuse of data. The data may be managed and preserved, but if it cannot be discovered and used by its community, those management and preservation efforts are for naught.

As to user-generated content, experiments such as those by the Library of Congress in Flickr[1] have let the data curators tap into

[1] http://www.flickr.com/photos/library_of_congress/

the expertise in the communities of interest and elicit user-centric, relevant terms that have the potential to increase retrieval and provide a richer experience for the users of the collections.

3.1. Categories of Metadata

There are a number of categories of metadata which should be addressed to serve management, preservation, and discovery.

Names

Identification of people and organizations must be unambiguous and will hopefully support changes of name during the life of the person or organization.

Subjects

Both keywords and classification as well as date coverage. Classification includes not only formal, recognized classification schemes, but also informal, personal, community and emerging classifications.

Geospatial

It is important to record geospatial metadata when describing data gathered in specific areas. Geospatial information can be expressed as coordinates (in many different coordinate systems), place names, regions, postcodes, and discovery often depends on being able to translate from one form to another.

Bibliographic

The commonly understood metadata elements that used to represent document-like objects, such as: resource language; type of document ; title ; author(s); affiliation or contact details of author(s); date of publication; page count and page numbers; document index, table of contents; sources cited/referenced within document/bibliography; theme; related documents. As these are many of the most commonly required metadata elements they offer enormous potential for reducing the effort of manual metadata creation.

Factual

Readily extractable factual metadata includes: authoritative file type; date/time of deposit; depositing user; unique identifiers; file size; dimensions (of images); location (such as from GPS devices). Much of this information can be extracted unambiguously at the time of ingest. This encompasses the metadata that features most prominently in technical, structural, and preservation metadata.

User-Supplied

Non-validated metadata, including annotations and identification of entities, places, and dates, which is supplied by the community of users. For cultural institutions, annotations can contribute valuable metadata for search and retrieval, which in turn can increase the visibility of the data they expose via their digital library systems.

3.2 Functional Tool Requirements

The following functional requirements apply to an overall environment supporting metadata creation for the management of quality data:

- The extraction and creation of descriptive metadata from file content and headers, including entity recognition.
- The extraction and creation of technical and preservation metadata from file headers.
- Transcription form text-based and multimedia file formats.
- Read from community standard file types.
- The ability to set up different profiles for different file types in a tool, and integrate various tools as appropriate for a variety of file types and workflows.
- Support for diacritics and special characters.
- Generate and export metadata in standard formats.
- Support automatic and semi-automatic quality control routines, error checking, and validation of encoding against schemas by processes and humans.
- Integrate access to name authority and geographic files or web services for the lookup, matching and disambiguation of entity names.
- System should support automatic linking of metadata records, including referencing and cross-referencing between related items.
- Enable users to submit transcriptions an annotations which can be collected and imported into data management tools.
- Support user/organizational customizability and flexibility.

3.2 Census of Applicable Tools

Given the appearance and disappearance of metadata tools, this is a non-comprehensive list of potentially useful metadata extraction tools.

3.2.1 File Formats, Properties, and Forensics

- DROID http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
- FITS http://code.google.com/p/fits/
- Fiwalk http://afflib.org/software/fiwalk
- GNU Libextractor http://www.gnu.org/software/libextractor/
- JHOVE http://sourceforge.net/projects/jhove/index.html
- JHOVE2 http://www.jhove2.org/
- Sleuthkit http://www.sleuthkit.org/sleuthkit/
- wvWare http://wvware.sourceforge.net/
- List of freely available forensics tool http://forensiccontrol.com/resources/free-software/

3.2.2 Entity Identification and Extraction

- AlchemyAPI Entity Extraction http://www.alchemyapi.com/api/entity/
- AlchemyAPI Entity Extraction http://www.alchemyapi.com/api/entity/
- CDL Date Normalization Utility http://www.cdlib.org/services/dsc/projects/docs/datenorm_documentation.pdf
- Names Project http://names.mimas.ac.uk/
- OpenCalias Entity Extraction http://www.opencalais.com/
- Stanford POS http://nlp.stanford.edu/software/tagger.shtml

- Stanford Named Entity Recognition http://nlp.stanford.edu/software/CRF-NER.shtm
- Textpresso http://www.textpresso.org/
- Unlock http://unlock.edina.ac.uk/home/

3.2.3 Optical Character Recognition and Automated Transcription

- ABBYY Fine Reader http://finereader.abbyy.com/
- ELAN http://tla.mpi.nl/tools/tla-tools/elan/
- EXMARaLDA http://www.exmaralda.org/en_index.html
- Tesseract http://code.google.com/p/tesseract-ocr/
- OCRopus https://code.google.com/p/ocropus/
- XTrans http://www.ldc.upenn.edu/tools/XTrans/
- Yuma.min.js http://yuma-js.github.com/

3.2.4 Integrated Toolkits

- ADAM http://www3.imperial.ac.uk/bioinfsupport/resources/software/adam
- Archivists' Toolkit http://www.archiviststoolkit.org/
- BitCurator http://www.bitcurator.net/
- Curators Workbench https://github.com/UNC-Libraries/Curators-Workbench
- Google Refine (aka Freebase Gridworks) http://code.google.com/p/google-refine/
- Helping Interdisciplinary Vocabulary Engineering (HIVE) http://ils.unc.edu/mrc/hive/
- ICA-AtoM https://www.ica-atom.org/
- IngestList http://ingestlist.sf.net
- MetaGeta http://code.google.com/p/metageta/
- NARA File Analyzer https://github.com/usnationalarchives/File-Analyzer
- National Library of New Zealand Metadata Extraction Tool http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool
- Nesstar http://www.nesstar.com/

3.2.5 Services to Train Tools, Validate Controlled Terminology, Normalize, and Enrich Metadata

- DBpedia http://dbpedia.org/About
- Freebase http://www.freebase.com/
- GeoNames http://www.geonames.org/
- Library of Congress Vocabularies http://id.loc.gov
- Virtual International Authority File http://viaf.org/
- Wikipedia http://www.wikipedia.org

3.2.6 User Supplied Annotation, Transcription, and Metadata

- FromThePage http://beta.fromthepage.com/
- Proofread Page https://www.mediawiki.org/wiki/Extension:Proofread_Page
- Scripto http://scripto.org/

3.3 Metrics to Test Tools:

There are two approaches to evaluating metadata assignment: automatic evaluation by automatic computer program, and human evaluation by expert catalogers. An automatic evaluation requires a set of documents where expert-assigned metadata values are known, and the degree of similarity between the automatically-assigned values and the expert-assigned values is measured. A human evaluation involves having a group of expert catalogers rate the appropriateness of the metadata assigned.

There are intrinsic metrics to measure the effectiveness of metadata extraction tools: measure the quality of auto-generated metadata that the tool.

- Completeness
- Exact Match Accuracy
- Precision; Recall
- False positives
- Error rate
- Summary length
- Summary Coherence
- Summary Informativeness
- Content Word Precision and Content Word Recall
- Content Similarity
- Likert Scale.

There are also extrinsic metrics that measure the efficiency and acceptability of the auto-generated metadata in relation to expert manual metadata assignement:

- Learning Accuracy
- Cost-based evaluation
- Time saved.

There is a greater potential accuracy in the extraction of technical and preservation metadata than for descriptive requiring intellectual discretion, such as *subject* and *description*, especially entity recognition. *Coverage* metadata, which is used for *temporal* or *spatial* subject-like metadata, have the lowest potential for accuracy due to the varied and potentially imprecise nature of recording date. *Rights metadata* is highly unlikely to be extractable.

4. DATA MANAGMENT

There are readily available tools for researchers to create Data Management Plans to meet funder requirements, such as DMPOnline[2] and the DMPTool[3]. How does an organization translate a researcher's hoped-for data sustainability goals into a feasible, actionable data management activity in its infrastructure?

The Digital Curation Centre has an exceptionally useful list of tools for assessing data management needs[4], but not so much for automated processes for management. What there are, however, are frameworks for auditing data collections and assessing data management needs, such as the Data Asset Framework[5] and the Keeping Research Data Safe (KRDS) Benefits Analysis Toolkit[6].

The data management activities to maintain data quality and sustainability that can be automated are those around the auditing

[2] https://dmponline.dcc.ac.uk/
[3] https://dmp.cdlib.org/
[4] http://www.dcc.ac.uk/resources/external/tools-services
[5] http://www.dcc.ac.uk/resources/repository-audit-and-assessment/data-asset-framework
[6] http://www.beagrie.com/krds.php

of files on storage to confirm continued fixity, auditing of file formats against format action plans, and the batch migration of files to new formats as appropriate. These are activities are performed in the context of an institution's data management and preservation infrastructure and tools. In this context, a stewardship institution's assessment and selection criteria for data management tools and infrastructure should include automation of these functions.

5. CONCLUSIONS

No selection, metadata extraction, or data management activities in support of data quality can be fully automated, nor are there tools to support all these activities, especially selection. Selection must involve, to some extent, personal and subjective review. Metadata extraction can be run through automatic processes, but must be followed by human review to evaluate and enrich the results unless minimal automated metadata is acceptable. User-generated metadata is, by definition, supplied by people. Data sustainability management in support of data quality can be automated, but not all data management efforts can be automated.

REFERENCES

Dalton, J., Can Structured Metadata Play Nice with Tagging Systems? Parsing New Meanings from Classification-Based Descriptions on Flickr Commons. In J. Trant and D. Bearman (eds). *Museums and the Web 2010: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2010. Consulted August 20, 2012. http://www.archimuse.com/mw2010/papers/dalton/dalton.html

Flynn, Paul, Li Zhou, Kurt Maly, Steven Zeil, and Mohammad Zubair. (2007). Automated Template-Based Metadata Extraction Architecture. In D.H.-L. Goh et al. (Eds.): *ICADL 2007, LNCS 4822*, pp. 327–336, 2007.

Greenberg, Jane. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, 6(4), 59–82.

Greenberg, Jane, Kristina Spurgin, and Abe Crystal. (February 17, 2005). *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project*. http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Intrallect Ltd. (24 August 2009). *Automatic Metadata Generation: Use Cases and Tools/Priorities. Synthesis Report on Automated Metadata Generation and its Uses*. http://www.intrallect.com/index.php/intrallect/content/download/960/4029/file/synthesis_report.pdf.

Intrallect Ltd. (24 August 2009). *Automatic Metadata Generation: Use Cases and Tools/Priorities. Guidance on different automated metadata generation approaches for service providers in HE*. http://www.intrallect.com/index.php/intrallect/content/download/961/4032/file/guidance_report.pdf.

Kern, Roman, Jack Kris, Maya Hristakeva, and Michael Granitzer. (2012). TeamBeam — Meta-Data Extraction from Scientific Literature. *Dlib Magazine*, Volume 18, Number 7/8. http://www.dlib.org/dlib/july12/kern/07kern.html

Miglitz, James. (2008). *Automated Metadata Extraction*. Masters Thesis, Naval Postgraduate School. http://cisr.nps.edu/downloads/theses/08thesis_migletz.pdf

Patton, M., Reynolds, D., Choudhury, G. S., & DiLauro, T. (2004). Toward a metadata generation framework: A case study at the John Hopkins university. *D-Lib Magazine*, *10*(11). http://www.dlib.org/dlib/november04/choudhury/11choudhury.html.

G. W. Paynter. Developing practical automatic metadata assignment and evaluation tools for internet resources. In M. Marlino, T. Sumner, and F. M. S. III, editors, ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CA, USA, June 7-11, 2005, Proceedings, pages 291–300. ACM, 2005.

Polfreman, Malcom and Shrija Rajbhandari. (29 October 2008). *MetaTools - Investigating Metadata Generation Tools Final report*. http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf

Springer, M., B. Dulabahn, P. Michel, B. Natanson, D. Reser, D. Woodward, & H. Zinkham (2008). For the Common Good: The Library of Congress Flickr Project. Last Updated: December 09, 2008. http://www.loc.gov/rr/print/flickr_report_final.pdf.

Voss, J (2007). "Tagging, Folksonomy & Co - Renaissance of Manual Indexing?" In *ISI 2007 Proceedings*. Cologne, Germany: 10th international Symposium for Information Science. Last Updated: April 30, 2007. http://arxiv.org/abs/cs/0701072. D

Data Quality: The Need for Automated Support

Prasenjit Mitra
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
pmitra@psu.edu

C. Lee Giles
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
giles@ist.psu.edu

ABSTRACT

In this paper, we discuss issues related to data quality especially as they pertain to those in a digital library such as CiteSeerX. We also identify some functionalities that are desirable in tools that can be used to improve the quality of data and metadata in these digital libraries and propose such tools be developed.

Categories and Subject Descriptors

[**Information Systems**]: Information Systems Applications – Data Mining – *Data cleaning*, Digital Libraries & Archives.

General Terms

Algorithms, Management, Documentation, Design, Reliability.

Keywords

Data quality, Digital libraries, Tool Design.

1. INTRODUCTION

Digital library search engines are known for ubiquity, heavy use, and data quality issues. Google Scholar, an invaluable resource for many, has had many well-discussed errors such as author disambiguation, ghost authors, over counting of citations, etc [9]. Here we present the types of data errors that we find in CiteSeerX and other digital libraries.

1. Automatically extracted information does not have sufficient accuracy due to errors in parsers.
2. The same string refers to multiple real-world objects and differentiating among these real-world objects is important, aka, name disambiguation errors.
3. Insufficient context is present in one document or item.
4. Completeness or coverage is a problem. Given the vastness of the world-wide-web, we do not know where to get the documents. Given resource constraints, we cannot afford to crawl the whole web.
5. Unknown provenance of the data resulting in mis-interpretations of the data and missing metadata.
6. Insufficient documentation that results in data and metadata ambiguity and erroneous usage.

1.1 Information Extraction Problem

The CiteSeerX digital library [7] is constructed by automatically crawling the public web for publically available scientific and academic papers mostly in computer and information science. Papers harvested are then ingested and indexed using an open source automatic metadata extraction process and text indexer [5]. For data that is automatically generated, a significant amount of errors are due to the erroneous process of metadata extraction. For large digital libraries such as CiteSeerX, extracting such data manually is impossible and would significantly reduce the scale of the operation. For example, CiteSeerX extracts metadata such as Dublin Core from papers using a PDF to text extractor. Then, it detects tables [6] and figures from the papers. We have an automated parser that extracts author names and affiliations of the authors [4] plus finds and extracts the references [8]. If possible, the CiteSeerX parser tries to extract the venue of where the paper was published.

The parser is a source of a large number of problems. Papers do not have a fixed format and authors do not often follow any standard norms when they write the paper. Therefore, the parsers miss author names, erroneously mark as part of the author names one word from the next field such as institution, extract wrong venue information or year of publication information, erroneously miss part of the title especially for long titles, or extraction errors appear in words that are wrongly extracted in the titles.

1.2 Information Linking and Disambiguation

This problem has been referred to in the literature using many different names with object deduplication, record linkage, and name disambiguation the most common ones.

The earliest work on this problem started with attempting to identify the same records of individuals even when they have moved. The U.S. Census Bureau has long suffered from this problem. They proposed solutions based on attributes associated with the records to try to match and detect similar records. The problem with such methods was the fact that there was no key that could accurately link the data together.

Geo-coding is the process of assigning a latitude-longitude co-ordinate to a place name. In this area, the problem of geographical name disambiguation is a serious one. There are between 30 to 50 Springfield's in the U.S.A. One does not know which city is being mentioned and thus geo-coding, i.e., finding the latitude and longitude of the city and showing it on a map is difficult. The problem gets even worse in social media where people do not refer to the canonical names, abbreviate names, etc.

In CiteSeerX, the problem of name disambiguation arises because of different authors having the same name [3]. For example, David Johnson is a common name as is Wei Wang. Typically, the different David Johnson's or Wei Wang's can be differentiated and disambiguated using the venues of publication, co-authors, their institutions, etc. However, there are three Wei Wangs who all work in data mining and authors do change institutions. Disambiguating among them is a problem. If we cannot disambiguate, then calculations such as citation impact factors, and any analytics computed from using such "dirty" data would be erroneous.

1.3 Insufficient Context Problem

Oftentimes, the interpretation of data depends upon the context. Data quality depends upon how much of the context has been captured and is made available with the data. When the context is not available, the data may be erroneously interpreted. This is especially acute when analyzing social text streams where the context is implicit because authors assume that their friends and readers have read the previous posts by the same author or have a certain cultural context to interpret the posts correctly.

In the case of the CiteSeerX digital library, the problem of insufficient context appears when the metadata extraction results in erroneous or missing metadata. For example, the name of a co-author may be missing. Since the documents are collected from the web, sometimes the venue of the paper is missing. If the co-author information and the venue were being used for author disambiguation, the lack of proper context resulting from erroneous information extraction results in errors.

We believe that tools that augment and flesh out the context of a document should be developed. Such tools need to be scalable in order to be effectively run on large repositories and data sets and to augment the context of the document before further analysis is enabled. Then future analysis can be used the context to make decisions.

1.4 Completeness or Coverage Problems

The incompleteness of the web results in incomplete context associated with a document. For example, all the works of an author may not be available to a digital library. Suppose a researcher has started working in a new area. She may have published two papers --- one with her old co-authors and another alone. If we only have the paper she authored alone, we would not be able to detect that the author of the single author paper is the same as the author who had worked with the co-authors in another area. We may think that there are two different authors because the topics are different and there is no linking of co-authors. The affiliation could have been different if the work was started while the author was on sabbatical at a different university or laboratory.

On the metadata level, often due to quality of extraction errors, data related to some fields are missing. At times, these errors propagate in the system when the next metadata field is mistaken to be the missing field.

1.5 Provenance

The provenance of the data is important since the eventual quality of the data extracted can depend upon it. When there are errors, having the provenance permits a check back to verify and validate the data. Whether the data was bad at the source or errors crept in during the extraction and loading process can be determined.

1.6 Documentation

Another important indicator of the quality of the data is the quality of the documentation associated with it. The context of the data, its type and format information, and its semantics need to be documented and the high-quality metadata needs to be maintained so that we can interpret the data accurately when necessary. Documentation is very important so as to avoid semantic errors. When fully documented, the reader understands the semantics of the data accurately.

Automatic tools that check whether adequate and accurate documentation is associated with the data and prompt the user for documentation and where desired prevents data entry without proper documentation may be very useful in enhancing the quality of the data.

2. Desiderata for Data Quality Tools

The data warehousing community has an entire industry that deals with data quality tools. These data quality tools are used and the process of validating the data takes about 37.5% of the time needed to populate a data warehouse from constituent databases [1]. We need similar tools for curating data in digital libraries and web information repositories.

Data entered into data warehouses are checked for reasonableness, data type, etc [2]. Such checks using automated tools (necessary because of the large scale of the repositories) are not systematically performed in document repositories. For example, in CiteSeerX, added rules that check for the reasonableness of years of publication, the number of pages in the document, the number of authors, author names, institution names could considerably enhance the quality of the metadata. We posit that the design and implementation of such tools are of vital importance in order to improve the data quality in digital libraries.

2.1 Transformation Tools

Transformation tools that automatically identify differing formats and transform the data into canonical forms can be very useful. For example, in CiteSeerX, different documents may have their references formatted differently. Identifying the differences in the formats automatically and converting them properly will avoid errors in the future and is useful for enhancing the quality of the data. While transformations are necessary to handle the source data, what transformations were applied and the provenance information pointing out what data was converted and stored where must be recorded in order to enable proper interpretation and auditing of the entire data transformation processes.

2.2 Dealing with Uncertainty

While the goal of designing and deploying tools that extract data and metadata of high quality is to be applauded, it is quite likely that we will never reach the state where all our data is accurate. Thus, tools that detect these errors and if possible corrects such errors is desirable. For example, CiteSeerX may extract an author name erroneously. However, if we find several references to a paper with the same title and venue and publication year or several co-authors matching, with the aid of proper tools the system may detect that the extraction of the author name was erroneous. The tool can then correct the error while keeping track of the fact that this field was automatically corrected. Even after such corrections are made, wherever possible, we cannot expect the data and metadata to be perfect. Therefore, tools that use the data and the metadata need to be cognizant of the uncertainty inherent in the data, present a true picture to the users of the data and the metadata, and when combining one set of data with another be cognizant of the fact that when aggregated, the uncertainty related to the data can magnify the uncertainty associated with the data beyond acceptable levels in certain cases.

2.3 Annotation and The Wisdom of the Crowds

One of the important ways in which today's Internet systems improve the quality of the data is by using the wisdom of the crowds. For example, in CiteSeerX, we have observed that the rate at which users correct information, mostly in their own papers is more than the rate at which they provide missing data. We believe that the wisdom of the crowds can improve the data quality. Wikipedia is perhaps the best example of this phenomenon. Even though due to errors committed by editors of Wikipedia either on purpose or by mistake, the quality of the data suffers, the community mostly corrects these errors rather quickly at least in the case of pages that are visited the most often and are therefore arguably more valuable.

Examples of data correction may involve the end-users correcting the author names, the titles, the venues, and the years of publication of the data. End-users can assist us in identifying the captions and reference statements associated with tables, figures, and algorithms in these digital documents. They can clean up malicious vandalism that occurs when we allow end-users to correct or edit some pages despite our best efforts.

3. Conclusions/recommendations

Data quality issues exist in today's digital libraries. In order to improve the quality of the data in the libraries and reduce the errors in analyses that depend upon the data such as bibliometrics, we need automated and semi-automated tools that are scalable yet can detect and correct errors resulting in poor data quality. Such tools should work in a variety of systems and, so that they are readily available, be open source.

4. Acknowledgements

The authors acknowledge partial support from NSF CISE (under Grant Nos. 0958143 and 0845487) and Chemistry (under Grant No. 0535656), and Dow Chemical plus useful discussions with co-panelists.

5. REFERENCES

[1] Atre, S. 1998. Rules for Data Cleansing. *ComputerWorld*.

[2] Neely M. P. 1998. Data Quality Tools for Data Warehousing – A Small Sample Survey. Technical Report. Centre for Technology in Government. Available at: http://www.ctg.albany.edu/publications/reports/data_quality_tools.

[3] Treeratpituk, P., Giles, C.L. 2009. Disambiguating authors in academic publications using random forests. In *Proceedings of the 2009 Joint International Conference on Digital Libraries*: 39-48.

[4] Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. 2003. Automatic Document Metadata Extraction using Support Vector Machines. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*: 37-48.

[5] Teregowda, P.B., Councill, I., Fernández R., J.P., Kasbha, M., Zheng, S., Giles. C.L. 2010. SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web. In *Proceedings of the 1st USENIX Conference on Web Application Development*.

[6] Liu, L., Bai, K., Mitra, P., Giles, C.L. 2007. TableSeer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2007),* 91-100.

[7] Giles, C.L., Bollacker, K., Lawrence, S. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of DL'98 Digital Libraries, The 3rd ACM Conference on Digital Libraries, 89-98*.

[8] Isaac G. Councill, C. Lee Giles, Min-Yen Kan. 2008. ParsCit: an Open-source CRF Reference String Parsing Package. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*.

[9] Watters, A. 2011. Problems with Google Scholar Citations. *Inside Higher Ed*.

Automating Data Curation Processes

Reagan W. Moore
University of North Carolina at Chapel Hill
216 Manning Hall
Chapel Hill, NC 27599-3360
01 919 962 9548
rwmoore@renci.org

ABSTRACT

Scientific data quality can be abstracted as assertions about the properties of a collection of data sets. In this paper, standard properties are defined for data sets, along with the policies and procedures that are needed to enforce the properties. The assertions about the collection are verified through periodic assessments, which are also implemented as policies and procedures. Data quality curation can then be defined as the set of policies and procedures that verify the scientific data quality assertions. The assertions made by the creators of a collection are differentiated from the assertions about data quality needed by the users of the data. The transformation of data into a useable form requires well-defined procedures that need to be considered as part of the data curation process. The automated application of both digital curation and data transformation procedures is essential for the management of large scientific data collections.

Categories and Subject Descriptors

D.4.7 [**Operating Systems**]: Organization and Design – *distributed systems*

General Terms

Management, Design, Verification.

Keywords

Policy-based data management.

1. Data quality

Scientific data quality is dependent on the specification of a scientific research context. The creators of a scientific data set are typically driven by a research question, and choose quality criteria that are necessary for exploration of a research issue. The criteria may include properties that each data set must possess (such as physical units), or properties that are related to the entire collection (such as completeness and coverage). The properties can be turned into assertions that the data set creators make about their collection. Scientific data quality is quantified by the collection creators by explicitly verifying compliance with the desired properties.

The types of properties that are associated with scientific data sets can be loosely categorized as:

- Data format (e.g. HDF5, NetCDF, FITS, …)
- Coordinate system (spatial and temporal locations)
- Geometry (rectilinear, spherical, flux-based, …)
- Physical variables (density, temperature, pressure)
- Physical units (cgs, mks, …)
- Accuracy (number of significant digits)
- Provenance (generation steps, calibration steps)
- Physical approximations (incompressible, adiabatic, …)
- Semantics (domain knowledge for term relationships)

Additional properties can be derived from these categories. Thus the relevant time period may be defined in the temporal coordinate, and allowed transformations may be implicit in the type of variables and physical approximations that were made in creating the data collection. The additional properties may be evaluated by applying procedures that generate the desired information, which in turn can be applied to the data sets as metadata.

Data curation corresponds to identifying the properties claimed by the data set creators, verifying that the desired properties are consistently present throughout the data set, logging any discrepancies, and assembling an archival information package. Each desired property requires the evaluation of knowledge to determine its presence, and the creation of information that is associated with each data set as metadata. By examining the metadata, a user of the collection can determine whether the properties are present that are needed for the user's research initiative. Data quality from the user's perspective is determined by compliance with the properties that are needed when incorporating a data set into the user's analysis environment or data collection.

Data quality is a mapping from assertions that the creators of a collection make about their data sets, to requirements by a user for the appropriateness of the data sets for use in their own research. Both persons may have equally valid but incommensurate criteria for data quality.

2. Data, Information, and Knowledge

Data curation can be thought of as an active process that requires assessment of the information and knowledge context associated with a collection. The OAIS model uses the terms "representation information" and "knowledge community" to express the requirement that a targeted community be able to understand and manipulate a collection based on the representation information. In order to automate the generation of representation information,

Curating for Quality Workshop, September 11-12, Arlington, VA.

we need a definition of representation information that is computer actionable. We define:

- Data - consists of bits (zeros and ones)
- Information - consists of labels applied to data
- Knowledge - defines relationships between labels
- Wisdom - defines when and where knowledge relationships should be applied (relationship on relationships)

If we have a set of computer actionable processes that apply representation information, we can automate data curation actions. Since scientific data collections may be comprised of hundreds of millions of files and contain petabytes of data, automation is essential for building a viable preservation environment.

Policy-based data management systems provide computer actionable mechanisms for defining information, applying knowledge, and governing policy execution.

- Information is treated as labels that are applied to data sets as metadata. Each data set may have both system defined and user defined metadata attributes that are persistently maintained. System metadata consists of pieces of information that are generated when processes are applied to data sets. An example is a process that creates a replica. The location of the replica is system metadata that is associated with the data set.

- Knowledge is treated as procedures that evaluate relationships. While information is treated as static metadata that is maintained persistently, knowledge is treated as an active process that involves the execution of a procedural workflow. To simplify creation of knowledge procedures, basic functions called micro-services are provided that encapsulate well-defined actions. The micro-services can be chained together into a workflow that is executed whenever the associated knowledge is required.

- Wisdom is applied through policy enforcement points that determine when and where knowledge relationships should be evaluated. Each external action is trapped by a set of policy enforcement points within the data grid middleware. At each policy enforcement point, the data grid checks whether a policy should be applied. The policy enforcement points can control what is done before an action is executed, can control the action itself, and can control what is done after an action takes place. A simple example is the control of what happens when a file is ingested into a collection. The data grid middleware may transform the data set to an acceptable archival format, extract provenance metadata, generate a checksum, and replicate the data set.

This defines the minimum system components that are needed to automate curation processes. Fortunately, policy-based data management systems implement the above mechanisms for managing information, generating knowledge, and applying wisdom.

3. POLICY-BASED DATA MANAGEMENT

The integrated Rule Oriented Data System (iRODS) is used to build data curation environments [1]. The system is sufficiently generic that the iRODS middleware is used to implement all stages of the data life cycle. This approach to data curation is based on the following principles:

- The **purpose** for creating the collection determines the **properties** that should be maintained including data quality.
- The **properties** of the collection determine the **policies** that should be enforced.
- The **policies** control the execution of **procedures** through computer actionable rules.
- The **procedures** apply the required knowledge relationships and generate **state information** through computer executable workflows.
- The **state information** (metadata) is saved persistently.
- **Assessment criter**ia can be evaluated through periodic execution of policies that query the **state information** and verify that re-execution of the procedures generates the same result.

This provides an end-to-end system that enforces the required curation policies, persistently manages the representation information, and enables validation of data quality assessments.

The iRODS data grid provides representation information about the preservation environment through the set of policies and procedures that are applied. This representation information quantifies the data curation policies. The system can be queried to discover which policies are being applied. The procedures can be re-run to verify that the system is maintaining the quality metrics correctly.

A simple example is an integrity criterion that asserts that the data sets have not been corrupted. One approach is to save a checksum that is formed by manipulating all of the bits in the file. If any of the bits have been corrupted, the checksum will change. The original checksum for the file (created at the time of ingestion) can be saved as persistent state information. The policy that governs the creation of the checksum can be re-run at any point in the future, generating a new checksum. The original value and the most recently created value can be compared to verify the integrity of the file.

3.1 Knowledge Scale

An important question is the whether it is feasible to quantify information, knowledge, and wisdom as metadata, policies/procedures, and policy enforcement points. Will the number of entities remain bounded, or will the amount of system representation information become larger than the collection size?

Applications of the iRODS data grid typically maintain:

- About 220 attributes associated with files, users, collections, storage systems, policies, and procedures. Note that system level metadata is needed not only for files, but also for the preservation environment itself.

- About 74 policy enforcement points for controlling the execution of policies. Policy enforcement may be imposed before an action is executed, to control an action, and after an action is executed.

- About 250 micro-services for implementing procedures [2]. Examples include micro-services to query the metadata catalog, loop over the result set, read a file, create a checksum, store new descriptive metadata attributes, replicate a file, etc.

- About 20 rules that enforce collection properties such as data quality. This number typically corresponds to about 20 properties that are desired for a collection. Note that the system is capable of supporting thousands of rules, but most applications choose a small subset. Examples might be rules that maintain integrity (replicate a file, validate checksums), maintain authenticity (manage provenance information), track chain of custody (audit trails), and track original arrangement (physical file path).

When enforcing assertions about data quality, a data management system needs a computer actionable rule that controls the extraction of the desired property, and a computer executable workflow that applies the required relationships.

4. CURATION APPLICATIONS

We can conduct a thought experiment to decide how we can automate data quality assessments about a collection, based on the properties of scientific data collections listed in Section 1. An immediate question is whether a specific data quality property requires the extraction of metadata from within each data set, or whether the information must be provided through an external mechanism. A related question is whether the properties will be uniform throughout the data collection, or whether some properties will be unique to a sub-set of the files. Another possibility is that the desired data quality property has to be determined through examination of the actual data, such as detection of missing values. The types of processing that are applied to verify data quality will vary dramatically based on the type of data, desired properties for a collection, and purpose behind the generation of the data set.

The following examples are intended to demonstrate that data quality inherently is dependent upon the execution of procedures that verify the presence of desired properties either within each data set or within a collection. Data quality curation can be defined in terms of the data quality procedures that are executed by either the creator of a collection or by users of a collection. These procedures may be quite different and result in different definitions of the quality of a data collection.

Quality assessment for **data format** tends to be evaluated for each data set within a collection. Each data type has a standard structure that can be verified. The HDF5, NetCDF and FITS data formats package metadata with the data. The metadata can be extracted and registered as queriable attributes within a collection. Given a standard structure for the data format, micro-services can be created that evaluate the structure, verify the structural components are consistent with the specification, and ensure that the data can be read from the structure in the future. An expectation is that the micro-service that analyzes and manipulates the data structure will be executable on future architectures. A quality assessment procedure that is executed today should also be executable in the future on future operating systems. Data grids provide this capability through virtualization of standard I/O operations.

Quality assessment for the **coordinate system** is typically information that is evaluated for each data set. For gridded data, the spatial and temporal location may be explicitly stored, or may be inferred from spatial dimension arrays. An example of the importance of correctly assigning the coordinate system occurs when satellite data is geo-registered. The quality of the data depends upon the ability to correlate a pixel in a satellite image with a point on the ground. The algorithm that does the correlation is an essential component of a data quality assessment that may need to be reapplied in the future. This particular geo-referencing procedure is typically done by the creator of the data collection.

The **geometry** associated with the coordinate system is rarely explicitly captured, and typically is provided as external metadata. In plasma physics, experimental data for the poloidal flux within toroidal plasma devices is typically mapped to a flux-based non-orthogonal curvilinear coordinate system. The resulting coordinate system is then used to evaluate the stability of the configuration to magneto-hydrodynamic instabilities. The generation of the flux-based coordinate system requires the application of an algorithm. The data quality of the resulting interpretation of the plasma stability is strongly tied to the accuracy of the toroidal geometry representation. In this case a procedure that is applied by a user determines the data quality.

Scientific data sets are generated with well-defined **physical variables**. The set of variables desired by a user may require combinations of the variables present within the data set. An example of a system that extracts data from a data set based on physical variables is the OpenDAP and THREDDS environment. It is possible to extract physical variables from a data set, without retrieving the entire file. The quality of the physical variables depends more on the transformations that may be needed to convert to desired quantities. Thus the conversion from velocity to vorticity depends on how well the conversion routine approximates the curl operation, which in turn depends upon the spatial resolution provided by the coordinate system and the number of spatial points needed to implement a curl operation. The transformation function accuracy is even more important when interpolating data for use in differential equations. In practice, data analyses can introduce numerical artifacts if the degree of the interpolation function is not commensurate with the solution order of the differential equation. In this case, data quality requires self-consistent treatment of both data and analyses by the user of the data collection.

The quality of the **physical units** depends mainly on the consistency across the data collection. If some variables are in feet/pound/second units and some are in meter/kilogram/second units, the data will easily be misinterpreted and may lead to incorrect analyses. An example of poor physical units quality was the crash of a Mars rover.

The quality of the **measurement accuracy** (number of significant digits) is important for determining the allowed transformations. The data accuracy may be so poor that the desired physical effect cannot be separated from noise in the data. However, averages of the data may be sufficient to track changes over long time periods, or to track effects that appear from superposition of many data sets. An example is the association of quasars with galactic centers, by superimposing thousands of quasar images. In this case, the users of a data collection could generate meaningful research results even though each individual data set lacked the required measurement accuracy.

The **provenance** of the data needs to include descriptions of all processing steps that were applied to the data. The standard example is the processing of satellite data by NASA. The data have to be calibrated, turned into physical variables from raw sensor data, and then projected onto a coordinate system. Each processing step requires the application of a procedure that significantly transforms the data. Assertions about data quality are then driven by the accuracy of the transformations, as well as by the original accuracy and resolution of the raw data. In this case, the quality assessments are done by the creators of the data collection. However, if a calibration is revised, the data quality becomes highly suspect and the transformations must be re-done for new assertions about data quality.

Transformations applied to data also may depend upon **physical approximations** that are used to simplify the analysis. The physical approximations may be associated with type of physical flow (compressible or incompressible), type of equation of state, type of assumed particle distribution functions, etc. For a consistently derived data set, similar physical approximations need to be applied across all transformations performed upon the data.

A related issue is the set of physical constraints that are enforced when the data are manipulated. Do the numerical algorithms enforce physically conserved properties, such as energy, mass, and momentum? A simple example is the projection of telescope images to a standard coordinate system. To conserve the intensity, spherical trigonometric functions need to be used to project each pixel. This was done in the 2-micron All Sky Survey to generate a unifying mosaic of the night sky. If the algorithms had applied trigonometric functions, the intensity would have been blurred.

Another set of physical constraints is the set of assumptions for how missing data points will be handled. Is the missing data marked as missing, or are interpolation functions used to approximate the missing data? A standard example is the generation of a uniform world weather model that incorporates observational data. A numerical weather simulation is run forward in time based on the observations for 6 hours. The result is then compared with new observations. Forcing functions are derived such that running the simulation a second time will generate the actual observations seen at the end of the 6-hour run. This interpolates the weather onto a uniform grid in space and time, effectively interpolating across all missing data points. The interpolation accuracy depends upon the physical approximations that were made in the weather model. Each time the physical approximations are improved, a re-analysis is needed to generate a better interpolation onto the uniform grid in space and time. These analyses are typically performed by the creators of the data collection.

A second example is the analysis of radar data to generate precipitation estimates. Re-analyses are done based on improvements in the physical model for reflection of radar waves by water. The quality of the data set is driven by the quality of the physical model.

Semantics (and ontologies that describe how semantic terms are related) can lead to data quality control issues. Each domain defines a standard set of semantic terms that describe physical phenomena. Each research group tries to refine that description of domain knowledge to improve the understanding of the underlying physical world. The semantics used by a research group evolve to track their improved understanding of physical reality. Thus semantic terms, as used by a research group, may have nuances of meaning that are not known to the rest of the community. This results in different definitions of data quality, based on the understanding of the underlying physics.

5. SUMMARY

Data quality curation inherently requires the application of procedures to verify or create required data set properties. An analysis of the data quality of a collection requires a detailed understanding of the curation procedures. In policy-based data management systems, the data curation procedures can be preserved, and re-executed in the future to verify an assertion about the data quality that is made by the creators of the collection. However, the users of a data collection may have different required properties for data quality that in turn depend upon application of additional procedures. An assessment of data quality by the users of a collection may generate a different interpretation of the relevance of the data for their research project. Data quality assessments require a mapping between assertions made by the creators of a collection, and the collection properties needed by the users of a collection. This requires the ability to control application of procedures, sharing of procedures, re-execution of procedures, and preservation of procedures. Data quality curation can be mapped to the procedures that are used to verify assertions about a data collection.

6. ACKNOWLEDGMENTS

The topics presented here were supported by the NSF Office of Cyberinfrastructure OCI-0848296 grant, “NARA Transcontinental Persistent Archive Prototype”, (2008-2012), the NSF ITR 0427196 grant, “Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives” (2004-2007), the NSF SDCI 0721400 grant, "SDCI Data Improvement: Data Grids for Community Driven Applications” (2007-2010), the NSF OCI-1032732 grant, "SDCI Data Improvement: Improvement and Sustainability of iRODS Data Grid Software for Multi-Disciplinary Community Driven Application," (2010-2013), and the NSF Cooperative Agreement OCI-094084, “DataNet Federation Consortium”, (2011-2013).

7. REFERENCES

[1] Rajasekar, R., Wan, M., Moore, R., Schroeder, W., Chen, S.-Y., Gilbert, L., Hou, C.-Y., Lee, C., Marciano, R., Tooby, P., de Torcy, A., and Zhu, B.. 2010. *iRODS Primer: Integrated Rule-Oriented Data System*, Morgan & Claypool. DOI= 10.2200/S00233ED1V01Y200912ICR012.

[2] Ward, J., Wan, M., Schroeder, W., Rajasekar, A., de Torcy, A., Russell, T., Xu, H., Moore, R. 2011. *The integrated Rule-Oriented Data System (iRODS 3.0) Micro-service Workbook*, DICE Foundation, November 2011, ISBN: 9781466469129, Amazon.com

Data Quality for New Science: Process Curation, Curation Evaluation and Curation Capabilities

Andreas Rauber
Vienna University of Technology
Favoritenstrasse 9, 1040 Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

In order to fully support the potential of data-driven science, eScience, the 4th Paradigm, and other similar concepts, we face significant challenges in curating the data, ensuring its authenticity, accessibility, proper reusability and repurposing in different contexts. So far, the primary focus in these areas has been on documentation and preserving the actual data. This position paper argues for an approach focusing on the curation of the actual processes involved in the collection, pre-processing and use of data, capturing process contexts and the actual processes together with the data. We further present an approach on how to validate and measure conformance of a re-activation of any such process to ensure and prove authenticity and validity. Last, but not least, we argue in favor of a capability and maturity based view of data and process curation, rather than mere auditing and certification, and the establishment of supporting (IT-)processes.

General Terms

E-Science, Research Infrastructures, Process Preservation, Context Information, Evaluation Framework, Enterprise Architectures, Maturity Model

1. INTRODUCTION

Like all digital data, research data is exposed to threats of digital obsolescence, i.e. when the digital objects become unusable. This may occur on three different levels - the bit level, the logical level, and the semantic level. While a range of solutions and best practice experience exists for bit-level preservation, most of digital preservation research focuses on logical preservation, i.e. ensuring that the file formats that the information is provided in remains accessible by current software versions. For research data, this challenge in some aspects is both harder as well as easier than for many conventional objects: on the one hand, research data is frequently represented in some form of numeric representation that is both rather stable in terms of accessibility, with simpler format specifications, a clearer separation between data and functionality, i.e. no embedded code, and thus simpler transformation settings for data migration. On the other hand, research data preservation at the logical level is more complex, as in many cases both data formats as well as preservation requirements are rather unique to each data set, with characteristics of data sets ranging both from individual data sets with massive volumes of data items to myriads of rather small data sets, each with their own and very specific designated community. Yet, the most serious challenge to data curation arises at the semantic level, ensuring the authenticity and correct interpretability of data. Conventionally, this comprises capturing as much information about the data, its preprocessing and use as well as actions performed on the data during curation activities as possible in order to establish provenance and interpretability.

We claim, however, that several aspects related to data curation, specifically with a focus on ensuring its quality, are not receiving sufficient attention in current R&D. This paper summarizes some of our current considerations and areas of research focus with respect to data curation both at the Vienna University of Technology[1] as well as at Secure Business Austria[2], most prominently in the research projects SCAPE[3], TIMBUS[4], APARSEN[5] as well as some new initiatives on data curation and evaluation to be launched.

First, establishing context of data is focused strongly on documentation, i.e. documenting intention, data capture, and potential processing steps and many others. Yet, specifically with respect to data (pre-)processing, pure documentary approaches are probably not sufficient: as the processing modules and processes become more complex, the risk of either not fully documenting the process or of the process as implemented not perfectly following the intended process grows. As a result, erroneous pre-processing software, processing steps not obeyed due to misunderstanding or lack of diligence etc. may lead to artifacts being introduced into the data, or lead to incoherent results when trying to repeat experiments under identical conditions. We thus argue that capturing and curating the (pre-)processing processes

[1] http://www.ifs.tuwien.ac.at/dp

[2] http://www.sba-research.org/research/data-security-and-privacy/digital-preservation

[3] http://www.scape-project.eu

[4] http://timbusproject.net

[5] http://aparsen.digitalpreservation.eu/

is in many cases an integral part of data curation. It also enables re-running earlier experiments with new data under identical conditions. We thus are currently working on new approaches for process and process context capture, documentation, preservation and re-activation [10, 8, 9].

Second, once the processes are curated as part of the data, mechanisms, strong emphasis must be placed on establishing whether any re-activation of research data is actually faithful to the original with regards to a set of determined significant properties. We feel there is a lack of established mechanisms and frameworks, both at the data/process capture as well as re-activation phases, to determine whether all essential aspects offered by a new viewing application, after a transformation, or even when opening objects in an emulated environment. In fact, it can be shown that both migration as well as emulation approaches are rather identical in character, and need to be evaluated in very similar manners [5]. We thus are currently investigating more formal frameworks for documenting and verifying identity of digital objects on re-use with respect to established properties [3, 4].

Third, data curation requires the consistent application of well-defined processes in a highly repeatable, consistent, well-documented manner to ensure trustworthiness. While these may partially be handled by institutions whose primary focus is data curation, we see a shift in such operations occurring as part of other primary business operations. This will result in a shift from current thinking of operational data on the one hand vs. dedicated archival data holdings on the other to a merged operational data repository with integrated preservation capabilities. It will also require an integration of curation activities into standard (IT) operations. Thus, models and standards from data curation will need to be merged with concepts from IT Governance and Enterprise Architectures to allow a consistent view on curation activities as part of a institutions operations. Beyond audit and certification establishing conformance to specific requirements, capabilities and maturity models may offer a more flexible and realistic approach to establishing the competences and improving them, guiding investment and ensuring proper alignment with an institutions objective. We are thus reviewing ways to align the two worlds of IT Governance and Digital Curation, defining capabilities and establishing maturity models to allow for process evaluation and improvement. [1, 2].

The following sections review some of the initial concepts developed clarifying their scope and outlining future directions.

2. FROM DATA PRESERVATION TO PROCESS CURATION

While preserving the data is an essential first step for any sustainable research efforts, the data alone is often not sufficient for later analysis of how this data was obtained, pre-processed and transformed. Results of scientific experiments are often just the very last step of the whole process, and to be able to correctly interpret them by other parties or at a later point in time, also these processes need to be preserved. Thus, one needs to go beyond the classical concerns of Digital Preservation research, and consider more than the preservation of data. The following passages and example are adopted from [9] detailing our approach to process preservation on a simple example from the music retrieval domain.

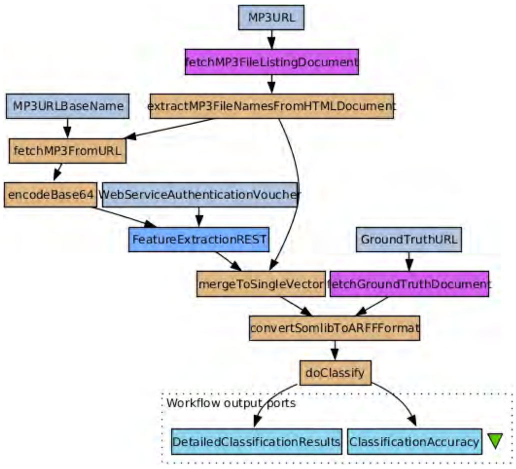


Figure 1: Musical genre classification, including fetching of data, modelled in the Taverna workflow engine

To move towards more sustainable E-Science processes, we recommend implementing them in workflow execution environments. For example, we are currently using is the Taverna workflow engine [11]. Taverna is a system designed specifically to execute scientific workflows. It allows scientists to combine services and infrastructure for modeling their workflows. Services can for example be remote web-services, invoked via WSDL or REST, or local services, in the form of pre-defined scripts (e.g. for encoding binaries via Base64), or user-defined scripts.

Implementing such a research workflow in a system like Taverna yields a complete and documented model of the experiment process – each process step is defined, as is the sequence (or parallelism) of the steps. Further, Taverna requires the researcher to explicitly specify the data that is input and output both of the whole process, as well as of each individual step. Thus, also parameter settings for specific software, such as the parameters for the classification model or feature extraction, become explicit, either in the form of process input data, or in the script code.

Figure 1 shows an example of a music classification experiment workflow modeled in the Taverna workflow engine. We notice input parameters to the process such as the URL of the MP3 contents and the ground truth, and also an authentication voucher which is needed to authorize the use of the feature extraction service. The latter is a bit of information that is likely to be forgotten frequently in descriptions of this process, as it is rather a technical requirement than an integral part of the scientific process transformations. However, it is essential for allowing re-execution of the process, and may help to identify potential licensing issues when wanting to preserve the process over longer periods of time, requiring specific digital preservation measures.

During an execution of the workflow, Taverna records so-

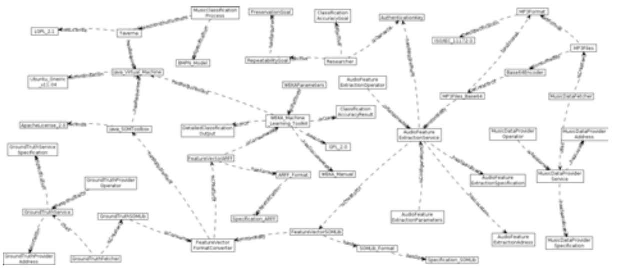

MusicClassification Process
PreservationGoal
Classification AccuracyGoal
AuthenticationKey
MP3Format
LGPL_2.1
Taverna
ISO/IEC_11172-3
MP3Files
BMPN_Model
RepeatabilityGoal
Researcher
Base64Encoder
Java_Virtual_Machine
Ubuntu_Gnenic v11.04
WEKAParameters
AudioFeature ExtractionOperator
MP3Files_Base64
MusicDataFetcher
Classification AccuracyResult
ApacheLicense_2.0
Java_SOMToolbox
DetailedClassification Output
WEKA_Machine Learning_Toolkit
AudioFeature ExtractionService
MusicDataProvider Operator
MusicDataProvider Address
GroundTruthService Specification
GPL_2.0
FeatureVectorARFF
ARFF_Format
WEKA_Manual
AudioFeature ExtractionSpecification
MusicDataProvider Service
GroundTruthProvider Operator
GroundTruthService
GroundTruthSOMLib
Specification_ARFF
AudioFeature ExtractionParameters
FeatureVectorSOMLib
AudioFeature ExtractionAdress
MusicDataProvider Specification
FeatureVector FormatConverter
SOMLib_Format
GroundTruthProvider Address
GroundTruthFetcher
Specification_SOMLib

Figure 2: Context Model of musical genre classification process

called *provenance data*, i.e. information about the creation of the objects, on the data transformation happening during the experiment. Taverna uses its proprietary *Janus* format, an extension on the Open-Provenance Model[12] that allows capturing more details. Such data is recorded for the input and output of each process step. It thus allows to trace the complete data flow from the beginning of the process until the end, thus enabling verification of the results obtained. This is essential for being able to verify system performance upon re-execution, specifically when any component of the process (such as underlying hardware, operating systems, software versions, etc.) have changed.

Curation of business or E-Science processes requires capturing the whole context of the process, including e.g. different or evolved enabling technologies, different system components on both hardware and software levels, dependencies on other computing systems and services operated by external providers, the data consumed and generated, and more high-level information such as the goals of the process, different stakeholders and parties. The context of information needed for preserving processes is considerably more complex than that of data objects, as it not only requires dealing with the structural properties of information, but also with the dynamic behavior of processes. Successful curation of an eScience process requires capturing sufficient detail of the process, as well as its context, to be able to re-run and verify the original behavior at a later stage, under changed and evolved conditions. We thus need to preserve the set of activities, processes and tools, which all together ensure continued access to the services and software which are necessary to reproduce the context within which information can be accessed, properly rendered and validated.

To address these challenges, we have devised a context model to systematically capture aspects of a process that are essential for its preservation and verification upon later re-execution. The model consists of approximately 240 elements, structured in around 25 major groups. It corresponds to some degree to the representation information network [7], modeling the relationships between an information object and its related objects, be it documentation of the object, constituent parts and other information required to interpret required to interpret the object. This is extended to understand the entire context within which a process, potentially including human actors, is executed, forming a graph of all constituent elements and, recursively, their representation information. The model is implemented in the form of an ontology, which on the one hand allows for the hierarchical categorization of aspects, and on the other hand shall enable reasoning, e.g. over the possibility of certain preservation actions for a specific process instance. While the model is very extensive, it should be noted that a number of aspects can be filled automatically – especially if institutions have well-defined and documented processes. Also, not all sections of the model are equally important for each type of process. Therefore, not every aspect has to be described at the finest level of granularity. Figure 2 gives an overview on the concrete instances and their relations identified as relevant aspects of the process context for the music classification process discussed above.

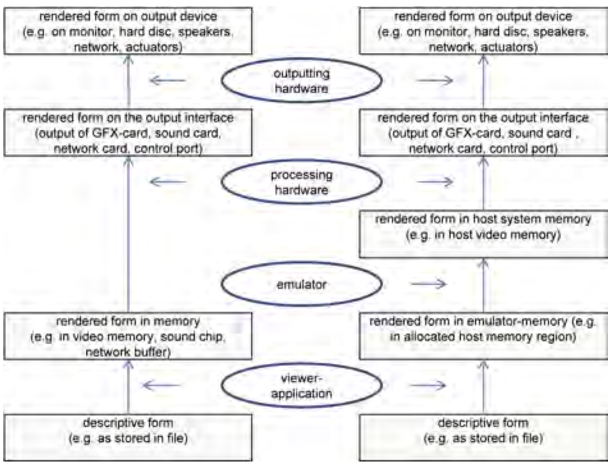
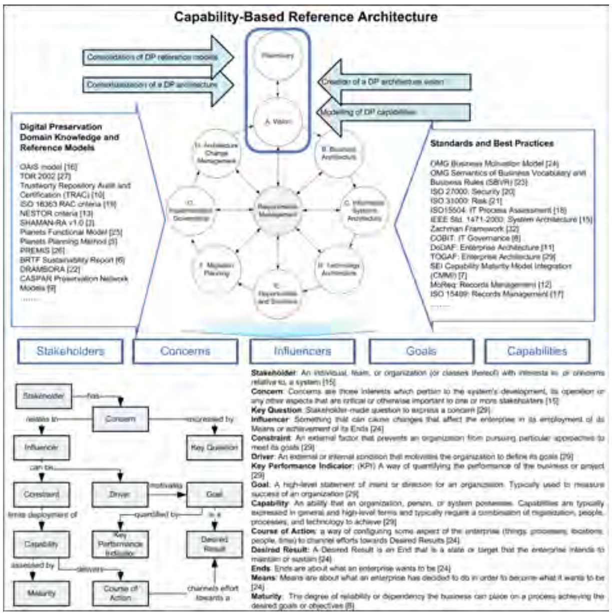

rendered form on output device (e.g. on monitor, hard disc, speakers, network, actuators)
rendered form on output device (e.g. on monitor, hard disc, speakers, network, actuators)
outputting hardware
rendered form on the output interface (output of GFX-card, sound card, network card, control port)
rendered form on the output interface (output of GFX-card, sound card , network card, control port)
processing hardware
rendered form in host system memory (e.g. in host video memory)
emulator
rendered form in memory (e.g. in video memory, sound chip, network buffer)
rendered form in emulator-memory (e.g. in allocated host memory region)
viewer-application
descriptive form (e.g. as stored in file)
descriptive form (e.g. as stored in file)

Figure 3: Different forms of a digital object in a system's memory. On the left the layers in an original system are shown, on the right the layers in the system hosting the emulator are shown.

3. EVALUATING PROCESS RE-ACTIVATION

A critical aspect of re-using digital information in new settings is its trustworthiness, especially its authenticity and faithful rendering (with rendering being any form of representation or execution and effect of a digital object, be it rendering on a screen, an acoustic output device, or state changes on ports, discs etc.). Establishing identity or faithfulness is more challenging than commonly assumed: current evaluation approaches frequently operate on the structural level, i.e. by analyzing the preservation of significant properties on the file format level in case of migration of objects. Yet, any digital object (file, process) is only perceived and can only be evaluated properly in a well-specified rendering environment within which faithfulness of performance need to be established. In emulation settings, this evaluation approach is more prominently present, yet few emulators support the requirements specific to preservation settings. we thus argue that, actually, migration, emulation and virtually all other approaches to logical/structural data preservation need to be evaluated in the same way, as they are virtually no different from each other as all need to be evaluated in a given rendering/performance environment. [5].

We also devise a framework for evaluating whether two versions of a digital object are equivalent [3]. Important steps in the this framework include a (1) description of the original environment, (2) the identification of external events influ-


Capability-Based Reference Architecture
Consolidation of DP reference models
Contextualisation of a DP architecture
Creation of a DP architecture vision
Modelling of DP capabilities
Preliminary
A. Vision
B. Business Architecture
C. Information Systems Architecture
D. Technology Architecture
E. Opportunities and Solutions
F. Migration Planning
G. Implementation Governance
H. Architecture Change Management
Requirements Management
Digital Preservation Domain Knowledge and Reference Models
OAIS model [16]
TDR 2002 [27]
Trustworty Repository Audit and Certification (TRAC) [10]
ISO 16363 RAC criteria [19]
NESTOR criteria [13]
SHAMAN-RA v1.0 [3]
Planets Functional Model [25]
Planets Planning Method [5]
PREMIS [26]
BRTF Sustainability Report [6]
DRAMBORA [22]
CASPAR Preservation Network Models [9]
....
Standards and Best Practices
OMG Business Motivation Model [24]
OMG Semantics of Business Vocabulary and Business Rules (SBVR) [23]
ISO 27000: Security [20]
ISO 31000: Risk [21]
ISO15504: IT Process Assessment [18]
IEEE Std. 1471-2000: System Architecture [15]
Zachman Framework [32]
COBIT: IT Governance [8]
DoDAF: Enterprise Architecture [11]
TOGAF: Enterprise Architecture [29]
SEI Capability Maturity Model Integration (CMMI) [7]
MoReq: Records Management [12]
ISO 15489: Records Management [17]
....
Stakeholders
Concerns
Influencers
Goals
Capabilities
Stakeholder
has
Concern
relates to
expressed by
Influencer
Key Question
can be
Constraint
Driver
motivates
Goal
limits deployment of
quantified by
is a
Capability
Key Performance Indicator
Desired Result
assessed by
delivers
Maturity
Course of Action
channels effort towards a

Stakeholder: An individual, team, or organization (or classes thereof) with interests in, or concerns relative to, a system [15]
Concern: Concerns are those interests which pertain to the system's development, its operation or any other aspects that are critical or otherwise important to one or more stakeholders [15]
Key Question: Stakeholder-made question to express a concern [29]
Influencer: Something that can cause changes that affect the enterprise in its employment of its Means or achievement of its Ends [24]
Constraint: An external factor that prevents an organization from pursuing particular approaches to meet its goals [29]
Driver: An external or internal condition that motivates the organization to define its goals [29]
Key Performance Indicator: (KPI) A way of quantifying the performance of the business or project [29]
Goal: A high-level statement of intent or direction for an organization. Typically used to measure success of an organization [29]
Capability: An ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve [29]
Course of Action: a way of configuring some aspect of the enterprise (things, processes, locations, people, time) to channel efforts towards Desired Results [24]
Desired Result: A Desired Result is an End that is a state or target that the enterprise intends to maintain or sustain [24]
Ends: Ends are about what an enterprise wants to be [24]
Means: Means are about what an enterprise has decided to do in order to become what it wants to be [24]
Maturity: The degree of reliability or dependency the business can place on a process achieving the desired goals or objectives [8]

Figure 4: Using TOGAF to integrate reference models creating a uniform view [1]

encing the object's behavior, (3) the decision on what level to compare the two objects, (4) recreating the environment, (5) applying standardized input to both environments, and finally (6) extracting and (7) comparing the significant properties. Even though the framework focuses mostly on emulation of environments, the principles are also applicable specifically for entire processes, and will work virtually unchanged also for migration approaches, when complex objects are transformed e.g into a new file format version.

A further component of the framework is the identification at which levels to measure the faithfulness of property preservation, as depicted in Figure 3. A rendered representation of the digital object has to be extracted on (a) suitable level(s) where the significant properties of the object can be evaluated. For some aspects, the rendering of an object can be performed based on its representation in specific memories (system/graphics/sound card/IO-buffer), for others the respective state changes at the output port have to be considered while for yet others the actual effect of a system on its environment needs to be considered, corresponding to delineating the boundaries of the system to be evaluated. (note that identity on a lower level does not necessarily correspond to identity at higher levels of the viewpath - in some cases significant effort are necessary to make up for differences e.g. on the screen level when having to emulate the visual behavior of cathode ray screens on modern LCD screens.) [13] An example of applying this framework to evaluation of preservation actions is provided in [4]

4. A CAPABILITY MODEL APPROACH TO DIGITAL CURATION

The types of institutions facing data curation challenges expands beyond the cultural heritage domain to include settings where curation is not the primary business goal. Rather, availability of data and processes is seen as an essential driver, be it due to legal/compliance requirements, as a contribution to business value, or other motivations. In settings where curation is not the main focus, it needs to be aligned with other core activities, integrating smoothly with its primary operations.

Data (and process) curation in research settings may be a typical example when curation is not delegated to a specific institution designated to preserve the data, but when preservation is happening as part of the research (and continued re-use) process. Moving beyond the more traditional data creation and use vs. data archiving approach we may want to aim at integrating all processes that revolve around data smoothly (and transparently for most actors) with curation activities.

To reach this goal, perspectives and approaches from fields such as Enterprise Architectures, Information Systems, Governance, Risk and Compliance may help in achieving a different view on data curation. This will assist in integrating digital curation as part of more generic (IT) operations while also offering a chance to make the needs and benefits of digital curation contributions to the overal value chain of an institution explicit. An overview of such an integrated view based on TOGAF [14], merging different models with the Shaman reference architecture is depicted in Figure 4. We also think that a process-based view on data curation rather than a data-centric view may help to better understand responsibilities, risks and costs involved to meet specific goals. It should also offer a more flexible basis for assessing the capabilities of an institution with respect to data curation, the level of maturity aimed at for specific capabilities, and allow more targeted actions to be planned in order to achieve them.

To this end we have further started modeling curation as a set of capabilities, with a range of maturity levels, as well as a clear specification of drivers and constraints, and their impact on an organization. An example of maturity levels for the capability *Preservation Operation* is depicted in Tab. 1. A detailed discussion of this approach is provided in [1, 2].

5. CONCLUSIONS

Ensuring quality in data curation for research is both simpler as well as more complex than ”standard” digital preservation. While it is in many respects similar to any kind of (more traditional, document-centric) data preservation, it raises significant challenges that require solutions going beyond what is currently available as state of the art solutions. While several aspects are predominantly extensions to cover e.g. new/specialized data formats, several challenges are rather unique in their importance to ensure the quality and authenticity of research data.

On the one hand, processes are an essential part of data provenance. Ensuring that any processing steps can be repeated, either on original data for verification and analysis purposes, or on new data to assure identical conditions, poses significant challenges in maintaining entire processing environments available and usable.

With the preservation of more complex environments, particular challenges emerge with respect to verifying the au-

	Awareness and Communication	**Policies, Plans and Procedures**	**Tools and Automation**	**Skills and Expertise**	**Responsibility and Accountability**	**Goal Setting and Measurement**
1	Management recognizes the need for preservation operations. There is inconsistent and sporadic communication.	Some operations are carried out, but they are not controlled. No useful documentation is produced about procedures and actions.	Some tools may be employed by individuals in an unsystematic ad-hoc manner.	There is no common awareness of which skills and expertise are required for which tasks.	There is no common awareness of responsibilities.	There is no clear awareness of goals; operations solely react to incidents and are not tracked.
2	Management is aware of the role of operations for authenticity and provenance. No formal reporting process exists, but there is some documentation about process results. Reports are delivered by individuals.	Some operational procedures emerge, but they are informal and intuitive. Operations rely on individuals; different procedures are followed within the organization. QA is recognized as a process, but mostly carried out ad-hoc and manual.	Automated tools are beginning to be employed by individuals based on arising needs and availability. Their usage is unsystematic and incoherent.	Staff obtain their operational skills through hands-on experience, repeated application of techniques and informal training by their peers.	Responsibility for operations emerges, but is not documented. Accountability is not defined.	There is individual awareness of short-term goals to achieve in operations, but no consistent goal definition or measurement.
3	Management understands the role of operations for authenticity and provenance. There are guidelines about statistics and reporting procedures, but they are not consistently enforced.	There is a defined process for all operations that relies on standardized plans. The processes and rules used are defined by available components, services and skills. QA and metadata management are not driven by business goals.	Plans are deployed according to specifications, but the process of initiating operations is mostly manual. No integrated system exists for tracking the state and results of operations.	A formal training plan has been developed that defines roles and skills for the different sets of operations, but formalized training is still based on individual initiatives.	Responsibility for operations is assigned, but accountability is not provided for all operations.	Operational goals are specified, but no formal metrics are defined. Measurements take place, but are not aligned to goals. Assessment of goal achievement is subjective and inconsistent.
4	Management fully understands the role of operations for authenticity and provenance and how they relate to business goals in the organization. Reporting processes are fully specified and adhered to.	Plans are fully deployed as operational activities, and the compliance of all operations to goals and constraints specified in plans is fully monitored. All Operations are actively monitoring state of operations.	An automated system exists to control automated operations, and automated components are widespread, yet not fully integrated.	Required skills and expertise are defined for all roles, and formal training is in place.	Responsibility and accountability for all operations is clearly defined and enforced.	A measurement system is in place and metrics are aligned with goals. Compliance monitoring is supported and compliance enforced in all operations.
5	Operations are continuously improving. An integrated communication and reporting system is fully transparent and operates in real time.	Extensive use is being made of industry good practices in plan deployment, analysis, actions, metadata, QA, and reporting.	All operations are fully integrated, status is constantly available in real-time.	Operators have the expertise, skills and means to conduct all operations. Continuous skills and expertise assessment ensures systematic improvement.	A formal responsibility and accountability plan is fully traceable to all operations.	Compliance is constantly measured automatically on all levels. Continuous assessment drives the optimization of measurement techniques.

Levels: 1: Initial/Ad-Hoc, 2: Repeatable but Intuitive, 3: Defined, 4: Managed and Measurable, 5: Optimized [6]

Table 1: Maturity Levels for the capability *Preservation Operation* [1]

thenticity of the performance/rendering of a process or data object in such a preserved environment. Formal models for these, as well as assistance in identifying and capturing the essential aspects needed for subsequent verification still represents a significant hurdle, with even more severe difficulties emerging from the need of automating any such validation in more generic settings.

Last, but not least, we feel that a shift from data-centric views of traditional approaches to depositing data somewhere for long-term curation needs to be superseeded by a view where curation processes are integrated into the operational environments. Furthermore, rather than auditing whether a specific sets of requirements is met by an institution tasked with curation we feel that a capability and maturity model based approach offers more flexibility to focus on essential aspects of data curation for a wide set of institutions.

Still, the considerations above cover only a small subset of the quite significant research challenges that continue to emerge in the field of digital curation. We thus strongly encourage the community to contribute to an effort of collecting and discussing these emerging research questions in a loosely organized form. To this end, following the Dagstuhl Seminar on Research Challenges in Digital Preservation[6], a Digital Preservation Challenges Wiki[7] has been created, where we invite contributions and discussion. As a follow-up to the Dagstuhl seminar, a workshop on DP Challenges[8] will be held at iPRES 2012 in Toronto focusing on the elicitation and specification of research challenges.

Acknowledgements

Part of this work has been supported by TIMBUS project, co-funded by the European Union under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940, and the SCAPE project, under grant agreement no. 270137, and APARSEN, grant agreement no. 269977.

6. REFERENCES

[1] C. Becker, G. Antunes, J. Barateiro, and R. Vieira. A capability model for digital preservation: Analysing concerns, drivers, constraints, capabilities and maturities. In *8th International Conference on Preservation of Digital Objects (IPRES 2011)*, Singapore, November 2011.

[2] C. Becker, G. Antunes, J. Barateiro, and R. Vieira. Control objectives for dp: Digital preservation as an integrated part of it governance. In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, New Orleans, Lousiana, US, October 2011.

[3] M. Guttenbrunner and A. Rauber. A Measurement Framework for Evaluating Emulators for Digital Preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 2012.

[4] M. Guttenbrunner and A. Rauber. Evaluating an emulation environment: Automation and significant key characteristics. In *Proceedings of the 9th conference on Preservation of Digital Objects (iPRES2012)*, Toronto, Canada, October 1–5 2012.

[6] http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=10291

[7] http://sokrates.ifs.tuwien.ac.at

[8] http://digitalpreservationchallenges.wordpress.com/

[5] M. Guttenbrunner and A. Rauber. Evaluating emulation and migration: Birds of a feather? In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries*, Taipei, Taiwan, November 12–15 2012.

[6] IT Governance Institute. *COBIT 4.1. Framework – Control Objectives – Management Guidelinces – Maturity Models*. 2007.

[7] Y. Marketakis and Y. Tzitzikas. Dependency management for digital preservation using semantic web technologies. *International Journal on Digital Libraries*, 10:159–177, 2009.

[8] R. Mayer, S. Pröll, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of the 9th conference on Preservation of Digital Objects (iPRES2012)*, Toronto, Canada, October 1–5 2012.

[9] R. Mayer and A. Rauber. Towards time-resilient MIR processes. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, October 8-12 2012.

[10] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, LNCS, Cyprus, September 2012. Springer.

[11] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna, reloaded. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 471–481. Springer, June 2010.

[12] L. Moreau, J. Freire, J. Futrelle, R. E. Mcgrath, J. Myers, and P. Paulson. *Provenance and Annotation of Data and Processes*, chapter The Open Provenance Model: An Overview, pages 323–326. Springer, 2008.

[13] G. Phillips. Simplicity betrayed. *Communications of the ACM*, 53(6):52–58, 2010.

[14] The open Group. *TOGAF Version 9*. Van Haren Publishing, 2009.

Position Paper: Data Curation for Quality

Kristin M. Tolle
Microsoft Research

Abstract:

Increasingly, funding agencies are beginning to require data management plans for for projects involving the collection of scientific data. To support this effort in environmental science, National Science Foundation has sponsored DataONE[1], a consortium of data repositories with the mission to "ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data". Beyond data management, data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for re-use over time, and includes authentication, archiving, management, preservation, retrieval, and representation [1]. This paper discusses the implications of these data management efforts from the perspective of data curation that ensures data reuse and quality and describes a new tool, DataUp, which is designed to help address the pain points of data curation for Environmental Scientists.

Introduction:

Data sharing and curation have become critical to both scientists and private and public agencies that support their work. As pointed out by Fry et al. "It is becoming increasingly clear that effective and efficient management and reuse of research data will be a key component in the knowledge economy in years to come, essential for the efficient conduct of research and its dissemination and use." [2] Often scientists' data management plans include just basic data storage—usually locally. This means that their data may or may not be visible to scientist studying the same phenomena or useable by others in the future, even within their own lab. So not only is this data not making it into public repositories, if it is preserved it is done so with little or no associated metadata [3].

Data curation is defined as "the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education." [4] What has been beneficial to enabling data curation is the building of data repositories and making them available to scientific users as is the requirement of funding agencies for data management plans from grant awardees. However, these approaches have not, to date, solved the data curation crisis—where thousands of scientists are creating millions of dataset every year—largely stored on local data storage. In [5], Lyon states "It is acknowledged that a huge cultural change is required in order to realise this vision [of data curation], both amongst researchers and publishers. Researchers are perceived to have not yet embraced or fully understood the principles of [Open Access]."

[1] http://www.dataone.org/

The DataUp project came together as a result of a joint dialogue between the Moore Foundation, Microsoft Research and the California Digital Library; all agreed that there needed to be tools to bridge the gap between scientists that are developing multitudes of small datasets but not storing them in a databank where they can be used by other scientists—often referred to as the “long tail” of science [6].

The first step towards solving a problem is developing an understanding of users’ specific needs. According to a study of ecologists funded by Microsoft Research, despite a willingness and desire to do so, environmental scientists encounter many barriers to placing their data in publically accessible repositories. Not the least of which is the lack of knowledge of the existence of (semi-)public repositories where they might be able to place their data for preservation and reuse. And though more than half of those surveyed were aware of the uses of metadata, they were unlikely to understand what associative metadata is necessary to enable them to be domain and repository compliant. So even if they make it past the first barrier, the second one limits reuse.

Another barrier is that scientists do not consistently name the fields in their data or adhere to common best practices in formatting data for reuse. These and other difficulties make understanding the quality of the data even more difficult. Therefore any proposed solution for data curation must take into consideration many of the basic hurdles and work to enhance quality, citation and reuse. Such a comprehensive solution should, in turn, enable faster and more efficient research that can cut across related disciplines, potentially increasing the pace and quality of scientific advancement [7].

Our data shows that easy-to-use tools are needed to enable high-quality data curation as well as bridge the gap between tabular data in the wild, present data management plans and data repositories. Present technology is barrier and that can be could turn into a bridge to facilitate data curation lifecycle. The benefit of achieving this goal is stated by Lord and MacDonald in [8] is that “digital technologies enable sophisticated collaboration and sharing within and between disciplines (where some of the most fruitful work lies). Proper retention of digital data is essential to demonstrate validity, and for respect of legal and ethical values.”

This paper will report addition feedback from several surveys of environmental scientists, as well as solicited feedback from repositories and publishers working to curate and preserve scientific data collected over a six month period. It will touch on a tool that we have developed to help facilitate the process of data curation and also make some general recommendations on methods that will facilitate ease of use, compliance as well as quality and reuse for the sometimes conflicting needs of both users and repositories.

User Data Requirements Gathering

Several months were spent collecting requirements from the environmental science community. Assessment of the communities’ needs focused on scientists and included assessing the needs of libraries and data centers. Data center and library assessments were collected via in-person conversations and interviews, and web-based surveys.

Feature Discussion:

Gathering feedback with and from the California Digital Library as well as from discussions with other repositories, we were able to collect the following list of features for an "ideal" data curation tool. The list below is a high level set of feature requirements that we are working to translate into an application for use in the community:

- Data management/curation needs to be built into the tools that scientists presently and frequently use (e.g., Excel®).
- Unique digital object identifiers (DOIs) are valuable for data sharing, publication and citation. They should be readily be assigned to datasets for future reference at the time a dataset is uploaded into a repository.
- A citation for the data that can be inserted into publications should also be generated for the user on publishing their data to a repository.
- Metadata should be defined with the data—and automatically generated when possible to facilitate compliance. For those fields that cannot be generated but are required by the repository users should be prompted for this information prior to or during upload.
- Repository metadata requirements need to be visible to users who are uploading data into repositories and there needs to be a "validation" mechanism for users prior to upload to ensure data searchability and reuse.
- Data collected needs to be "cleaned" and a set of best practices for data sharing should be exposed to the user *prior to upload* so that they can be educated in the tool on how best to format their data for sharing.
- Repositories should be able to create selective (lightweight) metadata that can be mapped to an associated domain or set of domains and specify which applies to what type of data in addition to a set of minimum requirements to be met prior to upload.
- Users should be able to select a domain, upload their data and validate only the metadata that is relevant to their data domain and/or repository.
- Repositories need to be "advertised" within tools. Those repositories that are available and relevant to them should be easily discoverable and selectable. If a login is required, the ability to request access to a repository should be integrated and seamless.
- Rationalization (mapping between) of metadata from other domains should occur without users of one domain needing to be conversant in the metadata requirements of other domains.

The DataUp Tool

The DataUp project's goal is a step towards providing tools that help researchers document, organize, preserve, and share their scientific data. As previously indicated, for tractability purposes we focused on assisting Earth, environmental, and ecological scientists. That said, this is an extremely diverse set of scientists and who collect and use all manner of data, from visual images and satellite data, from static field collected data sets to continuously monitored data generated by sensors. In some cases they combine these data sets through further analysis.

In this section, we will describe the DataUp add-in for Excel® and web application. Both the add-in and the web application perform four main tasks: (1) perform a best practices check to ensure good data organization, (2) help guide the user through creation of metadata for their Excel® file, (3) help the user obtain a unique identifier for their dataset, and (3) connect the user to a DataONE repository, where their data can be deposited and shared with others.

Originally the plan was to one application: an Excel® add-in to assist with uploading data to repositories. Owing to the difficulty of maintaining an add-in which is likely to need code changes at each new software release of Microsoft Office®, we also investigated, initially as a migration path, a web application which could be hosted on Azure in the cloud or at a repository. We weighed the pros and cons of each as shown in Table 1 below.

Table 1: Web Application vs. an Add-in

	Add-In	**Web Application**
Skilled Developers	Harder to find.	Easier to find.
Testing Resources and Tools	Manual UI testing only. Several platforms to test.	Many tools available. HTML5 resources aid UI testing.
Office Performance	Office starts more slowly.	Office not affected.
Office Compatibility	Windows only. Add-in code for each Office version. Can only support Excel® 2007 and later.	Can handle files from **any** platform.
Versioning / User Experience	Fixed bugs require download and re-install.	Bugs are fixed in one location and all users updated.
Offline use	Yes.	No.

Both options have clear benefits to users. We decided that this decision was the same as any other user requirement. So we polled a similar audience via social media and email for their input to decide which option was preferred. The response was split at nearly 50% on whether they would prefer one type of solution over the other.

Installing software on a computer is not always possible on organizationally owned machines. Users may not have permission to download and install an add-in. Citing this as a problem, several respondents recommended we go with a Web Application. A representative of one of the repositories commented, “A web app can incorporate new functionality (new data sources, new metadata standards etc.) with minimal user involvement.”

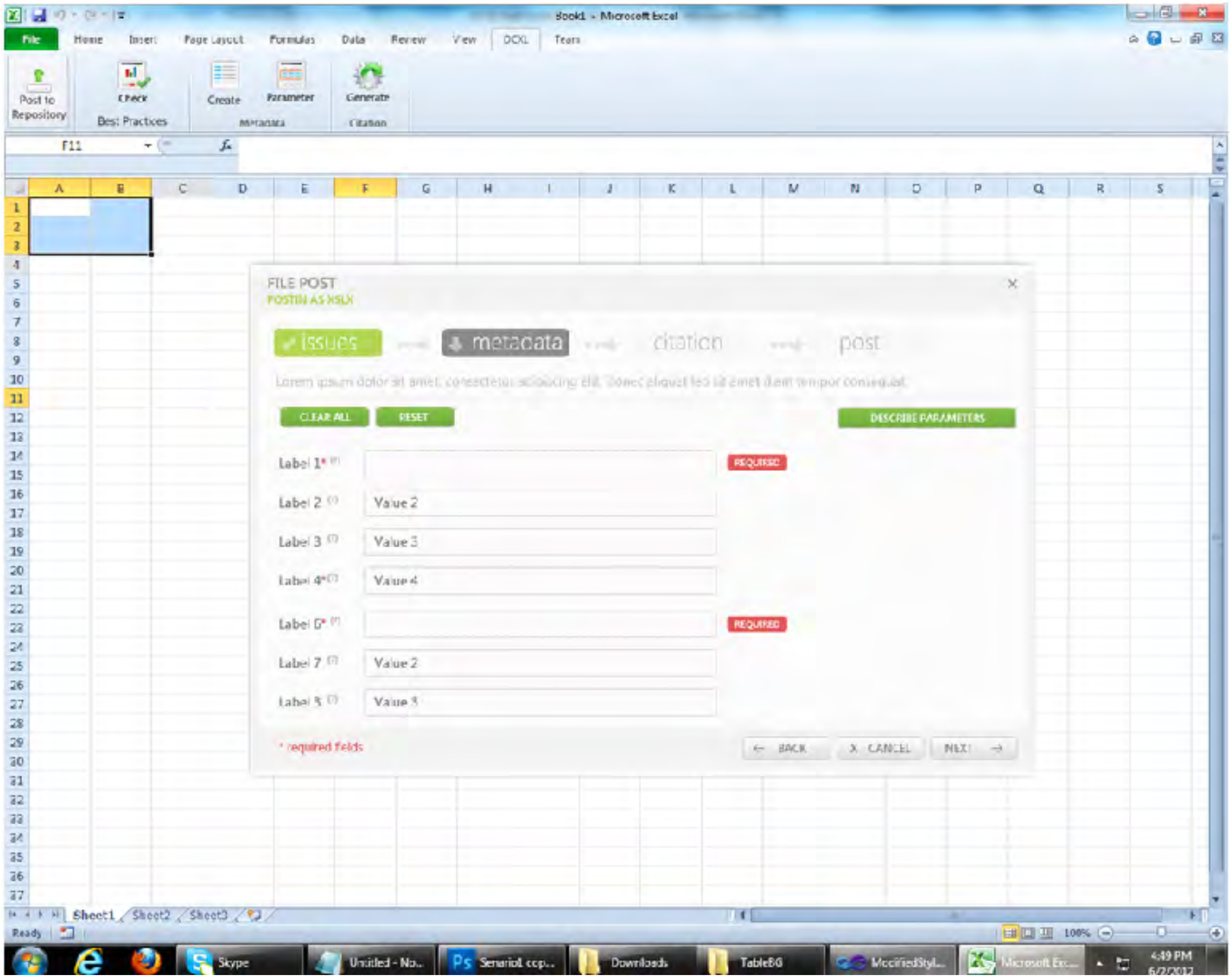
To keep it all within Excel® turned out to be another critical feature. As one person from DataONE commented, “…my strong preference is to enable Excel operation for data and metadata capture offline.”

Some of those polled were also of mixed opinions. In the end we decided it was feasible to provide both solutions and extended our development timeline. Providing both in the short term, we believe, will provide a solution to the broadest number of users for a long period of time.

Installable Add-in for Microsoft Excel

Users can download from the Outercurve Foundation[2] and install DataUp add-in. Once installed, it becomes an integral part of the MS Excel® software on their machine that enables them to clean and curate their data files as well as post them to the ONEShare repository.

Users can start by using each of the functions that preceed uploading data to a repository individually in the DataUp ribbon: Checking for Best Practices and Compatibility, Generating Metadata (both extracted from the data itself and user defined), or generating a data citation with a unique data identifier. Alternatively, they can click the Upload Data button and be stepped through these processes. The application provides the flexibility of doing these in either an ordered or non-ordered way and retains the information regards of which way the user chooses to interact with the user interface.



FILE POST
POSTIN AS XSLX

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec aliquet leo sit amet diam tempor consequat.

Label 1*

Label 2 Value 2

Label 3 Value 3

Label 4* Value 4

Label 6*

Label 7 Value 2

Label 5 Value 3

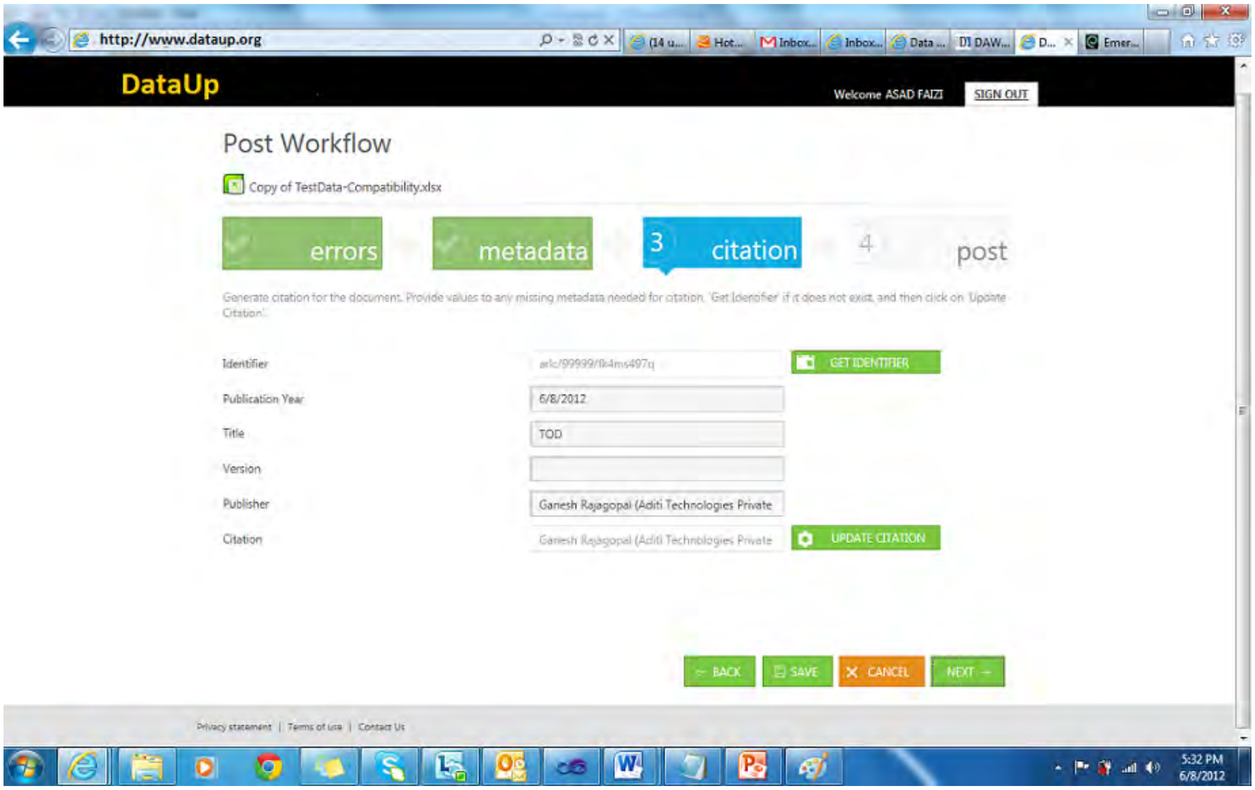
* required fields

[2] http://www.outercurve.org/

Web Application, Deployed in the Cloud

The web application version of DataUp is an Azure hosted service in the cloud. This solution allows users access to the same functionality as the add-in without requiring a software installation on their computer. The user can simply use their browser to get to the DataUp.org website to access the tool. Once they register with the service they are placed into a similar experience as the Add-in’s stepped “Upload” process. Should the user wish to revisit a step they can do so at any time in the upload process by using the top row of tabs.

As with the Add-in, the web services version is also available on the OuterCurve Foundation website. This portion of the code is placed in the open source domain to enable a repository to download, customize and host web application and use it locally by adapting the system to match their hosting environment. The web application was originally developed and deployed on Windows Azure with SQL Server. It is licensed under Apache 2.0 open source licensing.



Post Workflow

Copy of TestData-Compatibility.xlsx

Generate citation for the document. Provide values to any missing metadata needed for citation. 'Get Identifier' if it does not exist, and then click on 'Update Citation'.

Identifier	ark:/99999/fk4ms497q
Publication Year	6/8/2012
Title	TOD
Version	
Publisher	Ganesh Rajagopal (Aditi Technologies Private
Citation	Ganesh Rajagopal (Aditi Technologies Private

Conclusions and Future Directions

- Data preservation, curation and sharing are critical for the advancement of scientific discovery.
- Scientists and researchers have awareness and understanding of the significance of data curation and sharing, and want to share data beyond their immediate groups.
- They are, however hampered by the lack of tools and services designed to promote data curation and sharing and collaboration.

- Quality must be integrated into the data creation and management process.

Acknowledgements:

The study on Excel® usage by environmental scientists was conducted by Carly Strasser[3] under the supervision of Patricia Cruse[4] from the University of California Digital Library[5] and was co-funded by the Gordon and Betty Moore Foundation[6].

References:

[1] Data Curation Centre, UK, http://www.dcc.ac.uk/resources/curation-lifecycle-model/ , Accessed June 2012.

[2] Fry, J. et. al. (2008). "Identifying the Benefits of Curating & Sharing Research Data", Produced by UK Higher Education and research institutes. Accessed at Joint Information Systems Committee (JISC): http://www.jisc.ac.uk/publications/documents/databenefitsfinalreport.aspx.

[3] Abel, et al, Scientific Collaborations for Extreme-Scale Science, Workshop Report, December 6-7, 2011, Gaithersburg, MD, DOE.

[4] University of Illinois Graduate School of Library and Information Science, http://www.lis.illinois.edu/academics/programs/ms/datacuration, Accessed June 2012.

[5] Liz Lyon, Chris Rusbridge, Colin Neilson, and Angus Whyte, "JISC Final Report: Disciplinary Approaches to Sharing, Curation, Reuse and Preservation", http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

[6] Peter Murray-Rust's Blog: A Scientist and the Web; "Publishing Data: The long-tail of science", http://blogs.ch.cam.ac.uk/pmr/2011/08/14/publishing-data-the-long-tail-of-science/, Accessed June 2012.

[7] Sonnenwald, D. (2008). Scientific Collaboration. Annual Review of Information Science and Technology, 41, pgs 643-681.

[8] Lord, P. & MacDonald, A. (2003). e-Science Curation Report-Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision. Prepared for the JISC Committee for the Support of Research (JCSR). Twickenham, UK, The Digital Archiving Consultancy Limited.

[3] http://www.linkedin.com/pub/carly-strasser
[4] http://www.cdlib.org/contact/staff_directory/pcruse.html
[5] http://www.cdlib.org/
[6] http://www.moore.org/

Curating for Data Quality at the Protein Data Bank: Ensuring Data Quality to Enable New Science

Jasmine Y. Young
Center for Integrative Proteomics Research
Rutgers, The State University of New Jersey
Piscataway, NJ 08854
1.848.445.4920
jasmin@rcsb.rutgers.edu

John Westbrook
Center for Integrative Proteomics Research
Rutgers, The State University of New Jersey
Piscataway, NJ 08854
1.848.445.4919
jwest@rcsb.rutgers.edu

Helen M. Berman
Center for Integrative Proteomics Research
Rutgers, The State University of New Jersey
Piscataway, NJ 08854
1.848.445.4667
berman@rcsb.rutgers.edu

ABSTRACT

The Protein Data Bank (PDB) is the single archive for 3D macromolecular structures. The archive serves as a primary and critical resource for research in structural biology and in drug discovery worldwide. As scientists require consistent and highly accurate data for their research, the quality of the data in the archive is regularly reviewed. This paper describes the processes and tools used by the Worldwide PDB (wwPDB) to maximize data quality of individual structures and across the archive. This includes the development of a new deposition and annotation system that focus on the improvement of data quality and effectiveness of the curation processes.

General Terms

Management, Documentation, Standardization, Verification

Keywords

PDB, biomacromolecular structure, data quality, curation

1. INTRODUCTION

The Worldwide Protein Data Bank (wwPDB; wwpdb.org) consists of four organizations located in the USA, Europe, and Japan that together collect, curate, and disseminate the single global PDB archive of biomacromolecular structures [1; 3]. The wwPDB members are the Research Collaboratory for Structural Bioinformatics (RCSB) PDB at Rutgers, The State University of New Jersey [2]; Protein Data Bank Europe (PDBe) at the European Bioinformatics Institute in the United Kingdom [14]; Protein Data Bank Japan (PDBj) at the Institute for Protein Research at Osaka University, Japan [7]; and the BioMagResBank at the University of Wisconsin-Madison [10].

The wwPDB provides a global resource for the advancement of research and education in biology and medicine by curating, integrating, and disseminating biological macromolecular structural information in the context of function, biological processes, evolution, pathways and disease states. In carrying out this function, the wwPDB implements standards and develops appropriate technologies to support evolving science.

All the data in the PDB archive are freely and publicly available. The archive contains atomic coordinates that are substantially determined by experimental measurements on actual sample specimens containing biological macromolecules. Currently, coordinate sets produced by X-ray crystallography, Nuclear Magnetic Resonance (NMR), electron microscopy (3D EM), neutron diffraction, powder diffraction, fiber diffraction, and solution scattering can be deposited. Experimental data collected during the course of the structure determination process (structure factors for X-ray crystallography, restraints and chemical shifts for NMR) are deposited along with the atomic coordinates.

As an example, the structure determination pipeline for X-ray crystallography is illustrated in Figure 1. The target macromolecule is first selected, then expressed, purified and crystallized. The crystal is subsequently used to collect X-ray diffraction data and the atomic coordinates for the structure are produced by refinement software. Parameters and software used during this process are an integral part of the quality profile and need to be captured as completely and accurately as possible. During data annotation, the atomic structures are further curated with structural and functional information such as biological processes, pathway, and ligand interactions. Validation reports summarizing key structural features and anomalies are generated and reviewed by the structure authors and the wwPDB.

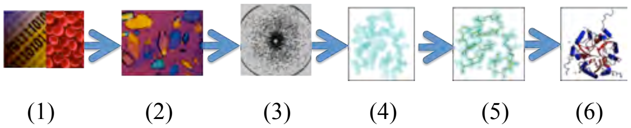


Figure 1. X-ray structure determination pipeline: (1) target selection, (2) crystallomics, (3) data collection, (4) structure solution, (5) structure refinement, (6) functional annotation.

All wwPDB deposition sites have formalized many aspects of PDB annotation policies and procedures to ensure the uniformity of data format and data files. The PDB format guide and annotation policies and procedures are documented on the wwPDB website (wwpdb.org).

2. APPROACH TO DATA QUALITY

2.1 Data Content

The data collected include experimental data, 3D atomic coordinates, information about the composition of the structure (sequence, chemistry, etc.), information about the experiment performed, details of the structure determination steps, and author contact information. In addition to data collected from authors, the wwPDB provides functional annotation, such as structural features and protein modification, and derived calculations such as sequence database cross-reference, ligand binding sites, biological assemblies, and geometric deviations.

2.2 Structure Data Quality Issues

Providing accurate chemical descriptions is a major focus of annotation. During processing, the deposited coordinates are dissected into individual chemical components and represented as their neutral or free form. These components are compared against standardized dictionaries to support uniformity across the archive [4]. The correlation of model coordinates and experimental data are calculated to evaluate the goodness of fit.

The chemical information in the PDB is experimentally-derived and therefore subject to observational and modeling restraints. In many cases, the component structure is obvious and the definition straightforward. However, because of the diverse nature of non-polymer components, the definition may not be clear for a variety of reasons. There are instances, for example, when the stability of a component is dependent on its interactions with a biopolymer which creates a perturbation relative to the "free standing" form. This may then result in an anomaly relative to the ideal geometry for the perceived component. At 1.3 Ångstrom resolution (Figure 2a), which is near atomic resolution in X-ray experiments, the atom density can be observed and the model can be unambiguously fitted. However, at a more typical 2.2 Ångstrom resolution (Figure 2b), only a block of density can be observed, and the model fitted to the density map is a more subjective interpretation. Human analysis of these and other exceptions is required to ensure data quality of the deposition.

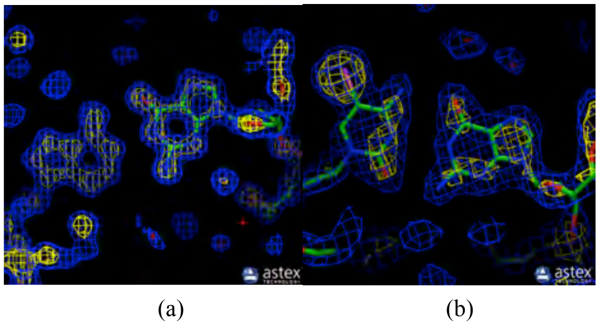


Figure 2. Model coordinates fitted to electron density: (a) at 1.3 Å resolution (PDB ID 3dnb) and (b) at 2.21 Å resolution (PDB id 6bna).

2.3 Validation

The wwPDB is committed to using the highest standards of curation and validation to process PDB structures. wwPDB members document data features in the PDB files that permit users to make informed decisions regarding quality. These features include self-consistency with respect to sequence and coordinates, sequence and taxonomy cross-references, ligand chemistry, model geometric, and model correlation to the experimental data using community-accepted algorithms [8; 11] as part of the structure annotation processes. The wwPDB members maintain close relationship with related data resources such as UniProtKB, Norine and NCBI databases for sequence and taxonomy cross-references.

Validation reports are produced at the end of annotation pipeline, and highlight outstanding issues that require corrections. These validation reports provide an assessment of structure quality without revealing coordinate data, thereby protecting author intellectual property. The wwPDB encourages journal editors and referees to request these reports from depositors as part of the manuscript submission and review process. A PDF version of the validation report is generated during the annotation process for depositors to include with their journal submissions. The reports are date-stamped, and display the wwPDB processing site logo. These reports are currently required by the *Journal of Biological Chemistry* and *Acta Crystallographica* volumes D and F.

2.4 Maintaining Data Uniformity Archive-wide

Improving the quality of data in the archive is another annotation focus. To insure the data are represented in the best way possible, wwPDB members regularly review the archive to correct errors and inconsistencies. These remediation efforts result in the creation of a new set of data files. To date remediation efforts have taken place in 2007 [6; 9], 2008 and 2011, and are documented on the wwPDB website. The remediation helps to identify systematic errors that inform updates and improvements to the tools used for annotation and processing. Current remediation efforts are focused on the uniform representation of carbohydrate and protein modifications.

2.5 Community Involvement: wwPDB Task Forces

The wwPDB members work with community experts to develop best practices in data quality standards. Method-specific Validation Task Forces (VTF) have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks. Several workshops have been held for X-ray, NMR, 3D EM and Small Angle Scattering (SAS). The outcome and recommendations from X-ray and 3D EM have been published. [12; 13]

These recommendations will be incorporated into the validation tools used by the wwPDB for data annotation [5].

3. PROCESSES AND TOOLS TO IMPROVE DATA QUALITY AND EFFICIENCY

3.1 PDBx Exchange Dictionary

The PDB has developed and uses the PDB exchange/macromolecular Crystallographic Information File (PDBx/mmCIF) data dictionaries [4] to describe the information content of PDB entries. The PDBx dictionary documents semantics; provides a software accessible resource for standardization (e.g. data relationships, data type consistency, boundary values, controlled vocabularies, etc.); enables extensibility in description of evolving science and technology. Recent enhancements to the dictionary address the increasing complexity and size of deposited structures and support efforts in improved file consistency.

3.2 Capture and Processing of Quality Files

Due to rapid growth of the number and complexity of structures deposited to the PDB, curation processes are constantly under review for the improvement of efficiency and data quality assurance. The deposition tool (ADIT) used by the RCSB PDB and PDBj has been improved through the years to capture more complete data including chemistry information and to include validation as part of deposition procedure. The RCSB PDB data harvesting tool (pdb_extract) [15] has been developed to extract data from refinement output files and generate coordinates and structure factor files that are ready for effective PDB deposition. Additionally, the stand-alone RCSB PDB validation server has been improved to provide PDF validation reports similar to the report produced during the PDB annotation pipeline for depositors to review and correct the data before PDB submission.

Processing efficiency supports data quality during curation. Several approaches have been taken to improve processing efficiency. Currently all of the data processing components and report generation steps in the curation pipeline are integrated into one online tool. This allows annotators to spend more time focused on the content of the files while running routine processes. Batch processing has also been implemented to improve curation efficiency.

Further improvements are still needed for effective curation and better data quality at the PDB.

3.3 New Tools

The wwPDB members are developing a new Deposition and Annotation (D&A) system. The goal of this project is to provide a common, enhanced system for deposition and processing of PDB entries globally that makes use of the best practices across the partner sites and focuses on data quality as the enabler of efficiency. The shared processes and tools will support the current and projected increases in complexity and experimental variety of submissions as well as the projected increases in deposition throughput. The D&A system features interactive graphical user interfaces that will engage the depositor in providing more complete and higher quality submissions. The annotation pipeline likewise employs graphical user interfaces that enhance efficiency. Both pipelines are supported by text entry validation and summary validation reports on the structure data.

The new D&A system is comprised of modules that address each of the major processing functions. The Ligand Module, for examples, supports the processing and annotation chemical components. In addition to the graphical user interface for annotator review of ligand data, the module supports batch chemistry assignments, improves standardization and geometric validation functionality. This module has been implemented into the current annotation pipeline at the RCSB PDB. Benchmark studies have shown a significant improvement with the processing efficiency up 70%. The total number of entries processed per full time employee has increased from 526 to 706 from year 2010 to 2011.

User testing of the new D&A system will take place in 2013.

4. ACKNOWLEDGMENTS

The wwPDB is supported by: RCSB PDB (NSF DBI 0829586, NIGMS, DOE, NLM, NCI, NINDS, and NIDDK), PDBe (EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, and EU), PDBj (NBDC-JST) and BMRB (NLM).

5. REFERENCES

[1] BERMAN, H.M., HENRICK, K., and NAKAMURA, H., 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol 10*, 12, 980.

[2] BERMAN, H.M., WESTBROOK, J.D., FENG, Z., GILLILAND, G., BHAT, T.N., WEISSIG, H., SHINDYALOV, I.N., and BOURNE, P.E., 2000. The Protein Data Bank. *Nucleic Acids Res. 28*, 235-242.

[3] BERNSTEIN, F.C., KOETZLE, T.F., WILLIAMS, G.J.B., MEYER JR., E.F., BRICE, M.D., RODGERS, J.R., KENNARD, O., SHIMANOUCHI, T., and TASUMI, M., 1977. Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535-542.

[4] FITZGERALD, P.M.D., WESTBROOK, J.D., BOURNE, P.E., MCMAHON, B., WATENPAUGH, K.D., and BERMAN, H.M., 2005. 4.5 Macromolecular dictionary (mmCIF). In *International Tables for Crystallography*, S.R. HALL and B. MCMAHON Eds. Springer, Dordrecht, The Netherlands, 295-443.

[5] GORE, S., VELANKAR, S., and KLEYWEGT, G.J., 2012. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Cryst D68*, 478-483.

[6] HENRICK, K., FENG, Z., BLUHM, W.F., DIMITROPOULOS, D., DORELEIJERS, J.F., DUTTA, S., FLIPPEN-ANDERSON, J.L., IONIDES, J., KAMADA, C., KRISSINEL, E., LAWSON, C.L., MARKLEY, J.L., NAKAMURA, H., NEWMAN, R., SHIMIZU, Y., SWAMINATHAN, J., VELANKAR, S., ORY, J., ULRICH, E.L., VRANKEN, W., WESTBROOK, J., YAMASHITA, R., YANG, H., YOUNG, J., YOUSUFUDDIN, M., and BERMAN, H.M., 2008. Remediation of the Protein Data Bank Archive. *Nucleic Acids Res 36*, Database issue, D426-D433.

[7] KINJO, A.R., SUZUKI, H., YAMASHITA, R., IKEGAWA, Y., KUDOU, T., IGARASHI, R., KENGAKU, Y., CHO, H., STANDLEY, D.M., NAKAGAWA, A., and NAKAMURA, H., 2012. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res 40*, Database issue (Jan), D453-460. DOI= http://dx.doi.org/gkr811 [pii] 10.1093/nar/gkr811.

[8] KLEYWEGT, G.J. and JONES, T.A., 1996. Phi/psi-chology: Ramachandran revisited. *Structure 4*, 12 (Dec 15), 1395-1400.

[9] LAWSON, C.L., DUTTA, S., WESTBROOK, J.D., HENRICK, K., and BERMAN, H.M., 2008. Representation of viruses in the remediated PDB archive. *Acta Cryst. D64*, 874-882.

[10] MARKLEY, J.L., ULRICH, E.L., BERMAN, H.M., HENRICK, K., NAKAMURA, H., and AKUTSU, H., 2008. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR 40*, 3 (Mar), 153-155. DOI= http://dx.doi.org/10.1007/s10858-008-9221-y.

[11] PARKINSON, G., VOJTECHOVSKY, J., CLOWNEY, L., BRÜNGER, A.T., and BERMAN, H.M., 1996. New parameters for the refinement of nucleic acid containing structures. *Acta Crystallogr. D52*, 57-64.

[12] READ, R.J., ADAMS, P.D., ARENDALL, W.B., III, BRUNGER, A.T., EMSLEY, P., JOOSTEN, R.P., KLEYWEGT, G.J., KRISSINEL, E.B., LUTTEKE, T., OTWINOWSKI, Z., PERRAKIS, A., RICHARDSON, J.S., SHEFFLER, W.H., SMITH, J.L., TICKLE, I.J., VRIEND, G., and ZWART, P.H., 2011. A new generation of crystallographic validation tools for the Protein Data Bank.

Structure 19, 10 (Oct 12), 1395-1412. DOI= http://dx.doi.org/S0969-2126(11)00285-1 [pii]

10.1016/j.str.2011.08.006.

[13] SCHUMACHER, M.A., MIZUNO, K., and BACHINGER, H.P., 2006. The crystal structure of a collagen-like polypeptide with 3(S)-hydroxyproline residues in the Xaa position forms a standard 7/2 collagen triple helix. *J Biol Chem 281*, 37 (Sep 15), 27566-27574. DOI= http://dx.doi.org/M602797200[pii]10.1074/jbc.M602797200.

[14] VELANKAR, S., ALHROUB, Y., ALILI, A., BEST, C., BOUTSELAKIS, H.C., CABOCHE, S., CONROY, M.J., DANA, J.M., VAN GINKEL, G., GOLOVIN, A., GORE, S.P., GUTMANAS, A., HASLAM, P., HIRSHBERG, M., JOHN, M., LAGERSTEDT, I., MIR, S., NEWMAN, L.E., OLDFIELD, T.J., PENKETT, C.J., PINEDA-CASTILLO, J., RINALDI, L., SAHNI, G., SAWKA, G., SEN, S., SLOWLEY, R., SOUSA DA SILVA, A.W., SUAREZ-URUENA, A., SWAMINATHAN, G.J., SYMMONS, M.F., VRANKEN, W.F., WAINWRIGHT, M., and KLEYWEGT, G.J., 2011. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res 39*(Nov 2), D402-D410 DOI= http://dx.doi.org/gkq985 [pii]10.1093/nar/gkq985.

[15] YANG, H., GURANOVIC, V., DUTTA, S., FENG, Z., BERMAN, H.M., and WESTBROOK, J., 2004. Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr 60*, 1833-1839.

Position Papers: 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 will:

- 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)

Generic Data Quality Metrics – what and why

Kevin Ashley
Digital Curation Centre
Appleton Tower
Crichton St, EDINBURGH, Scotland
+44 131 651 3823
Kevin.Ashley@ed.ac.uk

ABSTRACT

Data quality is often discussed as if were a single-dimensional scalar measure applying to a particular dataset. In fact much existing research recognizes that there are a number of generic dimensions to data quality and that some of them are inversely related to each other (such as timeliness or accuracy versus cost.) Different disciplines place different emphasis on and assign different values to these measures; often they give them discipline-specific names. This has the unfortunate effect of making it difficult to train data managers in generic issues of data quality and to recognize when tools developed to solve a quality issue in one discipline can be effective in another. It also poses challenges to the integration of data in multi-disciplinary research. Even assuming that machine-readable quality metadata is provided (a rarity), non-generic expressions of quality prevent the integration of data with similar quality measures. Wider awareness amongst the data curation community of such generic measures as those described by Wang and Strong would lead to improvements in all these areas – education and training, tool reuse and development, and data interoperability.

Categories and Subject Descriptors

E.0 [**Data**]: General

General Terms

Measurement, Standardization

Keywords

Keywords are your own designated keywords.

1. INTRODUCTION

This position paper revisits issues I described at greater length in a briefing paper [Ashley 2011] I produced for the GRDI2020[x] project on data curation & quality in the context of global research data infrastructures. In it I mainly address the 4th question that workshop attendees have been asked to consider – metrics of data quality. But I also touch on parts of question 1 (criteria for quality) and 2 (costs); in particular I argue that cost should simply be seen as another dimension of quality rather than as something separate from quality. Finally, my recommendations for machine-readable generic assertions of data quality touch on question 3, which relates to techniques which make curation 'effective and painless.' Machine-readable quality metadata won't make curation painless, but it will make much data reuse much more painless and in that sense it will make the curation far more effective. The end purpose of curation is almost always to enable reuse.

2. WHAT DO WE MEAN BY QUALITY?

Quality is a property of data on which it initially appears difficult to get disagreement – everyone agrees that quality is a good thing. Researchers believe that they produce high quality data and offer that to domain or subject data centres for others to reuse. Researchers seeking data go to domain-specific data centres in the belief that they are sources of high-quality data. The data centres will say that they operate stringent controls to ensure that they select data of high quality, apply processes that improve its quality further (or at least check quality assertions) and finally make the even-higher-quality data available. All of these actors in the data lifecycle are right about data quality up to a point. Unfortunately they aren't all speaking about the same thing. This would not be a problem if their assertions were explicit, but they usually aren't. This can lead to misunderstandings, inefficiencies, lack of interoperability or bad research. The latter, in the worst case, can lead to bad policy and decision making outside the research community. Yet the problem of expressing and communicating assertions about data quality is by no means intractable.

Studies, often carried out with data users outside the world of research, show that we have different, tacit assumptions about what data quality means and that we often treat it as a one-dimensional measure. Data is low quality or medium quality or high. If pressed, we might even quantify a particular dataset on a 5 or 10 point scale. But in a particular situation perhaps I actually want data that is comprehensive; you are concentrating on producing data that is timely; and the data producer is trying their hardest to produce data that is accurate. Each of these is a good measure of data quality for a specific purpose. They aren't the same measure and they may be in conflict with each other. For this reason amongst others it is also useful to think of cost as just another quality parameter. You are rich and you want timely, accurate data. Typically processes that produce accurate data take more time. Money can reduce the effect but not eliminate it. Cost is not an issue for you and therefore is an irrelevant, though measureable, parameter of quality. You are still forced to balance the relative importance of accuracy against timeliness. I am poor; although I want the same things as you I can't afford them. For me, cost is a third quality variable to balance against timeliness and accuracy. Quality parameters aren't always in conflict in this way; for instance, documented provenance and accuracy are relatively independent. It's easy to have neither, either one alone or both in a single dataset.

Most people recognise these multiple dimensions of quality once they are pointed out; they aren't difficult concepts to grasp. In general, though, we aren't explicit about them either as consumers or producers. When we are explicit, we tend to express ourselves in very domain-specific ways. What we also tend to forget is that many of the qualities we seek are actually measurable surrogates for the things we really want but can't measure, such as truth.

The work of Wang and Strong [Wang and Strong, 1996] showed that it is possible to transcend domain-specific quality attributes and reduce them to a manageable number that still suffices for almost all use cases. They aren't the only people to have done this type of work but I have found their results more accessible than some of the alternatives. Based on interviews with data creators, users and managers in many settings they initially established a detailed and domain-specific list of 176 data quality attributes that they reduced to 15 generic parameters that cover almost everything that most people want. Not all of the parameters are applicable or measurable for every type of data, but that in itself is not a problem. What is important is that it is relatively straightforward to understand what precision means for a particular type of measurement in a particular dataset. Timeliness – the extent to which the data is contemporary – is measured in microseconds in some disciplines and in decades in others. But the issue for quality is whether I can get the data soon enough for the purpose I have in mind.

3. HOW DO WE COMMUNICATE QUALITY?

As I asserted above, in general we don't communicate either assertions of quality (as producers or distributors of data) or requirements of quality (as consumers.) But in general, if asked, we know what we want and we know what we can provide. Given the simple dimensions of Wang and Strong we can do so in a way that is comprehensible to those outside our domain of expertise.

When we do state things about data quality we tend to do so in ways that are domain-specific and highly textual rather than machine-readable. The description of one dataset may tell us which scientific instruments were used to take which measurements and perhaps even tell us more about the properties of the instruments – their age, or accuracy, or some sensor profile. Another dataset may tell us about experimental methods used or the laboratory in which the experiments took place. Yet another will describe the processed used to transform raw observational data into cleaned data which is made available for reuse. All of these are making assertions about provenance, and possible about other qualities such as accuracy and precision. They are doing so in ways that make it difficult to realize that this is the case.

Another way in which even the same quality assertions can be made differently relates to whether a process is documented or the intended outcomes. It is common in some disciplines to 'clean' data to spot obvious outliers and erroneous measurements or transcriptions of measurements. One dataset may say which software was used to do this cleaning and what parameters were applied. Another will simply document the intended results without saying how they were achieved.

Even if we manage to agree on generic words to describe one aspect of quality we may struggle to assign a measure to that aspect that has the same meaning in different contexts, and the work of Wang and Strong is of less help here. 'Comprehensiveness' is one example; the word 'coverage' is often used to mean a similar thing in many disciplines. But it isn't enough to say that a census, for example, has 98% coverage. Do we mean that we covered 98% of the population? Of the area being surveyed? Of the households in that area? Any social scientist would be very specific about which of those they meant. But they might then have difficulty looking for data on a topic like air quality and recognizing whether '98% coverage' for the air quality dataset meant the same thing.

In some cases we need to set bounds on particular quality parameters and say that we cannot use data which falls outside them. In others we may not set bounds but we need to know that the parameter has been measured in some way. You may need data that is precise to 4 significant figures; I can work with data of almost any degree of precision, but I do need to know what that precision is. These are different types of request and assertion about data quality and our systems need to be able to accommodate both.

Some of these things matter less when human beings are examining textual descriptions of an individual dataset to determine whether it meets their research needs. They matter more when we are dealing with very large collections of potential datasets or integrating data from many sources. At these times automated systems will need to make some or all of the decisions about whether data is of acceptable quality for our purposes.

4. ONE DATASET, MULTIPLE MANIFESTATIONS

In some cases it can be useful to provide multiple versions of a single dataset with different quality attributes. Some examples may be choosing between a dataset which is complete but partially inaccurate (because some measurements were flawed) and one which is accurate but incomplete, because the flawed measurements were removed. In other cases some users may need early access to raw, inaccurate data of unknown completeness and precision whereas others can wait for one or more of those parameters to be improved or measured.

This is not how many data workflows and lifecycles operate at present, particularly in the world of research. (Some commercial data providers, by contrast, do recognise the different needs of different markets and have multiple products for them.) Most of our systems, with some exceptions, apply a particular set of quality controls to each data set before finally making it available for reuse. Sometimes that's enough, but it often isn't. Data centres that take years to make data available because of the intense quality control processes they apply and the resource-intensiveness of those processes are only satisfying the users who can deal with those delays. It's possible that there are other research or non-research applications which might require access in minutes, hours or days. At present we don't tend to discover these requirements and it isn't easy for anyone to express them. We also can't always say if it is even possible to satisfy them.

Yet a world in which a single dataset can be available in different forms, at different times or costs, for different use cases is likely to be one where more research is possible than the present one.

.

5. TOOLS AND SKILLS

A number of tools already exist to deal with aspects of data quality. Some are designed to test whether data meets certain quality measures and some are designed to improve some of those measures. Many don't come from the world of research and those that do are aimed at specialist audiences even when they potentially have wider applicability.

Many tools will always remain highly domain-specific. An example is the suite of tools employed by journals of the International Union of Crystallographers as described in [McMahon]. In later presentations, McMahon has described how these tools have been used to spot fraudulent science by looking for characteristics of the data that appear only in unmodified experimental measurements. The techniques are highly specific to the instruments and the experiments performed, but these are highly uniform for these crystallographic journals.

But commercial data processing makes extensive use of more generic data cleaning methods which are applied to a wide variety of data sets and the development of these is an active topic of study amongst database specialists (see e.g. [Jia]). Some national initiatives archiving government data developed similar generic tools to both characterise data (spotting coded fields, for instance) and to identify certain error classes. AERIC, developed at the USA's NARA and chkdata which performed a similar task for the UK's NDAD are examples of such tools. Although the checks performed are relatively simplistic such toolchains were designed to work at scale and to produce both machine-readable and human-readable results. It's likely that they could have wider applicability for research data and it also seems likely that tools designed to perform certain types of check in specific disciplines could have much wider applicability also.

The same can be said of the skills that data managers need in managing data quality. Unfortunately many data managers currently receive their training in the context of specific disciplines and hence may not be aware that their knowledge could be of much wider use. The experience of the Digital Curation Centre in the development of DC101 (a generic training suite intended primarily for data managers) in 2008 shows that the management of data quality, amongst other skills, is something that can be taught in a highly generic fashion. This produces data managers who have greater potential job mobility and who are more able to deal with the demands presented by multi-disciplinary research.

6. FUTURE POSSIBILITIES

I believe that, based on what is said above, a number of potential gains can be realized with relatively modest effort and little additional research. Some others may require more in-depth research to be realized but are still near-term rather than long-term goals. Certain actions are required to achieve these gains which I list below:

- encourage agreement on generic data quality attributes, based on the work of Wang and Strong or a similar set;
- Devise mappings from existing discipline-specific quality attributes to the generic ones. In the process, identify which attributes are unimportant or cannot be measured in these disciplines;
- Express these attributes clearly in machine-readable form for datasets made available for reuse and develop mechanisms to render this in human-readable forms;
- Enhance existing dataset discovery mechanisms to allow the existence of a quality assertion or specific values assigned to an attribute to be a component of the search;
- Ensure that training for data managers considers data quality from a generic standpoint first before consider domain-specific realisations.

These measures alone will make data reuse, and specifically cross-disciplinary data reuse, much more practical than it is at present. Reuse which depends on the integration of many data sources will be made more reliable if one can have automated reassurance that the sources have compatible quality attributes, or alternatively that the quality attributes are quantified in a way that allows the combination to be carried out in a reliable fashion, compensating for any differences.

They will also allow better consideration of which attributes should be the focus of which uses, and ease the development of a data infrastructure in which multiple versions of data with different qualities can be made available at lower risk than is the case today.

Furthermore, a greater cohort of data managers and researchers who have a more generic understanding of data quality will lead to greater reuse of tools between disciplinary areas and the development of new tools which have wider applicability from the outset.

Some areas present more difficulty. I've said already that it is one thing to agree on a generic quality parameter, but quite another to agree on an appropriate quantifiable measure for it. For some, such as provenance, numeric measures are not appropriate and we can already identify good progress in generic means to express them. For measures such as provenance we also enter the potentially difficult area of quality parameters for quality parameters – how complete is the provenance information, for example, or how accurate ?

Getting everyone to be more specific about their requirements and their assertions is an eminently achievable goal, however, and I am confident that this workshop can make considerable progress in defining how this is to be done.

7. ACKNOWLEDGMENTS

I am grateful to Suzanne Embury who first drew my attention to the work of Wang and Strong and its potential for training data managers in generic techniques for managing data quality when we taught together on the Digital Curation Centre's DC101 course.

I would also like to acknowledge the encouragement and support provided by colleagues on the GRDI2020 project and the opportunity it gave me to clarify and publish my thoughts on possible future infrastructures for managing and communicating data quality in a cross-disciplinary fashion. The GRDI2020 project was funded by the European Commission under its FP7 programme.

8. REFERENCES

[1] Ashley, K. 2011. *Data Quality and Curation* Technical report of the GRDI2020 project. Available from http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=201287b9-9f1c-4626-8b62-44cfe418707a

[2] The GRDI2020 project http://www.grdi2020.eu/

[3] Jia, X. 2008 “From Relations to XML: Cleaning, Integrating and Securing Data” (2008) http://hdl.handle.net/1842/3161

[4] McMahon, B. 1996. The role of journals in maintaining data integrity: Checking of crystal structure data in Acta Crystallographica *J. Res. Natl. Inst. Stand. Technol*. 101 ,3 (May 1996)

[5] Wang, R.Y. and Strong, D.M. 1996 Beyond accuracy: What data quality means to data consumers *Journal of Management Information Systems*; Available from: http://web.mit.edu/tdqm/www/tdqmpub/beyondaccuracy_files/beyondaccuracy.html

Error Metrics for Large-Scale Digitization

Paul Conway
University of Michigan
4427 North Quad
105 S. State Street
Ann Arbor, MI 48109
+1 734 615-1419
pconway@umich.edu

Jacqueline Bronicki
University of Michigan
320 Hatcher North
Ann Arbor, MI 48109
+1 734 764-8742
bronick@umich.edu

ABSTRACT

The paper summarizes the methodology utilized in an ongoing project that is exploring quality issues in the large-scale digitization of books by third-party vendors – such as Google and the Internet Archive – that are preserved in the HathiTrust Digital Library. The paper describes the research foundation for the project and the model of digitization error that frames the data gathering effort. The heart of the paper is an overview of the metrics and methodologies developed in the project to apply the error model to statistically valid random samples of digital book-surrogates that represent the full range of source volumes digitized by Google and other third party vendors. Proportional and systematic sampling of page-images within each 1,000-volume sample produced a study set of 356,217 page images. Using custom-built web-enabled database systems, teams of trained coders have recorded perceived error in page-images on a severity scale of 0-5 for up to eleven possible errors. The paper concludes with a summary of ongoing research and the potential for future research derived from the present effort.

Categories and Subject Descriptors

I.4.1 [**Image Processing & Computer Vision**]: Digitization and Image Capture

General Terms

Measurement, Verification

Keywords

digitization quality; error model; Google Books; HathiTrust

1. INTRODUCTION

From Project Gutenberg to Google Books, the large-scale digitization of books and serials is generating extraordinary collections of intellectual content that are transforming the way society reads and learns. Questions are being raised, however, regarding the quality and usefulness of digital surrogates produced by third-party vendors and deposited in digital repositories for preservation and access. For such repositories and their communities of users to trust digital documents, repositories must validate the quality of these objects and their fitness for the uses envisioned for them. Information quality should be an important component of the value proposition that digital preservation repositories offer their stakeholders and users. [4]

The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. The literature on information quality, however, is relatively silent on how to measure quality attributes of very large collections of digitized books and journals, created as a combination of page images and full-text data by third party vendors. Lin [10] provides an excellent review of the state of digital image analysis (DIA) research within the context of large-scale book digitization projects and establishes a "catalog of quality errors," adapted from Doermann. [8] His research is most relevant because it distinguishes errors that take place during digitization [e.g., missing or duplicated pages, poor image quality, poor document source] from those that arise from post-scan data processing [e.g., image segmentation, text recognition errors, and document structure analysis errors]. Lin recognizes that, in the future, quality in large-scale collections of books and journals will depend on the development of fully automated analysis routines, even though quality assurance today depends in large measure upon manual visual inspection of digitized surrogates or the original book volumes. [9]

Quality judgments are by definition subjective and incomplete. From the perspective of users and stakeholders, information quality is not a fixed property of digital content (Conway 2009). Tolerance for error may vary depending upon the expected uses for digitized books and journals. Marshall argues that "the repository is far less useful when it's incomplete for whatever task the user has in mind." [11, p. 54] Baird makes the essential connection between quality measurement and expected uses in articulating the need for research into "*goal directed metrics* of document image quality, tied quantitatively to the reliability of downstream processing of the images." [2, p. 2] Certain fundamental, baseline capabilities of digital objects span disciplinary boundaries and can be predicted to be important to nearly all users. [7] Use-cases articulate what stakeholders and users might accomplish if digital content was validated as capable of service-oriented functions. [6] Individual users construct scenarios that articulate their requirements for digital content. [1]

For this research project, we define quality as the absence of errors in scanning and post-scan processing relative to expected uses. [5] Within the context of a large-scale preservation repository, the research adapts Stvilia's [12] model of intrinsic quality attributes and Lin's [10] framework of errors in book surrogates derived from digitization and post-scan processing. The overall design of the three-year research project consists of three overlapping investigative phases. Phase one defines and tests a set of error metrics (a system of measurement) for digitized books and journals. Phase two applies those metrics to produce a set of statistically valid measures regarding the patterns of error (frequency and severity) in multiple samples of volumes drawn from strata of HathiTrust. Phase three (ongoing) will engage stakeholders and users in building, refining, and validating the use-case scenarios that emerge from the research findings.

The research project utilizes content deposited in the HathiTrust Digital Library, which is a digital preservation repository launched in October 2008 by a group of research universities, including the Committee on Institutional Cooperation [the Big Ten universities and the University of Chicago] and the University of California system. At present [August 2012] HathiTrust consists of 10.4 million digitized volumes ingested from multiple digitization sources (primarily Google). HathiTrust is supported by base funding from its 66 institutional partners, and its governing body includes top administrators from libraries and information offices at investing institutions. [13][14] HathiTrust is a large-scale exemplar of a preservation repository containing digitized content; 1) with intellectual property rights owned by a variety of external entities; 2) created by multiple digitization vendors for access; and 3) deposited and held/preserved collaboratively. The findings of the research are broadly applicable to the challenges in duplication, collection development, and digital preservation that are common to all digital libraries.

2. ERROR MODEL

A three-tiered hierarchical error typology and associated value definitions are the keystones of the study. The error model **(Figure 1)** identifies error at the data, page, and volume levels and establishes hypotheses regarding the cause of each error (source, scanning, post-scan manipulation). Data and page-image errors are individually identifiable errors that affect the visual appearance of single bitmap pages. A particular error may be confined to a single page or repeated across a sequence in a volume. Whole volume-level errors apply to structural issues surrounding the completeness or accuracy of the volume as a whole, such as missing pages, duplicate pages, and ordering of pages. The development process for the error model was deeply iterative and involved substantial testing of individual error items and the meaning of narrative error definitions. The goal was to create a validated error model with clearly defined errors that could be repeatedly and consistently identified by coding staff in multiple settings.

2.1 Sources of error

The error model implies causality regarding one of three factors: the physical qualities of the source volume, the cluster of scanning activities that create a master bitmap image of two pages in an open book, and the suite of post-scan manipulation processes that produce the final deliverable image that users consult. One of the primary objectives of the data collection process is to gather data on errors without assuming the cause of error. Coders were instructed to “code what you see” rather than speculate on the cause of error.

2.2 Severity of error

The research team developed a severity scale for each of the eleven page-image errors to capture a more granular rating of each error. In order to train coding staff to uniformly assign severity, the team outlined four main definitions for coders to reflect upon when assigning severity: original content, error, reading ability, and inference. *Original Content* is defined as the text or image content on the page created through the original printing process. Original content excludes marginalia, annotations, and other library-added content (bar codes, call numbers, book plates, circulation aids) added by users after the acquisition of the volume by the library. *Error* is defined as variations from the expected appearance of Original Content. *Reading ability* is designated as the ability of a reviewer to interpret the letters, illustrations, and other information contained in the Original Content of a page. *Inference* is the degree to which an average reviewer cannot detect Original Content, but must use contextual information to determine letters, words, or other information that compose the Original Content. Using this understanding, the coder is expected to apply a level of severity from zero to five for all errors detected on the page upon review. **Figure 2** displays the operative severity scale used by the 12 part-time coders working in teams at the University of Michigan and the University of Minnesota.

3. METRICS FOR DIGITIZATION ERROR

The research hypothesizes a state of image and text quality in which digitized book and serial benchmark-volumes from a given vendor are sufficiently free of error such that these benchmark-surrogates can be used nearly universally within the context of specific use-case scenarios. In the development phase, the research explored how to specify the gap between benchmark and digitized volumes in terms of detectable error. The project developed a highly reliable and statistically sound data gathering and analysis system to measure error-incidence in HathiTrust volumes. The research team focused initially on sampled page-images within a digitized volume, followed by physical review of sampled volumes, and culminating with a whole volume review of the same sampled volumes. The scope of the project included review of 356,217 individually sampled pages from four distinctive samples, plus a second-stage review of entire volumes totaling 691,972 page-images.

3.1 Page-level data collection

A key component of our study is efficient coding of each digital page-image with an easy to use web application **(Figure 3).** The project built a highly efficient web-based application that could be used in multiple remote locations. The web application has a user interface that populates to a backend database with complex controls to minimize data entry error. The database records all coded values per sequence number relating to a unique volume, identified by a unique HathiTrust ID.

3.2 Physical book inspection

To supplement the data gathered on page-level and whole volume errors, the research team designed a process for inspecting physical volumes and correlating material and bibliographic characteristics with detected errors. A physical review of each sampled volume was conducted by current UMSI students. The physical review model was developed by the principal investigator based on prior standards and variables used by the preservation community to review physical volumes for damage and deterioration. The independent variables and their values were crafted into a brief online questionnaire and student volunteers were trained to identify and capture physical characteristics of the volume under the supervision of the principal investigator. The survey featured 11 questions regarding the quality of the book as a whole, 12 bibliographic data fields to be confirmed by the reviewer, and 4 metadata fields populated by the project programmer.

The project programmer created a stand-alone web-based interface designed with efficiency and mobility as the central features **(Figure 4).** The interface connects to a backend SQL database where a unique identifier could be used to map data gathered in physical inspection to page-level and whole volume error data. Reviewers were able to access the interface from various locations through a secure internet connection after they were authenticated by the system.

3.3 Whole volume error

The error model identifies five distinct whole volume error categories related to scanning and processing of digital volumes that relate to completeness and integrity of the volume. The five major binary error types are: missing page(s), duplicate page (s), out of order page(s), false page(s), and fully obscured page(s). No severity level is assigned to whole volume as the condition either exists or it does not.

A secure web-based application **(Figure 5)** has been developed to capture error coding at the whole volume level. All coded errors are captured in a central database for statistical analysis. To control for HathiTrust interface effects, the application was designed to have a minimalist thumbnail view interface while maximizing data collection efficiency. Each data coder is authenticated using unique ID and login, thus allowing the detailed logging of coding activity. The coder has access to an entire volume as sequenced in HathiTrust along with relevant metadata to enhance the ability to code error. The coder inspects several parts of the digital image as well as aspects of vendor-supplied metadata to determine if an error exists: page number as seen in digital image, page number as provided by vendor in the metadata, context of the text from page to page, and context of the volume as a whole.

4. APPLYING THE METRICS

4.1 Representativeness (two tier sampling)

The purpose of sampling is to gather a representative group of volumes to test and refine the error definition model and to make projections about error in a given strata population. The issue of representativeness was addressed in the sampling techniques applied during data collection phase. Under direction from the team statistician, the programmer developed a systematic random sampling algorithm to pull random samples from the HathiTrust Library with pre-determined sample parameters. The project co-PI, who is a distinguished scholar of statistical process control, determined that 1,000 volumes would be representative of sampling pools within HathiTrust and would allow for statistical comparison of sub-populations with small frequencies in important variables.

Within each 1,000 volume sample, the project team extracted a systematic random sample of 100 pages within each volume to predict the distribution of error within the volume as a whole. The sampling algorithm is applied to the image sequence number, the complete set of which serves as a proxy for the total number of pages in a given volume, cover to cover. The algorithm divides the total number of images within a volume by one hundred to establish a number that determines the sequential sampling interval value. A random number generator establishes where in the volume (between sequence number 1 and 10) to begin sequential sampling. This method ensures that the sample will be representative of the images at the front and ends of the volumes. Sequential sampling then selects pages according to the sampling interval value, rounded up or down accordingly, to determine which whole-sequence-number image should be chosen.

4.2 Data reliability and tests of significance

The research adapts analytical procedures designed to diagnose and address the challenge of detecting and adjusting for the fact that two human beings will see and record the same information inconsistently. The presence of significant levels of inter-coder inconsistency generates error in the statistical evaluation of the findings of quality review undertaken by multiple reviewers in a distributed review environment. One error review procedure entails multiple reviewers coding the severity of errors in the same volumes. Collapsing severity to a two-point scale (severe/not) allows for the testing of the null hypothesis that the pairs of reviewers code error severity in the same way, using Cohen's Kappa statistic as a measure of agreement. Similar tests assessing the frequency of errors detected utilize the Chi Square test of significance. The outcome of these analyses supports improved training of coders and establish the lower threshold of coding consistency in a distributed review environment.

4.3 Data gathered in the study

The project team established two data gathering teams, one group of four part time staff at the University of Minnesota and another group of between four and eight part time staff. The Project Manager developed training materials and a training routine to establish a consistent pattern of review behavior. **Table 1** displays the total number of volumes and pages reviewed by the combined coding teams and estimates the size of the populations represented by the random samples.

5. ONGOING RESEARCH

5.1 Cost of manual inspection

The Project Coordinator tracked very closely the expenditure of time and resources by paid coding staff. Additionally, the web-based review systems recorded the time spend by individual coders on page level and volume level review. This data will be processed to yield an assessment of the total cost of manual review processes as well as a comparison of the cost of the separate approaches to quality review (page-level versus whole volume).

5.2 Validating results from users

Ongoing research with two populations of users of digitized volumes seeks to validate the statistical findings with end-user needs and expectations. The two populations of study are digital humanities scholars (faculty and doctoral students), whose research requires close reading of published books; and library collection development staff who expect to use digitized volumes as replacements for or surrogates of physical volumes. The goal of the research is to identify needs-based thresholds of acceptance of detected error.

5.3 Potential for automatic error detection

Findings from page-level and volume level error will yield a prioritized list of scanning and post-scan procedures that result in error. Future research will explore the extent to which the most frequent and the most offending errors can be detected and corrected using image processing algorithms. Preliminary research has identified potentially valuable processing procedures for duplicate page images, and for warped or skewed page images. Fixing text anomalies might also be possible in certain cases. The challenges of correcting scanning artifacts in book illustrations are more problematical.

5.4 Tagging and rating error

A supplemental goal of the project is to address a priority need within the HathiTrust community of stakeholders: namely a tool for the efficient review of individual volumes on demand and the rating of these volumes in terms of the presence or absence of critically important errors. This work is ongoing and will become one of the principal deliverables of the grant project.

6. ACKNOWLEDGMENTS

Project planning supported by the Andrew W. Mellon Foundation; project implementation supported by a grant from the Institute for Museum and Library Services [LG-06-10-0144-10]. We wish to thank Ryan Rotter, Systems Programmer; Ken Guire, Statistician; Jeremy York, Associate Librarian; and co-PI Edward Rothman, Professor of Statistics, for their indispensible work on the project.

7. REFERENCES

[1] Alexander I. F. & Maiden N.A.M., eds. (2004). *Scenarios, Stories and Use Cases.* New York: John Wiley.

[2] Baird, H. (2004). "Difficult and Urgent Open Problems in Document Image Analysis for Libraries," *Proc. Of International Workshop on Document Image Analysis for Libraries*? (?): 25-32.

[3] Conway, P. (2009). "The Image and the Expert User." *Proceedings of IS&T's Archiving 2009*, Imaging Science & Technology, Arlington, VA, May 4-7, pp. 142-50.

[4] Conway, P. (2010). "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly* 80 (1): 61-79.

[5] Conway, P. (2011). "Archival Quality and Long-term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science* 11 (3): 293-309. [DOI: 10.1007/s10502-011-9155-0]

[6] Cockburn, A. (2000). *Writing Effective Use Cases.* Boston: Addison-Wesley.

[7] Crane, G. and Friedlander, A. (2008). *Many More than a Million: Building the Digital Environment for the Age of Abundance. Report of a One-day Seminar on Promoting Digital Scholarship, November 28, 2007.* Washington, D.C.: Council on Library and Information Resources.

[8] Doermann, D., Liang, J., and Li, H. (2003). "Progress in Camera-Based Document Image Analysis." *Proc. Seventh International Conference on Document Analysis and Recognition (ICDAR'03),* 3 (6): 606-616.

[9] Le Bourgeois, et al. (2004). "Document Images Analysis Solutions for Digital Libraries." *Proceedings of the First International Workshop on Document Image Analysis for Libraries* (DIAL'04), 23-24 Jan., Palo Alto, California, pp. 2-24.

[10] Lin, X. (2006). "Quality Assurance in High Volume Document Digitization: A Survey." *Proceedings of the Second International Conference on Document Image Analysis for Libraries* (DIAL'06), 27-28 April, Lyon, France, pp. 319-326.

[11] Marshall, C. C. (2003). "Finding the Boundaries of the Library without Walls." In. Bishop, A., et al. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Cambridge: MIT Press, pp. 43-64.

[12] Stvilia, B., et al. (2007). "A Framework for Information Quality Assessment." *Journal of the American Society for Information Science and Technology* 58 (12): 1720-1733.

[13] York, J. J. (2009). This library never forgets: Preservation, cooperation, and the making of HathiTrust digital library. *Proc. IS&T Archiving 2009*, Arlington, VA, pp. 5-10.

[14] York, J. J. (2010). Building a future by preserving our past: The preservation infrastructure of HathiTrust digital library." *76th IFLA General Congress and Assembly, 10-15 August*, Gothenberg, Sweden.

.

Level of Abstraction	*Possible Cause of Error*
LEVEL 1: DATA/INFORMATION	
1.1 Text: thick text [fill, excessive]	Source or post-processing
1.2 Text: broken text [character breakup]	Source or post-processing
1.3 Illustration: scanner effects [moiré patterns, gridding]	Scanning or post-processing
1.4 Illustration: tone, brightness, contrast	Scanning, post-processing, or source
1.5 Illustration: color imbalance, gradient shifts	Scanning, post-processing, or source
LEVEL 2: ENTIRE PAGE	
2.1 Blur [distortion]	Scanning or source
2.2 Warp [text alignment]	Post-processing
2.3 Skew [page alignment]	Scanning, post-processing, or source
2.4 Crop [gutter, text block]	Source or post-processing
2.5 Obscured [portions not visible]	Scanning or post-processing
2.6 Colorization [text bleed, low contrast]	Source or post-processing
LEVEL 3: WHOLE VOLUME	
3.1 Fully obscured [foldouts]	Scanning
3.2 Missing pages [one or more]	Original source or scanning
3.3 Duplicate pages [one or more]	Original source or scanning
3.4 Order of pages	Original source or scanning
3.5 False pages [not part of original content]	Scanning or post-processing

Figure 1. Model of error in large-scale digitization

0 - Default - Error is undetectable on the page.

1 - Error exists but has a negligible effect on the Original Content.

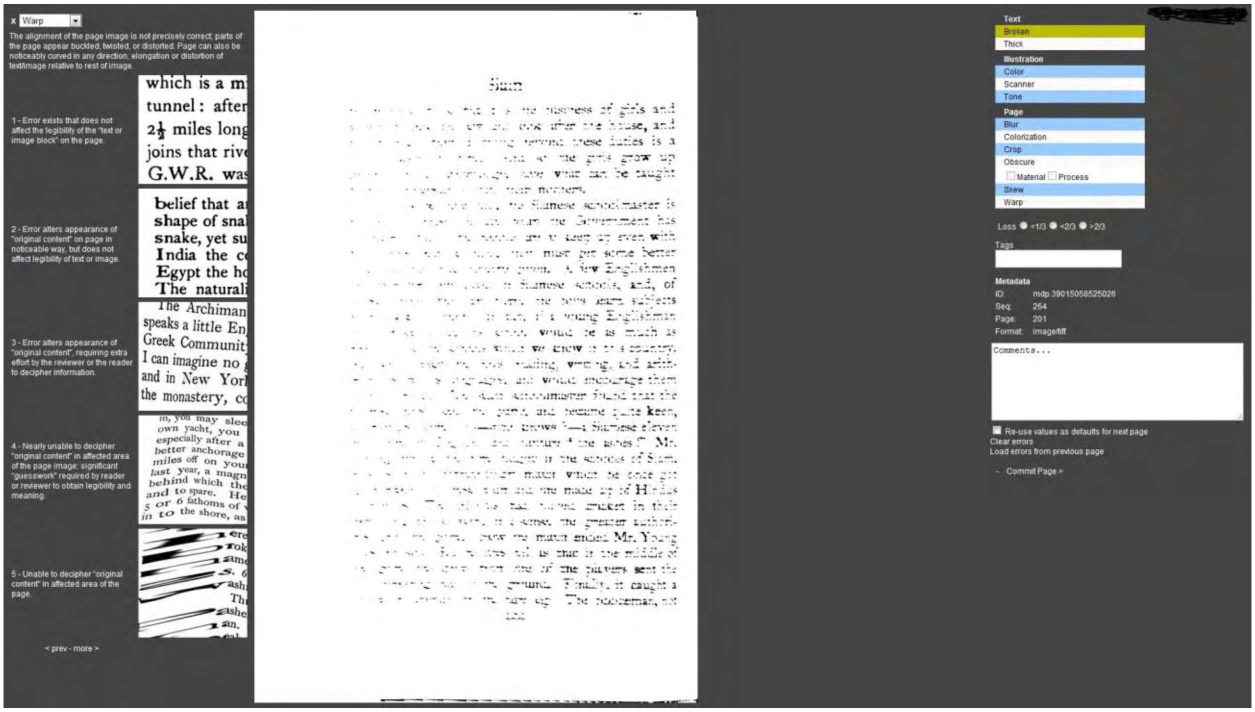
2 - Error clearly alters appearance of Original Content, but has a negligible effect on reading ability.

3 - Error clearly alters appearance of Original Content and has a clear negative impact on reading ability.

4 - Nearly unable to decipher Original Content in affected area of the page; significant inference required by reviewer to obtain legibility and meaning.

5 - Original Content in affected area of the

Figure 2. Severity scale



x Warp

The alignment of the page image is not precisely correct; parts of the page appear buckled, twisted, or distorted. Page can also be noticeably curved in any direction; elongation or distortion of text/image relative to rest of image.

1 - Error exists that does not affect the legibility of the "text or image block" on the page.

which is a m
tunnel : after
2½ miles long
joins that riv
G.W.R. was

2 - Error alters appearance of "original content" on page in noticeable way, but does not affect legibility of text or image.

belief that a
shape of sna
snake, yet su
India the c
Egypt the h
The naturali

3 - Error alters appearance of "original content", requiring extra effort by the reviewer or the reader to decipher information.

The Archiman
speaks a little En
Greek Communit
I can imagine no
and in New Yor
the monastery, c

4 - Nearly unable to decipher "original content" in affected area of the page image; significant "guesswork" required by reader or reviewer to obtain legibility and meaning.

in, you may slee
own yacht, you
especially after a
better anchorage
miles off on you
last year, a magn
behind which the
and to spare. He
5 or 6 fathoms of
in to the shore, as

5 - Unable to decipher "original content" in affected area of the page.

ere
rok
ame
s. 6
ash
Th
ashe
an.

< prev - more >

Siam

[illegible]

[redacted]

Text
- Broken
- Thick

Illustration
- Color
- Scanner
- Tone

Page
- Blur
- Colorization
- Crop
- Obscure
  - □ Material □ Process
- Skew
- Warp

Loss ● <1/3 ● <2/3 ● >2/3

Tags

Metadata

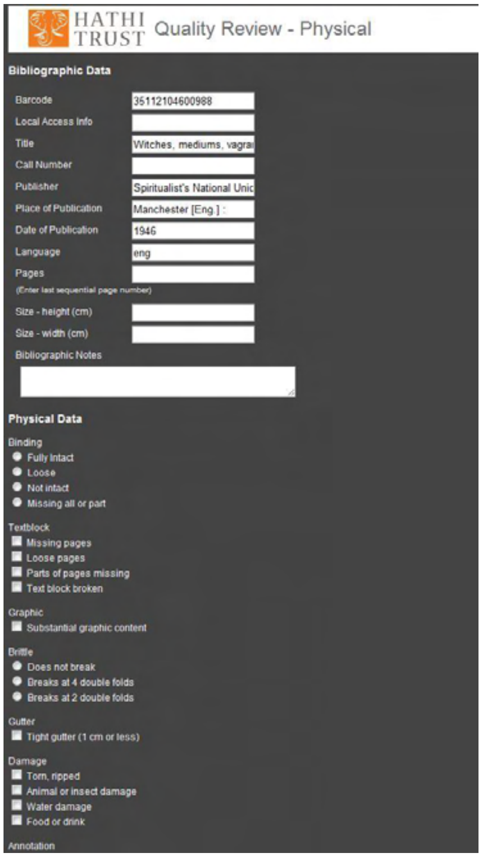
ID:	mdp.39015058525026
Seq:	264
Page:	201
Format:	image/tiff

Comments...

□ Re-use values as defaults for next page
Clear errors
Load errors from previous page

- Commit Page >

Figure 3. Interface for coding page-level error



HATHI TRUST Quality Review - Physical

Bibliographic Data

Barcode	35112104600988
Local Access Info	
Title	Witches, mediums, vagra
Call Number	
Publisher	Spiritualist's National Unic
Place of Publication	Manchester [Eng.] :
Date of Publication	1946
Language	eng
Pages (Enter last sequential page number)	
Size - height (cm)	
Size - width (cm)	
Bibliographic Notes	

Physical Data

Binding
- ○ Fully Intact
- ○ Loose
- ○ Not intact
- ○ Missing all or part

Textblock
- □ Missing pages
- □ Loose pages
- □ Parts of pages missing
- □ Text block broken

Graphic
- □ Substantial graphic content

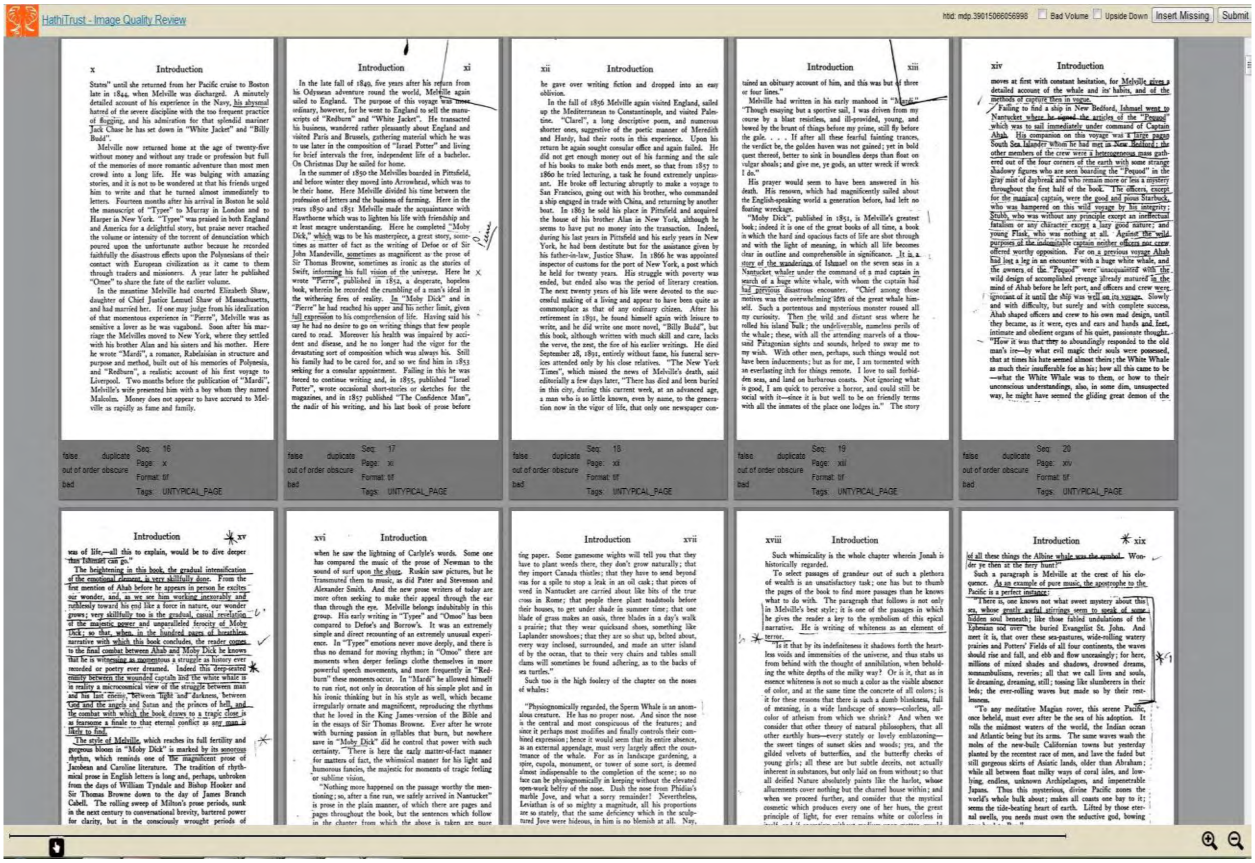
Brittle
- ○ Does not break
- ○ Breaks at 4 double folds
- ○ Breaks at 2 double folds

Gutter
- □ Tight gutter (1 cm or less)

Damage
- □ Torn, ripped
- □ Animal or insect damage
- □ Water damage
- □ Food or drink

Annotation

Figure 4. Physical review interface (partial view)



x Introduction

States" until she returned from her Pacific cruise to Boston late in 1844, when Melville was discharged. A minutely detailed account of his experience in the Navy, his abysmal hatred of the severe discipline with the too frequent practice of flogging, and his admiration for that splendid mariner Jack Chase he has set down in "White Jacket" and "Billy Budd".

Melville now returned home at the age of twenty-five without money and without any trade or profession but full of the memories of more romantic adventure than most men crowd into a long life. He was bulging with amazing stories, and it is not to be wondered at that his friends urged him to write and that he turned almost immediately to letters. Fourteen months after his arrival in Boston he sold the manuscript of "Typee" to Murray in London and to Harper in New York. "Typee" was praised in both England and America for a delightful story, but praise never reached the volume or intensity of the torrent of denunciation which poured upon the unfortunate author because he recorded faithfully the disastrous effects upon the Polynesians of their contact with European civilization as it came to them through traders and missioners. A year later he published "Omee" to share the fate of the earlier volume.

In the meantime Melville had courted Elizabeth Shaw, daughter of Chief Justice Lemuel Shaw of Massachusetts, and had married her. If one may judge from his idealization of that momentous experience in "Pierre", Melville was as sensitive a lover as he was vagabond. Soon after his marriage the Melvilles moved to New York, where they settled with his brother Alan and his sisters and his mother. Here he wrote "Mardi", a romance, Rabelaisian in structure and purpose and method, built out of his memories of Polynesia, and "Redburn", a realistic account of his first voyage to Liverpool. Two months before the publication of "Mardi", Melville's wife presented him with a boy whom they named Malcolm. Money does not appear to have accrued to Melville as rapidly as fame and family.

false duplicate
out of order obscure
bad

Seq: 16
Page: x
Format: tif
Tags: UNTYPICAL_PAGE

Introduction xi

In the late fall of 1849, five years after his return from his Odyssean adventure round the world, Melville again sailed to England. The purpose of this voyage was more ordinary, however, for he went to England to sell the manuscripts of "Redburn" and "White Jacket". He transacted his business, wandered rather pleasantly about England and visited Paris and Brussels, gathering material which he was to use later in the composition of "Israel Potter" and living for brief intervals the free, independent life of a bachelor. On Christmas Day he sailed for home.

In the summer of 1850 the Melvilles boarded in Pittsfield, and before winter they moved into Arrowhead, which was to be their home. Here Melville divided his time between the profession of letters and the business of farming. Here in the years 1850 and 1851 Melville made the acquaintance with Hawthorne which was to lighten his life with friendship and at least meagre understanding. Here he completed "Moby Dick," which was to be his masterpiece, a great story, sometimes as matter of fact as the writing of Defoe or of Sir John Mandeville, sometimes as magnificent as the prose of Sir Thomas Browne, sometimes as ironic as the stories of Swift, informing his full vision of the universe. Here he wrote "Pierre", published in 1852, a desperate, hopeless book, wherein he recorded the crumbling of a man's ideal in the withering fires of reality. In "Moby Dick" and in "Pierre" he had reached his upper and his nether limit, given full expression to his comprehension of life. Having said his say he had no desire to go on writing things that few people cared to read. Moreover his health was impaired by accident and disease, and he no longer had the vigor for the devastating sort of composition which was always his. Still his family had to be cared for, and so we find him in 1853 seeking for a consular appointment. Failing in this he was forced to continue writing and, in 1855, published "Israel Potter", wrote occasional short-stories or sketches for the magazines, and in 1857 published "The Confidence Man", the nadir of his writing, and his last book of prose before

Pierre

false duplicate
out of order obscure
bad

Seq: 17
Page: xi
Format: tif
Tags: UNTYPICAL_PAGE

xii Introduction

he gave over writing fiction and dropped into an easy oblivion.

In the fall of 1856 Melville again visited England, sailed up the Mediterranean to Constantinople, and visited Palestine. "Clarel", a long descriptive poem, and numerous shorter ones, suggestive of the poetic manner of Meredith and Hardy, had their roots in this experience. Upon his return he again sought consular office and again failed. He did not get enough money out of his farming and the sale of his books to make both ends meet, so that from 1857 to 1860 he tried lecturing, a task he found extremely unpleasant. He broke off lecturing abruptly to make a voyage to San Francisco, going out with his brother, who commanded a ship engaged in trade with China, and returning by another boat. In 1863 he sold his place in Pittsfield and acquired the house of his brother Alan in New York, although he seems to have put no money into the transaction. Indeed, during his last years in Pittsfield and his early years in New York, he had been destitute but for the assistance given by his father-in-law, Justice Shaw. In 1866 he was appointed inspector of customs for the port of New York, a post which he held for twenty years. His struggle with poverty was ended, but ended also was the period of literary creation. The next twenty years of his life were devoted to the successful making of a living and appear to have been quite as commonplace as that of any ordinary citizen. After his retirement in 1891, he found himself again with leisure to write, and he did write one more novel, "Billy Budd", but this book, although written with much skill and care, lacks the verve, the zest, the fire of his earlier writings. He died September 28, 1891, entirely without fame, his funeral services attended only by his close relatives. "The New York Times", which missed the news of Melville's death, said editorially a few days later, "There has died and been buried in this city, during this current week, at an advanced age, a man who is so little known, even by name, to the generation now in the vigor of life, that only one newspaper con-

false duplicate
out of order obscure
bad

Seq: 18
Page: xii
Format: tif
Tags: UNTYPICAL_PAGE

Introduction xiii

tained an obituary account of him, and this was but of three or four lines."

Melville had written in his early manhood in "Mardi," "Though essaying but a sportive sail, I was driven from my course by a blast resistless, and ill-provided, young, and bowed by the brunt of things before my prime, still fly before the gale. . . . If after all these fearful fainting trances, the verdict be, the golden haven was not gained; yet in bold quest thereof, better to sink in boundless deeps than float on vulgar shoals; and give me, ye gods, an utter wreck if wreck I do."

His prayer would seem to have been answered in his death. His renown, which had magnificently sailed about the English-speaking world a generation before, had left no floating wreckage.

"Moby Dick", published in 1851, is Melville's greatest book; indeed it is one of the great books of all time, a book in which the hard and opacious facts of life are shot through and with the light of meaning, in which all life becomes clear in outline and comprehensible in significance. It is a story of the wanderings of Ishmael on the seven seas in a Nantucket whaler under the command of a mad captain in search of a huge white whale, with whom the captain had had previous disastrous encounter. "Chief among those motives was the overwhelming idea of the great whale himself. Such a portentous and mysterious monster roused all my curiosity. Then the wild and distant seas where he rolled his island bulk; the undeliverable, nameless perils of the whale; these, with all the attending marvels of a thousand Patagonian sights and sounds, helped to sway me to my wish. With other men, perhaps, such things would not have been inducements; but as for me, I am tormented with an everlasting itch for things remote. I love to sail forbidden seas, and land on barbarous coasts. Not ignoring what is good, I am quick to perceive a horror, and could still be social with it—since it is but well to be on friendly terms with all the inmates of the place one lodges in." The story

false duplicate
out of order obscure
bad

Seq: 19
Page: xiii
Format: tif
Tags: UNTYPICAL_PAGE

xiv Introduction

moves at first with constant hesitation, for Melville gives a detailed account of the whale and its' habits, and of the methods of capture then in vogue.

Failing to find a ship in New Bedford, Ishmael went to Nantucket where he signed the articles of the "Pequod" which was to sail immediately under command of Captain Ahab. His companion on this voyage was a large pagan South Sea Islander whom he had met in New Bedford; the other members of the crew were a heterogeneous mass gathered out of the four corners of the earth with some strange shadowy figures who are seen boarding the "Pequod" in the gray mist of daybreak and who remain more or less a mystery throughout the first half of the book. The officers, except for the maniacal captain, were the good and pious Starbuck, who was hampered on this wild voyage by his integrity; Stubb, who was without any principle except an ineffectual fatalism or any character except a lazy good nature; and young Flask, who was nothing at all. Against the wild purposes of the indomitable captain neither officers nor crew offered worthy opposition. For on a previous voyage Ahab had lost a leg in an encounter with a huge white whale, and the owners of the "Pequod" were unacquainted with the wild design of accomplished revenge already matured in the mind of Ahab before he left port, and officers and crew were ignorant of it until the ship was well on its voyage. Slowly and with difficulty, but surely and with complete success, Ahab shaped officers and crew to his own mad design, until they became, as it were, eyes and ears and hands and feet, intimate and obedient organs of his quiet, passionate thought. "How it was that they so aboundingly responded to the old man's ire—by what evil magic their souls were possessed, that at times his hate seemed almost theirs; the White Whale as much their insufferable foe as his; how all this came to be —what the White Whale was to them, or how to their unconscious understandings, also, in some dim, unsuspected way, he might have seemed the gliding great demon of the

false duplicate
out of order obscure
bad

Seq: 20
Page: xiv
Format: tif
Tags: UNTYPICAL_PAGE

Introduction xv

seas of life,—all this to explain, would be to dive deeper than Ishmael can go."

The heightening in this book, the gradual intensification of the emotional element, is very skillfully done. From the first mention of Ahab before he appears in person he excites our wonder, and, as we see him working inexorably and ruthlessly toward his end like a force in nature, our wonder grows; very skillfully too is the gradual, casual revelation of the majestic power and unparalleled ferocity of Moby Dick; so that, when, in the hundred pages of breathless narrative with which this book concludes, the reader comes to the final combat between Ahab and Moby Dick he knows that he is witnessing as momentous a struggle as history ever recorded or poetry ever dreamed. Indeed this deep-seated enmity between the wounded captain and the white whale is in reality a microcosmical view of the struggle between man and his last enemy, between light and darkness, between God and the angels and Satan and the princes of hell, and the combat with which the book draws to a tragic close is as fearsome a finale to that eternal conflict as any man is likely to find.

The style of Melville, which reaches its full fertility and gorgeous bloom in "Moby Dick" is marked by its sonorous rhythm, which reminds one of the magnificent prose of Jacobean and Caroline literature. The tradition of rhythmical prose in English letters is long and, perhaps, unbroken from the days of William Tyndale and Bishop Hooker and Sir Thomas Browne down to the day of James Branch Cabell. The rolling sweep of Milton's prose periods, sunk in the next century to conversational brevity, bartered power for clarity, but in the consciously wrought periods of

xvi Introduction

when he saw the lightning of Carlyle's words. Some one has compared the music of the prose of Newman to the sound of surf upon the shore. Ruskin saw pictures, but he transmuted them to music, as did Pater and Stevenson and Alexander Smith. And the new prose writers of today are more often seeking to make their appeal through the ear than through the eye. Melville belongs indubitably in this group. His early writing in "Typee" and "Omoo" has been compared to Defoe's and Borrow's. It was an extremely simple and direct recounting of an extremely unusual experience. In "Typee" emotions never move deeply, and there is thus no demand for moving rhythm; in "Omoo" there are moments when deeper feelings clothe themselves in more powerful speech movements, and more frequently in "Redburn" these moments occur. In "Mardi" he allowed himself to run riot, not only in decoration of his simple plot and in his ironic thinking but in his style as well, which became irregularly ornate and magnificent, reproducing the rhythms that he loved in the King James·version of the Bible and in the essays of Sir Thomas Browne. Ever after he wrote with burning passion in syllables that burn, but nowhere save in "Moby Dick" did he control that power with such certainty. There is here the early matter-of-fact manner for matters of fact, the whimsical manner for his light and humorous fancies, the majestic for moments of tragic feeling or sublime vision.

"Nothing more happened on the passage worthy the mentioning; so, after a fine run, we safely arrived in Nantucket" is prose in the plain manner, of which there are pages and pages throughout the book, but the sentences which follow in the chapter from which the above is taken are pure

Introduction xvii

ting paper. Some gamesome wights will tell you that they have to plant weeds there, they don't grow naturally; that they import Canada thistles; that they have to send beyond seas for a spile to stop a leak in an oil cask; that pieces of weed in Nantucket are carried about like bits of the true cross in Rome; that people there plant toadstools before their houses, to get under shade in summer time; that one blade of grass makes an oasis, three blades in a day's walk a prairie; that they wear quicksand shoes, something like Laplander snowshoes; that they are so shut up, belted about, every way inclosed, surrounded, and made an utter island of by the ocean, that to their very chairs and tables small clams will sometimes be found adhering, as to the backs of sea turtles."

Such too is the high foolery of the chapter on the noses of whales:

"Physiognomically regarded, the Sperm Whale is an anomalous creature. He has no proper nose. And since the nose is the central and most conspicuous of the features; and since it perhaps most modifies and finally controls their combined expression; hence it would seem that its entire absence, as an external appendage, must very largely affect the countenance of the whale. For as in landscape gardening, a spire, cupola, monument, or tower of some sort, is deemed almost indispensable to the completion of the scene; so no face can be physiognomically in keeping without the elevated open-work belfry of the nose. Dash the nose from Phidias's marble Jove, and what a sorry remainder! Nevertheless, Leviathan is of so mighty a magnitude, all his proportions are so stately, that the same deficiency which in the sculptured Jove were hideous, in him is no blemish at all. Nay,

xviii Introduction

Such whimsicality is the whole chapter wherein Jonah is historically regarded.

To select passages of grandeur out of such a plethora of wealth is an unsatisfactory task; one has but to thumb the pages of the book to find more passages than he knows what to do with. The paragraph that follows is not only in Melville's best style; it is one of the passages in which he gives the reader a key to the symbolism of this epical narrative. He is writing of whiteness as an element of terror.

"Is it that by its indefiniteness it shadows forth the heartless voids and immensities of the universe, and thus stabs us from behind with the thought of annihilation, when beholding the white depths of the milky way? Or is it, that as in essence whiteness is not so much a color as the visible absence of color, and at the same time the concrete of all colors; is it for these reasons that there is such a dumb blankness, full of meaning, in a wide landscape of snows—colorless, all-color of atheism from which we shrink? And when we consider that other theory of natural philosophers, that all other earthly hues—every stately or lovely emblazoning—the sweet tinges of sunset skies and woods; yea, and the gilded velvets of butterflies, and the butterfly cheeks of young girls; all these are but subtle deceits, not actually inherent in substances, but only laid on from without; so that all deified Nature absolutely paints like the harlot, whose allurements cover nothing but the charnel house within; and when we proceed further, and consider that the mystical cosmetic which produces every one of her hues, the great principle of light, for ever remains white or colorless in

Introduction xix

of all these things the Albine whale was the symbol. Wonder ye then at the fiery hunt?"

Such a paragraph is Melville at the crest of his eloquence. As an example of pure music, the apostrophe to the Pacific is a perfect instance:

"There is, one knows not what sweet mystery about this sea, whose gently awful stirrings seem to speak of some hidden soul beneath; like those fabled undulations of the Ephesian sod over the buried Evangelist St. John. And meet it is, that over these sea-pastures, wide-rolling watery prairies and Potters' Fields of all four continents, the waves should rise and fall, and ebb and flow unceasingly; for here, millions of mixed shades and shadows, drowned dreams, somnambulisms, reveries; all that we call lives and souls, lie dreaming, dreaming, still; tossing like slumberers in their beds; the ever-rolling waves but made so by their restlessness.

"To any meditative Magian rover, this serene Pacific, once beheld, must ever after be the sea of his adoption. It rolls the midmost waters of the world, the Indian ocean and Atlantic being but its arms. The same waves wash the moles of the new-built Californian towns but yesterday planted by the recentest race of men, and lave the faded but still gorgeous skirts of Asiatic lands, older than Abraham; while all between float milky ways of coral isles, and low-lying, endless, unknown Archipelagoes, and impenetrable Japans. Thus this mysterious, divine Pacific zones the world's whole bulk about; makes all coasts one bay to it; seems the tide-beating heart of earth. Lifted by those eternal swells, you needs must own the seductive god, bowing

Figure 5. Interface for coding whole volume errors

Table 1. Summary of sample sizes

Sample Name	Criteria for Sample Selection	Sampling Pool Size	Number of Volumes Reviewed	Number of Pages Reviewed
Page-Level Samples				
Production Run #1	Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	1,000	93,858
Production Run #2	Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	1,000	86,439
Production Run #3	Internet Archive Digitized, Publication Date ≤ 1923, English Language, Monograph	850,000 Volumes	1,000	84,539
Production Run #4	Non-Roman Language/Script Digitized Content in HathiTrust 4 Main Language/Script Categories: Arabic, Asian, Cyrillic, Hebrew	1.29 Million Volumes	1,000	91,381
Whole Volume Error Samples				
Production Run #1a	Same Sampled Volumes from Production Run #1 Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	1,000	397,467
Production Run #2a	Same Sampled Volumes from Production Run #2 Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	1,000	294,505
Physical Review Samples				
Production Run #1b	Same Sampled Volumes from Production Run #1 Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	906	-
Production Run #2b	Same Sampled Volumes from Production Run #2 *Only University of Michigan Owned Volumes Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	584	-

Academic Libraries as Data Quality Hubs[*]

Michael J. Giarlo
Penn State University
E-017 Paterno Library
University Park, PA 16802
michael@psu.edu

ABSTRACT

Academic libraries have a critical role to play as data quality hubs on campus, based on the need for increased data quality to sustain "e-science," and on academic libraries' record of providing curation and preservation services as part of their mission to provide enduring access to cultural heritage and to support scholarly communication. Scientific data is shown to be sufficiently at risk to demonstrate a clear niche for such services to be provided. Data quality measurements are defined, and digital curation processes are explained and mapped to these measurements in order to establish that academic libraries already have sufficient competencies "in-house" to provide data quality services. Opportunities for improvement and challenges are identified as areas that are fruitful for future research and exploration.

Categories and Subject Descriptors

E.0 [**Data**]: General; H.4 [**Information Systems Applications**]: Miscellaneous; H.3.7 [**Digital Libraries**]: General

Keywords

data quality, digital curation, digital preservation, academic libraries, stewardship, e-science, research data, trust

1. SCIENTIFIC DATA AT RISK

Data quality is a pressing, not to mention costly, issue in industry; a 2002 study [16] calculated that over $600 billion per year was spent on "data quality problems" [9]. At the same time, data quality issues have become an area of growing attention within academia and academic libraries [11, 6, 14, 12], as scientific practices evolve to exploit robust campus cyberinfrastructure and as funding agencies, such as the National Science Foundation and the National Institutes of Health, increasingly require data management plans to protect and amplify the impact of their investments.

As computing costs have dwindled, computer processing speed, network throughput, and storage capacity have grown, resulting in an explosion of scientific data. Experiments, in some disciplines more than others, are producing more data than their principal investigators and research assistants can handle [4]. Due to the wealth of data that is being produced, scientific practice is changing; the gathering of data for one experiment may drive dozens or hundreds of other experiments around the world [12].

Data is more abundant than ever before, and no less important, and yet it is at risk [14, 11]. "The survival of this data is in question since the data are not housed in long-lived institutions such as libraries. This situation threatens the underlying principles of scientific replicability since in many cases data cannot readily be collected again" [11]. There are numerous examples in the literature of analog data enabling scientific inquiry decades and longer past the date it was gathered [1]; how do we as a society, and particularly we within academia, not only preserve this wealth of data for future science but ensure it is of high quality?

1.1 Curatorial Practice and Challenges

Cultural heritage organizations such as libraries and archives have been stewards of society's cultural and scientific assets for millennia, providing public access to high-quality collections, and they remain so in the Internet age. Though the activities involved are different for analog assets, "[s]tewardship of digital resources involves both preservation and curation. Preservation entails standards-based, active management practices that guide data throughout the research life cycle, as well as ensure the long-term usability of these digital re-

[*] Prepared for the NSF III #1247471 workshop, **Curating for Quality: Ensuring Data Quality to Enable New Science**, in Arlington, VA, USA

[1] Ogburn [14] cites Stephen Jay Gould's "The Mismeasure of Man" in which we learn that "analysis and critique of cranial measurements in the 1800s, twin studies in the 1950s, and the rise of IQ testing were possible because the data were still available for scrutiny and replication"

sources. Curation involves ways of organizing, displaying, and repurposing preserved data" [6].

Digital preservation and digital curation, though relatively new practices, are widely treated in the literature [12, 8, 10, 14, 11, 17, 6]. Digital curation aims to make selected data accessible, usable, and useful throughout its lifecycle. Digital curation subsumes digital preservation; without viable data, which digital preservation enables, there's nothing to be curated [2].

An oft-cited mantra on the practice of digital curation is that "curation begins before creation [of the data]" [15]. And yet, "[b]y the time knowledge in digital form makes its way to a safe and sustainable repository [such as those provided by academic libraries], it may be unreadable, corrupted, erased, or otherwise impossible to recover and use. Scientific data files may be especially endangered due to their sheer size, computational elements, reliance on and integration with software, associated visualizations, few or competing standards, distributed ownership, dispersed storage, inaccessibility, lack of documented provenance, complex and dynamic nature, and the concomitant need for a specialized knowledge base — and experience — to handle data. Data also may be endangered by the practices of scholars who regard their data as having little value beyond the confines of a small group, a specific project, or a specified period" [14].

1.1.1 *Post-Hoc Curation Considered...*

As digital curation is a new practice, and is generally centered within cultural heritage organizations (rather than within the research enterprise), *post-hoc* curation is an unfortunate fact of life; researchers lack the incentive, the resources, the time, or the expertise to curate their own data [3], and so its curation falls to other parties after the data has been created, and often after it has been "archived." For especially massive data sets, furthermore, it is difficult even to imagine, *e.g.*, a research institute or academic department having sufficient resources to curate their own data at scale.

The practice of *post-hoc* curation (vs. "sheer curation," or curation by researchers at the time of creation) is less than ideal for a number of reasons.

First, one of the goals of curation is to enable the usefulness of a digital resource over time, and one of the tactics applied is to provide sufficient context for a resource such that future users can understand what an object is, where it came from, why it is significant, and how to use it. Context is often provided via documentation, descriptive metadata, or both [6, 11, 8, 12]. The creator(s) of the data, **not** its *post-hoc* curators, are best equipped to provide this context; to get a sense of this distinction, consider the difference between the tasks of cataloging your own book collection and cataloging a complete stranger's book collection.

Second, building on the prior reason, is that *post-hoc* curation happens some time after the data have been created, possibly a long enough time to lose track of important information; capturing the context around a data set is best done while the data is still fresh in its creator's mind, *i.e.*, before or during its creation. Documentation or metadata that is created by a party other than the data's creator, especially when performed after the responsible parties have moved on to other challenges, will suffer from this lack of context.

"This [*post-hoc* curation] activity is to provide representational information and description. This is particularly problematic for academic libraries, since the data being generated at research and teaching institutions are incredibly varied. Many representational schemes for the data and metadata will be required. No one individual will have all of the required skills. Data curators will need to collaborate closely with the data providers to understand the data" [11]. Whether researchers will have sufficient time, resources, and inclination to collaborate with academic libraries on the work of curating research data at scale is yet to be seen.

Finally, possibly the most limiting reason: there is a misalignment between the scale of the need for on-campus data curation and the level of commitment by academic libraries to address this need (as measured by the amount of resources allocated to this need vs. other needs). Data curation efforts are often understaffed and underresourced, with many academic libraries devoting one full-time equivalent employee, if that, to this role, to say nothing of the level of administrative and staff support for this role.

Academic libraries, institutional will and administrative support notwithstanding, are nonetheless uniquely positioned to tackle the problem of data quality in e-science by virtue of their record of effective stewardship, their commitment to providing access to high-quality data over the long-term, and their expertise in digital preservation and digital curation practices, as "[digital] curation is a process that can ensure the quality of data and its fitness for use" [8]. It is worth examining this claim in the context of a framework for measuring data quality.

2. MEASURING DATA QUALITY

There are a number of theoretical frameworks quantifying data quality measures already established, and Knight's 2005 paper compares a selection of a dozen "widely accepted [information quality] Frameworks collated from the last decade of [information science] research" [5]. Common features are identified for data quality (or information quality), such as that it is a concept with multiple dimensions, wherein the overall quality is a function of successive indicators. Another common feature of data quality frameworks is the grouping of quality indicators into categories, classes, or levels corresponding to, *e.g.*, semiotic levels, layers of intrinsicity and extrinsicity, and the subjectivity / objectivity spectrum.

The following framework is distilled from Knight's comparison of quality frameworks, and constitutes "a series of quality dimensions which represent a set of desirable characteristics for an information resource" [8]. The framework is then applied to the domain of research data quality as viewed from my perspective, that of a digital preservation technologist

[2] This characterization of digital curation and digital preservation is a mere gloss; more may be found, for instance, on the Digital Curation Centre's website: `http://www.dcc.ac.uk/digital-curation`.

[3] Hereafter referred to as "sheer curation or curation at source" [8].

and practitioner of digital curation. It is not offered as a novel framework, nor a comprehensive one, but merely as a tool for understanding and evaluating the applicability of digital curation and preservation practices to the measure of data quality.

Trust

Evaluation of the extent to which data is trusted depends on a set of subjective factors, including whether the data is judged to be authentic, the uses to which the data is put, the subject discipline, the reputation of the party/ies responsible for the data, and the biases of the person who is evaluating the data [4].

Authenticity

Evaluation of the authenticity of data requires that data be understood. Authenticity in this context is a rough measure of the extent to which the data is judged to be "good science," answering questions pertaining to, *e.g.*, the reliability of the instruments used to gather the data; the soundness of underlying theoretical frameworks; the completeness, accuracy, and validity of the data; and ontological consistency within the data.

Understandability

Evaluation of the understandability of data requires that there be sufficient context (documentation, metadata, or provenance) describing the data, and that the data is usable.

Usability

Usability of data requires that data is discoverable and accessible; that data is in a usable file format; that the individual judging the data's quality has an appropriate tool to access the data; and that the data is of sufficient integrity to be rendered.

Integrity

Integrity of data assumes that the data can be proven to be identical, at the bit level, to some prior accepted or verified state. Data integrity may be required for usability, understandability, authenticity, trust, and thus overall quality, though this depends in part of the level of perturbation of integrity. Integrity changes will have varying effects depending on how significant the perturbation is, the file format, and where within the file the perturbation has occurred.

The relationship between the quality dimensions in this framework is analogous to that of the Semantic Web Layer Cake in that "each layer exploits and uses capabilities of the layers below" [1]. Viewed from the bottom up, this framework asserts that data integrity may be necessary but not sufficient for data quality; if the data lacks integrity, it may not be usable, and thus not understandable, authentic, or trustable — a very low measure of quality. On the other hand, unauthorized changes at the bit level may not effect the rendered data in any perceivable ways. Viewed from the top down, on the other hand, if an individual trusts a data set, she likely judges it to be of the highest quality even if it is not usable, understandable, or fixed in integrity.

3. APPLYING CURATION TO DATA QUALITY

Within the defined framework, how might the practice of curation help ensure data quality? Each of the indicators in this framework is evaluated within the context of the digital curation lifecycle [7].

3.1 Integrity

The curation lifecycle contains actions geared towards preservation of the digital asset, which includes bit-preservation via a number of possible tactics such as regular digital signature (or checksum) verification, replication, media refreshing, version management, and file-level backups. These tactics taken together should be sufficient to ensure that the data remains in the same state as originally processed. Assuming that the data was authentic to begin with [5], the effective practice of curation should provide data integrity.

3.2 Usability

Three of the seven sequential actions defined in the lifecycle model have a direct impact on the usability of data. First, the Create or Receive action [6] should include determination of an appropriate file format for the data, choosing a format that is judged to be widely accessible and preservable. The Access, Use, & Reuse action "[e]nsure[s] that data is accessible to both designated users and reusers, on a day-to-day basis", thus ensuring that the data is discoverable and made available to potential users of data. The Transform action, lastly, includes periodic evaluation of file formats and migration to new formats so data remain usable well after the original formats have been rendered obsolete.

3.3 Understandability

Context is provided for data, in order that users may understand the data, both in sequential actions within the curation lifecycle — those being Create or Receive and Preservation Action — and also within the full lifecycle action of Description and Representation Information. The generation, extraction, and application of metadata by machine agents and humans is thus a key part of the curation lifecycle, providing periodic management and addition of context to data. These actions make sure the data's purpose, impact, and provenance are established over the course of its lifecycle so that current and future users can make sense of data that they have discovered.

3.4 Authenticity and Trust

Authenticity and trust as dimensions of data quality are highly subjective. The curation process can document what instruments are used to generate data, but not how reliable

[4] Trust is a complex issue that though relevant is too far-reaching to be within the the scope of this position paper. It is nonetheless listed in the framework at the very top to establish that lower layers may be entirely discounted by an individual judging data quality if there are overriding trust issues. This topic is fertile for subsequent research

[5] Authenticity is evaluated higher up the stack.

[6] Again underscoring the mantra that "curation begins before creation"

a user judges those instruments to be; it can include metadata about the theoretical frameworks underlying the data, but not whether the frameworks are theoretically sound; it can clearly establish the parameters of the data, but it is up to the user to judge whether those are a complete or incomplete set of parameters. The context, provenance, and documentation provided by curation are thus critically important in arming users of data with the information they need to make quality judgments but are **not** capable of independently ensuring data authenticity or trust in data; that is entirely for the individual user to judge.

4. AREAS OF OPPORTUNITY

4.1 Curation Models

Given the issues with the practice of *post-hoc* curation raised above, it is worth examining alternative curation models. This is not to suggest that one model of curation is to be selected exclusively; a mix of *post-hoc* curation and curation-at-source models will likely be in place at most institutions.

The work required for doing curation at the source needs to be incentivized and integrated into the researcher's extant workflows. Unless there are clear and valuable incentives for researchers to spend time and thought on curatorial work, and unless curation can be made to fit into the way researchers currently work, curation will be an after-thought, and thus so will data quality.

These different curatorial models are not mutually exclusive and in fact it may be ideal to combine them, leveraging both the researcher's deep domain knowledge and the professional curator's commitment, expertise, and tools to preserve data quality over time.

4.1.1 Scaling Post-Hoc Curation

Curry has examined a number of successful community-based curation models, which may offer academic libraries a way to scale *post-hoc* curation and deal with the aforementioned deficiencies of this approach: "[d]ata curation teams have found it difficult to scale the traditional [*post-hoc* curation] approach and have tapped into community crowd-sourcing and automated and semi-automated curation algorithms" [8].

The rise of the "citizen science" paradigm, such as demonstrated in the Galaxy Zoo and Zooniverse projects [2, 4], suggests community crowd-sourcing as a tactic that may be used to complement an institution's curation model. These initiatives leverage the "wisdom of the crowd" in curating [7] massive data sets such as the astronomical image data in the original Galaxy Zoo project. Galaxy Zoo in particular has been wildly successful, attracting a user base numbering into the hundreds of thousands, who have worked together to classify hundreds of millions of records [4].

There are numerous incentives at play in crowdsourcing, such as access to broadly interesting and compellingly visualized data; competition; and a desire for the layperson to be involved with *bona fide* research with opportunities to make novel scientific discoveries despite limited domain expertise. Consider "Hanny's *Voorwerp* [3]," an astronomical body discovered in Galaxy Zoo's data set by an amateur astronomer. The *Voorwerp* is now being studied by more than one professional astronomer, studies that may never have happened if not for the serendipitous discovery of an untrained curator. There are numerous other collaborative or crowd-sourced curation efforts highlighted in Curry's chapter on community data curation [8].

Galaxy Zoo and other Zooniverse projects demonstrate aspects of a model that could be repurposed in academic libraries as libraries seek alternative models for research data curation that scale out.

As mentioned earlier, some combination of *post-hoc* curation and curation-at-source seems effective. The Galaxy Zoo project balances crowd-sourced curation with verification by trained astronomers [4], who verify samples of curatorial work over time, thus enabling network effects to take place — this form of training or correction is not unlike the balance between human correction and machine learning algorithms, or, *e.g.*, the reCAPTCHA [8] service. This sort of delegation of quality to the community is not unlike a principle found in the open source software world, which is that the more eyes are on a codebase, the more likely it is that defects will be found and corrected.

The challenges that face academic libraries in leveraging crowd-sourcing as part of an institutional data curation strategy, each of which bears more in-depth consideration or research, are finding or allocating sufficient resources to build tools; finding effective incentives to curate research data; building a community around the data that is large enough to realize the benefits of network effects; and coming up with a model that puts the "trust but verify" strategy, whereby a sampling of crowd-curated records is checked for quality (and corrected if need be), into effect at scale.

Curry [8] has identified a number of social and technical best practices around community curation, which may be useful in addressing these challenges: early and sustained stakeholder involvement; outreach beyond the existing community via multiple channels including both emerging social media and more traditional channels such as newsletters and mass email; connection of curation activities to tangible payoffs; an appropriate and clear governance model; community-standard data representations; balance between automated and human curation with the latter always overriding the former; and recording and displaying provenance events to provide additional context to crowd curators and users.

In addition to human curation, whether via trained curators or citizen curators in "the crowd," there is a growing number of increasingly sophisticated tools for automated curation which could be used as a less costly and more timely tier of curation (until such time as a human curator has time to curate a data set). Tools for automated curation such as for subject classification, part-of-speech tagging, semantic entity extraction, and characterization can provide

[7] Or, at least, classifying, cataloging, and otherwise annotating these data sets, even if it not inclusive of all activities within the curation lifecycle.

[8] `http://www.google.com/recaptcha`

much-needed context to enable some level of understandability, usability, authenticity, and trust. Automated curation can thus help with data quality in a way that scales in a less resource-constrained way than requiring intensive human curation of every data set.

4.2 Academic Libraries as Data Quality Hubs

Academic libraries have an opportunity to serve as data quality hubs on campus, extending their established digital curation and preservation services to the research enterprise, doing for e-science what libraries have a wealth of experience doing for other areas of scholarly communication. With the scramble to establish data management support services in the wake of the NSF's data management plan requirement, the timing is opportune to take advantage of the new and reinforced connections between libraries and researchers by offering new services around data quality.

Libraries that lack the resources to sustain a new university service around data quality, or libraries on campuses where other organizations (such as central IT) might be better resourced or positioned to provide such services, may play a less active but equally vital role. Libraries are in large part the centers of campus, where so much of the institution's research, publishing, and instruction come together. Librarians that serve as liaisons to academic departments and research institutes provide a crucial connection that libraries could use for outreach and marketing in the area of data quality services; though the libraries may not provide data quality services themselves, they may serve a consultative role, pointing at relevant services on campus and abroad, helping to "knit" them together for the research enterprise.

Libraries can also offer assistance in the form of instruction, not radically different from existing information literacy programs, particularly around practical tools and processes pertaining to personal digital curation [17]. Such instruction could be especially helpful at institutions where the culture is that of extreme decentralization or sparse collaboration.

There is a tremendous opportunity as well to offer workshops and otherwise emphasize the value of curation in providing data quality for e-science, and also to publicize the "curation begins before creation" mantra. The sooner libraries can insert themselves into the research process, the better the data quality situation will be on campus. Libraries need to figure out how to "hack" academic culture and scientific practice in such a way that curatorial skills are considered required within the new scientific process.

4.2.1 Helping Others to Help Us Help Others

New "data science" programs such as the certificate program at the University of Washington [13] give the author hope that there is some movement in this area. The focus on data gathering, analysis, and visualization is an important start; quality and curation, however, are noticeably absent. A more complete degree program in data science would effectively combine these topics with those within data curation and retention, pulling together domain-specific knowledge, scientific methodology, computer science techniques, and best practices from the information science, information technology, and cultural heritage realms to ensure effective management of data quality over time. The onus is on cultural heritage institutions such as academic libraries to make this happen, a daunting and enormous challenge to be realistic. It falls to us to make a convincing value-added argument regarding curation and preservation of data to researchers. Funding agencies like the NSF and NIH can help with this by continuing to require substantial data management plans, as can academic research offices and subject disciplines and institutes; forging or strengthening partnerships with these departments would be strategic for libraries on campus. This recommendation echoes one of the findings of the 2006 Association of Research Libraries report on data stewardship, namely that "[a] change in both the culture of federal funding agencies and of the research enterprise regarding digital data stewardship is necessary if the programs and initiatives that support the long-term preservation, curation, and stewardship of digital data are to be successful" [6].

4.2.2 Our Challenge

Are academic libraries adequately prepared for this role? A new suite of data quality services on campus may require not insignificant re-skilling and re-education of the workforce, and may also require some reorganization and redefinition of positions [12].

I agree strongly with Ogburn, who argues that "funding and planning for the care and retention of data must be built into the front end, not the back end, of the research process. Data files must be attended to while they are compiled and analyzed in order to keep them available for a reasonable life span. This will require librarians to be conversant with the language and methods of science, at the table for campus cyberinfrastructure planning, and working with researchers at the beginning stages of grant planning" [14].

Academic libraries **need** to be conversant with the language and methods of science and to be involved with advances in campus cyberinfrastructure. We have the expertise and the challenge of data quality is well within the traditional mission of libraries. The time has come for academic libraries to serve as data quality hubs on campus to enable a new generation of scientific discovery and inquiry for the good of our society.

5. REFERENCES

[1] `http://en.wikipedia.org/wiki/Semantic_Web_Stack`.

[2] `http://en.wikipedia.org/wiki/Galaxy_Zoo`.

[3] `http://en.wikipedia.org/wiki/Hanny{'}s_Voorwerp`.

[4] T. Adams. Galaxy zoo and the new dawn of citizen science. *The Guardian*, March 2012. `http://www.guardian.co.uk/science/2012/mar/18/galaxy-zoo-crowdsourcing-citizen-scientists`.

[5] S. Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8:159–172, 2005. `http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf`.

[6] Association of Research Libraries. *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering.* Association of Research Libraries, 2006.

http://www.arl.org/bm~doc/digdatarpt.pdf.

[7] Digital Curation Centre. DCC curation lifecycle model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[8] E. Curry, A. Freitas, and S. O'Riain. *The Role of Community-Driven Data Curation for Enterprises*, pages 25–47. Springer, 2010. http://3roundstones.com/led_book/led-curry-et-al.html.

[9] W. W. Eckerson. Data warehousing special report: Data quality and the bottom line. *Application Development Trends*, May 2002. http://adtmag.com/articles/2002/05/01/data-warehousing-special-report-data-quality-and-the-bottom-line_633729392210484545.aspx.

[10] C. Goble, R. Stevens, D. Hull, K. Wolstencroft, and R. Lopez. Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9:506–517, July 2008. http://bib.oxfordjournals.org/content/9/6/506.full.

[11] P. B. Heidorn. The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51:662–672, October 2011. http://dx.doi.org/10.1080/01930826.2011.601269.

[12] JISC. The data deluge. http://www.jisc.ac.uk/publications/briefingpapers/2004/pub_datadeluge.aspx, November 2004.

[13] U. of Washington Professional and C. Education. Winter 2013 | data science certificate. http://www.pce.uw.edu/certificates/data-science/web-winter-2013/, 2012.

[14] J. L. Ogburn. The imperative for data curation. *portal: Libraries and the Academy*, 10:241–246, 2010. http://muse.jhu.edu/journals/pla/summary/v010/10.2.ogburn.html.

[15] C. Rusbridge. Project data life course. http://digitalcuration.blogspot.com/2008/11/project-data-life-course.html, 2008.

[16] P. Russom. Liability and leverage - a case for data quality. *Information Management*, August 2006. http://www.information-management.com/issues/20060801/1060128-1.html.

[17] P. Williams, J. L. John, and I. Rowland. The personal curation of digital objects: A lifecycle approach. *Aslib Proceedings*, 61:340–363, 2009. http://dx.doi.org/10.1108/00012530910973767.

Towards Data Quality Metrics Based on Functional Requirements for Scientific Records

J. Caitlin Sticco
National Library of Medicine
38 Center Drive
Bethesda, MD 20894
(202) 480-4327
sticcojc@nlm.nih.gov

ABSTRACT

In this position statement, I propose defining the functional requirements of a scientific record in order to provide an evaluation framework for the usefulness of a dataset. Based on supporting these functions, I also propose a set of quality dimensions, metrics, and means of assessment that may be common to many disciplines. While many additional items may be proposed for specific disciplines, this set is simply offered as an illustration of a possible framework for quality control.

Categories and Subject Descriptors

E.0 [**Data**]: General

General Terms

Management, Measurement, Documentation, Design, Standardization

Keywords

Metadata, Data, Librarianship, Functional Requirements, Metrics, Quality Assessment

1. INTRODUCTION

Some of the questions proposed for these workshops suggest a wish to address usefulness, as well as quality as a purely objective trait. Let us draw distinctions between usefulness and quality. Quality measures are determinants of usefulness, but not all factors that make a dataset useful would necessarily be gathered under all definitions of quality. For example, data may be produced with utmost precision on the finest of instruments, but unannotated data removed from its original context is often useless. Additionally, data may be high quality, but essentially irrelevant for further use due to subject matter, licensing restrictions, or other factors. The uses to which the data may be put, that is, the functions of the record, should suggest the quality dimensions and criteria that are important to curators.

I propose defining the functional requirements of a scientific record in order to provide an evaluation framework for the usefulness of a scientific record. Such a framework would allow curators to identify those record functions that are most critical to the mission of their organization, and therefore support selection and policies based on supporting those critical functions. For example, a resource focused on allowing data reuse, like Protein Data Bank, is likely to be more intensely focused on making sure users can effectively compare data sets than an archive focused on preserving raw observations. They should consequently make different decisions on which version of data is useful and what types of maintenance processing are acceptable. I here propose a set of functional requirements for scientific records, with the intention that they be supplemented with details and more specific tasks by disciplinary authorities.

Based on supporting these functions, I also propose a set of quality dimensions, metrics, and means of assessment that may be common to many disciplines. While many additional items may be proposed for specific disciplines, this set is simply offered as an illustration of a possible framework for quality control. It was noted that this set of quality dimensions bears a strong resemblance to the data quality framework created by Wang and Strong. Although the two were created independently, I have since incorporated notes comparing the Wang and Strong framework to my own where appropriate. My contribution could be considered a supplement to [Wang and Strong 1996], with increased focus on computation and increased consideration given to data managers (instead of only data users). Additionally, this framework provides preliminary suggestions for actual metrics.

2. FUNCTIONAL REQUIREMENTS

With the intention of bridging a gap between broad functional roles such as those defined by the Functional Requirements for Bibliographic Records [Madison et al. 2007] and the detailed individual data elements in reporting schema like MIAME [Brazma et al. 2001], I here suggest a set of functional requirements for scientific records. Scientific functions given here are suggested from observation and personal experience in laboratories and libraries. Additional functions were suggested by comparing existing metadata schema in biomedicine, imaging, bibliography, and preservation. These tasks are designated administrative as opposed to scientific, in that they do not directly support scientific discovery. For example, discovery and analysis could continue even if authors were not credited, whereas discovery would be hampered irretrievably if a notebook were not kept well enough to correctly interpret the experimental results. Those functions designated as scientific are fundamental to scientific methodology.

Scientific functions include allowing users to:

- Correctly interpret experimental results (including diagnosis)
- Communicate and illustrate findings
- Replicate experiments
- Compare data sets appropriately

Administrative functions include allowing users to:

- Organize, sort, and aggregate information
- Search and retrieve information
- Control access
- Transfer intellectual property rights via licensing or other mechanisms and reuse data legally
- Credit authors, funders, or rights holders
- Prevent fraud or allegations of fraud, and other misconduct
- Preserve data

3. QUALITY DIMENSIONS AND METRICS

The specific quality criteria that allow record functions to be fulfilled will vary across disciplines, but some possible common dimensions, metrics, and assessment methods are suggested below as a framework.

A special consideration should be given to the differences between computer- and human-created data. As methods for automating data creation are increasingly available and reliable, we must consider that it is not enough to consider human performance a gold standard in all cases. Computer performance may simply represent a different set of errors. Computers are likely to be more consistent, for example, even while they lack the subtle semantics of a person. Multiple assessment methods may be necessary, and assessment methods may change over time.

3.1 Accuracy

Obviously, the accuracy of the dataset is paramount to its usefulness. Primary data must be checked thoroughly by review and replication, and processed data, where errors may be introduced at multiple points, must be rechecked for consistency. Wang and Strong include accuracy under the larger dimension of *Intrinsic Data Quality* with *believability, reputation*, and *objectivity*. However, their analysis is based on data consumer feedback. I believe these latter three characteristics are simply proxy mechanisms that data consumers use to evaluate the probability that data are accurate and useful. However, they are therefore useful to consider as metrics that could be representative of accuracy here.

3.1.1 Representative Metrics

- Reproducible results and consistency of results with earlier data or comparable data, especially verified reference sets
- For versioning, consistency with previous incarnations, or correction of known errors.
- Logical consistency, for data sets like ontologies.
- Plausibility. Specific rules would depend on disciple, but examples would include checks like gene sequences that are long enough to code a protein or physics particles that move slower than light speed.
- Known error rates for techniques or instrumentation.

3.1.2 Means of Assessment

- Peer review
- Reproduction
- Documentation for data collection and maintenance of process, policies, update schedule, and alteration transparency/consistency
- Automated checks for errors (e.g. checksums)
- Disciplinary authority or trust systems for sources
- Disciplinary rating criteria for quality.

3.2 Comprehensiveness

While it can be difficult in some cases to determine if a data set is complete or even to delimit one data set from another, completeness or comprehensiveness is still a useful quality criteria for many purposes. Even in cases where a data set might be said to be constantly expanding, as in the case of an annotated genome, a measure of relative comprehensiveness could be made. Data sets can be considered incomplete when missing data changes the interpretation of results, whether from data loss, withholding, or poor study design. Data sets can also be considered incomplete if they are missing sufficient metadata to support interpretation and reuse, such as descriptions of experimental procedures or study parameters. Wang and Strong use the broad category Contextual Data Quality to encompass what they term Completeness. They include several additional quality aspects that are very specific to the immediate needs of a user, and therefore do not lend themselves to general quality assurance metrics. In particular, Relevancy, Value-Added, and Appropriate Amount of Data can only be defined in relation to specific tasks, functions more specific than those proposed here. Timeliness is likely to also be defined in response to user needs; however, it at least can be supported by quantifiable metadata and policies supporting likely uses for the data.

3.2.1 Representative Metrics

- Data distribution anomalies such as interruptions in numbering or other sequences, statistical abnormalities, or lack of parallel reporting across study groups
- Relative completeness measures
- Discipline-specific reporting requirements
- General reporting requirements for administrative functions, such as bibliographic data and grant information

3.2.2 Means of Assessment

- Peer review
- Compliance with reporting guidelines and requirements

3.3 Accessibility

Otherwise high-quality data sets can have limited usefulness if they are difficult to access. License restrictions, for example, prevent bulk downloading of scientific journal articles for natural language processing. Data that is not actively curated or maintained can be difficult to find, decipher, and use.

Wang and Strong add to this concern the importance of access security, as well. It is also important that data sets with sensitive information, like patient data, be adequately protected. Data that is not correctly deidentified or restricted, for example, could be useless if it would be deemed unethical or illegal to use it.

3.3.1 Representative Metrics

- Restrictions on access and licensing
- Continuing availability of contacts and curators for maintenance and troubleshooting

3.3.2 Means of Assessment

- Compliance with licensing or sharing policies
- Documentation of access plans and contacts
- Rates of current use

3.4 Interoperability and Data Representation

Maximum interoperability between scientific records facilitates advanced computation, makes developing software tools easier and more efficient, improves the ability to make meaningful inferences by combining multiple data sources, and prevents the loss of information during conversion and preservation. Towards these ends, we may exert control on at least three aspects of data representation: vocabulary, syntax, and format. The complementary goals of partial compatibility and orthogonality translate to a design model that can be repeated over all of these areas, at multiple levels of granularity: modular, object-oriented design with standardized module links. The design principles that enable forward-compatibility (or future interoperability) are already well-outlined by the preservation community, but it is worth repeating them here: languages and file formats that are open, self-describing, and human readable provide the best possible base for successful conversion and interoperability later.

This category corresponds to the Wang and Strong dimension Representational Data Quality, covering both format and semantic interpretability. However, the Wang and Strong framework, created in 1996, lacks focus on facilitating computation, networking, and machine readability.

3.4.1 Representative Metrics

- Use of standards in vocabulary, syntax, and format
- Modular record or data model design with standards for linking independent modules
- Open format
- Self-describing format
- Human and machine readable formats
- Format currentness
- Documentation of state of processing
- Availability of tools that allow use of the data

3.4.2 Means of Assessment

- Automated validation against standards
- Peer review
- Rates of current use

3.5 Investment Value

While the quality measures above contribute to the value of datasets, additional measurements of dataset value are important. The types of metrics below are important to curators and funders for decision making about selection, return on investment, and future funding. I do not believe these considerations are adequately addressed by the Wang and Strong framework, as it is predicated exclusively on the needs of data users. Note that replacement cost estimates are likely to change over time as new scientific techniques replace older more expensive ones. Thus, these estimates would need to be reviewed and repeated for preservation purposes.

3.5.1 Representative Metrics

- Cost and time commitment of data acquisition, annotation, and maintenance
- Estimates of cost and time of dataset replacement (if possible).
- Projections of obsolescence or relevance to future research
- Uniqueness
- Size

3.5.2 Means of Assessment

- Funding tracking and budget reports
- Citation monitoring
- Modeling and projection, potentially based on similar fields

4. SUMMARY OPINION

Much discussion has been given to data quality as a contextual feature that depends on the use to which the data will be put. I suggest that it is necessary to establish the primary uses for a scientific record as a set of functional requirements before establishing the quality metrics to evaluate it. The data elements or other conditions needed to support each function of the record can then be determined and evaluated, based on those functions. Wang and Strong provided an excellent framework in 1996 for evaluating data quality that covers most of the broad quality dimensions it is necessary to examine, based on data user feedback. I believe this framework now requires supplementation to give more consideration to intensive computational activities with data sets and data management activities. Additionally, I have provided examples of practical metrics and means of assessment that might be applied to assess these qualities. This is not intended to offer a complete plan for assessment, but rather a starting place for contemplating related schemes.

5. AKNOWLEDGEMENTS

This research is supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health,

6. REFERENCES

[1] Wang, R and Strong, D. Beyond Accuracy: What Data Quality Means to Data Consumers. 1996. *Jrnl of Management and Information Systems*. 1996:12(4):5-34.

[2] Brazma, A, Hingamp, P, Quackenbush, J, et al. Minimum information about a microarray experiment (MIAME)-

toward standards for microarray data. 2001. *Nat. Genet.* 2001;29(4):365–371. DOI= http://doi:10.1038/ng1201-365

[3] Madison, O., Byrum, Jr, J., Jouguelet, S., McGarry, D., Williamson, N., Witt, M., Delsey, T., et al. 2007. *Functional Requirements for Bibliographic Records*. International Federation of Library Associations and Institutions.

Metrics for Data Quality

Douglas White
NIST
100 Bureau Drive, MS 8970
Gaithersburg, MD 20899
dwhite@nist.gov

Barbara Guttman
NIST
100 Bureau Drive, MS 8970
Gaithersburg, MD 20899
guttman@nist.gov

ABSTRACT

In this paper, we present an opinion on establishing metrics for data quality.

Keywords

data, quality, metrics, DiVisa, measurement

1. INTRODUCTION

Measurement of data quality can only be performed in relation to defined business rules or requirements. Quality is not absolute; it is subject to the environment in which it is discussed. Context dependence may have a role in determining a level of quality.

Data may be considered to have a high quality within the scope of the storage and retrieval mechanism in which it resides, but that measurement has limited applicability outside of the storage instance. The application of a context such as a set of government regulations can alter the known quality of data within a defined scope.

2. BACKGROUND

Data quality has been described via commonly used dimensions, e.g. accuracy, completeness, consistency, timeliness. Several dimension terms have been documented and aggregated with respect to internal or external data-related and system-related views. Many of the dimensions lend themselves well to measurement, while some have comparative assessment algorithms. The representation of multi-dimensional measurements as a unified “quality” is possible but the comparison is daunting.

Background material includes:

- *Towards a Vocabulary for Data Quality Management in Semantic Web Architectures,* Fuerber & Hepp
- *Anchoring Data Quality Dimensions In Ontological Foundations.* Wand & Wang
- *A Formal Definition Of Data Quality Problems,* Oliveira, Rodrigues, & Henriques
- *A Classification Of Data Quality Assessment Methods,* Borek, Woodall, Oberhofer & Parlikad

In the Perl (computer programming language) community, there is a concept of “Kwalitee,” which is used as an approximation of quality. Kwalitee increases as inconsistencies decrease. An inconsistency is a difference between documentation, test and implementation. Absence of inconsistencies does not imply kwalitee, which hearkens to the quote from Dijkstra: "Testing shows the presence of bugs, but never their absence." This might also run afoul of Godel’s incompleteness theorems.

3. OPINION

There are dimensions of data quality that can be measured or assessed. In the physical definition of measurement units, only time can be applied to data. All other assessments are Boolean values or comparative values as the result of a function applied to data (a test, statistics, relevance ranking, etc). If measurement and assessment is applied uniformly to data, it is possible to visualize multi-dimensional results.

NIST ITL Applied and Computational Mathematics Division, Scientific Applications and Visualization Group developed an information visualization tool called DiVisa.

DiVisa is a multi-dimensional visualization tool developed for researchers to understand the behavior of their data. From raw data, the user can interact with the visualization in order to obtain different "points of views" and thus to extract more information from the data. Geometrical forms such as squares, ellipses or lines are associated with data and visual attributes such as position, size, shape, color, stroke are used to represent different dimensions. Indeed, the researcher can easily modify the associations between data items and visual attributes, apply mathematical functions on and between items, subset and zoom in on areas, data ranges, or times of interest, superpose curves with transparency to compare them, and animate the visualization to show time series data. Moreover, the program can read any kind of data (simulation, statistics, text or numeric, etc.), and converters have been implemented to read several data formats without need for reformatting.

http://math.nist.gov/mcsd/savg/software/divisa/

Techniques for addressing data quality are important for digital curation, both to assess the quality of the holding as well as to assess the quality of tools used to store, manipulate and analyze the holdings. For example, secure hashes can be used to assess the integrity of stored digital objects, but it may be far more valuable to look at similarity digests. Data that has degraded may still have significant historical value.

The concept of multi-dimensional measurements for assessing the quality of digital holdings could address the use of comparisons of accuracy, reliability, timeliness, relevance, completeness, currency, consistency, precision, format, importance, usefulness, clarity, etc. For example, in the National Software Reference Library, we implement rules and perform tests within the scope of database field definitions. We also perform cross-referencing tests linking various pieces of our metadata. We are interested in exploring avenues for improving the methods we use to assess our holdings and in developing a more multi-dimensional approach.

Appendix 2. Biographies

Invited Participants

Dr. Micah Altman is Director of Research and Head/Scientist, Program on Information Science for the MIT Libraries, at the Massachusetts Institute of Technology. Dr. Altman is also a Non-Resident Senior Fellow at The Brookings Institution. Prior to arriving at MIT, Dr. Altman served at Harvard University for fifteen years as the Associate Director of the Harvard-MIT Data Center, Archival Director of the Henry A. Murray Archive, and Senior Research Scientist in the Institute for Quantitative Social Sciences.

Dr. Altman conducts research in social science, information science and research methods -- focusing on the intersections of information, technology, privacy, and politics; and on the dissemination, preservation, reliability and governance of scientific knowledge.

Kevin Ashley is Director of the UK's Digital Curation Center (http://www.dcc.ac.uk/). The DCC was established by JISC in 2004 to provide services, training, practical advice and guidance to research institutions on digital preservation with a special focus on research data curation. He serves on a number of advisory and guidance bodies and contributed the report on Data Quality and Curation which informed the GRDI2020 roadmap for research data infrastructure.

Kevin was formerly (1997-2010) Head of Digital Archives Department, University of London Computer Centre (ULCC) where he delivered digital preservation and repository services to organisations including the UK National Archives and the British Library. These included NDAD, which captured, preserved and provided access to Uk government datasets. He was a member of the RLG/NARA task force which developed TRAC and was chair of JISC's Repositories and Preservation Advisory Group. He was previously involved in the development and standardisation of network protocols, active in bodies such as ANSI, BSI and EWOS. He began his career in a medical research unit devoted to innovative uses of IT in the support of clinical research and practice.

Jackie Bronicki is the project coordinating librarian for a grant, funded by the Institute of Museum and Library Services, focusing on validating quality in large-scale digitization. Jackie is based out of Technical Services at the University of Michigan Library and works under the direct guidance of the Professor Paul Conway, Principal Investigator for the grant, from the School of Information at University of Michigan. She manages many aspects of the day to day operations of the research project with a focus on data collection to determine frequency and severity of error in digitization. She graduated from Rice University in 1997 with a Bachelor of Arts in Human Physiology, with an area interest in Biology and Biochemistry. In 2005, she completed her MLIS at Wayne State University with a specialization in medical librarianship. Before joining the project team at University of Michigan, she was a project coordinator for a large international dialysis study focused on collecting both qualitative and quantitative data from over 400 facilities in 13 countries.

Ruth Duerr is currently the manager of NSIDC's data stewardship program, and PI/Project Manager for several ongoing data management and cyberinfrastructure projects.

Her research interests involve nearly all aspects of data stewardship. Her recent activities include a NASA-sponsored activity that demonstrated the feasibility of using the PREMIS metadata standard with NSIDC data holdings at the data set level; testing mechanisms for improving the long-term recoverability of data in NASA's archive that are in outdated data formats; and a NOAA-sponsored project, working with the science community to introduce production of detailed metadata into the product development process.

Duerr also leads NSIDC efforts on two cyberinfrastructure-related efforts: Libre and the Data Conservancy. Duerr is also working with NASA's Technology Infusion working group to identify data stewardship-related technology and standards gaps, to assess the readiness levels of existing standards and technologies, and to recommend data stewardship-related technologies for adoption.

Ricky Erway is Senior Program Officer in OCLC Research. She coordinates the Research Information Management program, investigating how academic libraries can better serve their institutions' research missions. She also works on topics related to digitization (rights issues, public/private partnerships, increasing the scale of digitization of special collections, and managing born-digital materials). Ricky's wide-ranging expertise is often tapped for invited presentations at a variety of professional conferences and workshops. Prior to the combination of RLG and OCLC in July of 2006, Ricky was Manager of Digital Resources at RLG, responsible for CAMIO (the Catalog of Art Museum Images Online) and RLG Cultural Materials (digitized special collections from libraries, archives, and museums) and was a key player in the development of ArchiveGrid (a service that aggregates bibliographic records and finding aids describing archival and special collections). Before joining RLG, Ricky worked at the Library of Congress for nine years, the last five as associate coordinator of the American Memory program, aimed at significantly increasing public access to the special collections of the Library of Congress. Ricky has an MLS from the University of Wisconsin, USA.

Andrew Fiore Ph.D., is a data scientist at Facebook and a lecturer at the UC Berkeley School of Information. Fiore, a researcher in computer-mediated communication, has examined relationship formation through online dating, designed and prototyped novel interfaces for online social interaction, and analyzed social judgments in large-scale conversations. Previously, he was a visiting assistant professor at Michigan State University. He holds a Ph.D. from the UC Berkeley School of Information, master's degrees in Statistics from Berkeley and Media Arts and Sciences from MIT, and an undergraduate degree from Cornell.

Michael Giarlo is Digital Library Architect at the Pennsylvania State University. His primary roles are designing a technical architecture for durable access to the institution's digital assets, providing vision and strategy for the development of the architecture, building a development team, and fostering community around digital curation locally and abroad. He has been working in library technology since 1999, holding systems administration and software development positions primarily in support of digital libraries and repositories at the Library of Congress, Princeton University, the University of Washington, and Rutgers University. He earned both a bachelor's degree in linguistics and an MLIS at Rutgers. His top interests are APIs, IPAs, and AIPs.

Alan Hall - Mr. Hall is a Senior IT Team Lead for NOAA's National Climatic Data Center (NCDC) where his primary responsibilities are the stewardship of the Nation's Climate resource. NCDC is the world's largest archive of climate data. In addition, Mr. Hall is NOAA's representative of the White House Office of Science and Technology Policy Subcommittee on Networking and Information Technology Research and Development (NITRD) Program's Big Data Senior Steering Group. He is also a member of the National Science Foundations' Data Net Consortium focusing on a national framework for sharing science data across disparate disciplines.

Leslie Johnston has over twenty years experience in digitization and digital conversion, setting and applying metadata and content standards, and overseeing the development of digital content management and delivery systems and services. She is Chief of Repository Development at the Library of Congress, which includes managing technical architecture initiatives in the National Digital Information Infrastructure and Preservation Program. Previously, she served as the head of digital access services at the University of Virginia Library; Head of Instructional Technology and Library Information Systems at the Harvard Design School; the academic technology specialist for Art for the Stanford University Libraries; and as database specialist for the Getty Research Institute. She has also been active in the museum community, working for various museums, teaching courses on museum systems, editing the journal Spectra and serving on the board of the Museum Computer Network.

After receiving the PhD degree in Chemical Physics in 1969, **Michael Lesk** joined the computer science research group at Bell Laboratories, where he worked until 1984. From 1984 to 1995 he managed the computer science research group at Bellcore, then joined the National Science Foundation as head of the Division of Information and Intelligent Systems, and since 2003 has been Professor of Library and Information Science at Rutgers University, and chair of that department 2005-2008.

He is best known for work in electronic libraries, and his book "Practical Digital Libraries" was published in 1997 by Morgan Kaufmann and the revision "Understanding Digital Libraries" appeared in 2004. His research has included the CORE project for chemical information, and he wrote some Unix system utilities including those for table printing (tbl), lexical analyzers (lex), and intersystem mail (uucp). His other technical interests include document production and retrieval software, computer networks, computer languages, and human-computer interfaces. He is a Fellow of the Association for Computing Machinery, received the Flame award from the Usenix association, and in 2005 was elected to the National Academy of Engineering. He was the first chair of the NRC Board on Research Data and Information.

Matthew S. Mayernik is a Research Data Services Specialist in the library of the National Center for Atmospheric Research (NCAR)/University Corporation for Atmospheric Research (UCAR). He has a MLIS and Ph.D. from the UCLA Department of Information Studies. His work within the NCAR/UCAR library is focused on developing research data services. His research interests include research data management, data publication and citation, metadata practices and standards, cyberinfrastructure development, and social aspects of research data.

Jerome McDonough Associate Professor, Graduate School of Library & Information Science University of Illinois at Urbana-Champaign. Dr. McDonough has been on the faculty of the Graduate School of Library & Information Science since 2005. His research focuses on socio-technical aspects of digital libraries, with a particular focus on issues of metadata and description as well as digital preservation of complex media and software. Prior to joining the faculty at GSLIS, Dr. McDonough served as the head of the Digital Library Development Team for New York University. He has also been an active participant in metadata standards activities for digital libraries, having served as chair of the METS Editorial Board, as well as serving on the NISO Standards Development Committee and on the ODRL International Advisory Board.
Dr. McDonough completed his doctoral studies at the U.C. Berkeley School of Library & Information Studies in 2000. His dissertation, "Under Construction: The Application of a Feminist Sociology to Information Systems Design," investigated the construction of identity in graphical, computed-mediated communication systems and the influence that CMC system designers may yield on their users' presentation of self.

Prasenjit Mitra is an Associate Professor in the College of Information Sciences and Technology; he serves on the graduate faculty of the Department of Computer Sciences and Engineering and is an affiliate faculty member of the Department of Industrial and Manufacturing Engineering at The Pennsylvania State University. His major research interests are in exploring issues in information extraction, information integration and information visualization. His research is being supported by the NSF CAREER Award. Additionally, his research has been supported by the NSF, Microsoft Corporation, DoD, DHS, DoE, NGA, and DTRA. He obtained his Ph.D. in Electrical Engineering from Stanford University in 2004. Prior to that, he obtained his M.S. in Computer Science from The University of Texas at Austin in 1994 and his B. Tech. (Hons.) from the Indian Institute of Technology,Kharagpur in 1993. From 1995 to 2000, he was a Senior Member of the Technical Staff at Oracle Corporation in the Oracle Parallel Server and Languages and Relational Technologies groups in the Server Technologies division. He also serves in the Board of Advisors of Global IDs, Inc. Mitrahas co-authored over sixty articles at top conferences and journals. His work along (with his co-authors) resulted in a visual analytics system that was awarded the IEEE VAST '08 Grand Challenge award in the Data Integration area. He has served as the co-chair of three workshops including WIDM’09 and served on the PC of several conferences including SIGMOD, AAAI, IJCAI, WWW, CIKM, and ICDM.

Reagan Moore is the Director of the Data Intensive Cyber Environments Center at the University of North Carolina at Chapel Hill, professor in the School of Information and Library Science, and Chief Scientist at the Renaissance Computing Institute. Moore coordinates research efforts in development of policy-based data management systems that are used to support data grids, digital libraries, processing pipelines and persistent archives. Moore is the principal investigator for the development of the integrated Rule Oriented Data System. Moore has a B.S. in physics from the California Institute of Technology (1967), and a Ph.D. in plasma physics from the University of California, San Diego (1978).

Michael L. Nelson is an associate professor of computer science at OldDominion University. Prior to joining ODU, he worked at NASA Langley Research Center from 1991-2002. He is a co-editor of the OAI-PMH and OAI-ORE specifications and is a 2007 recipient of an NSF CAREER award. He has developed many digital libraries, including the NASA Technical Report Server. His research interests include repository-object interaction and alternative approaches to digital preservation. More information about Dr. Nelson can be found at: http://www.cs.odu.edu/~mln/

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems (ifs) at the Vienna University of Technology (TU-Wien), and a Key Researcher at Secure Business Austria (SBA), repsonsible for the Digital Preservation Team. He furthermore is president of AARIT, the Austrian Association for Research in IT and a Honorary Research Fellow in the Department of Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. He received his MSc and PhD in Computer Science from the Vienna University of Technology in 1997 and 2000, respectively. In 2001 he joined the National Research Council of Italy (CNR) in Pisa as an ERCIM Research Fellow, followed by an ERCIM Research position at the French National Institute for Research in Computer Science and Control (INRIA), at Rocquencourt, France, in 2002. From 2004-2008 he was also head of the iSpaces research group at the eCommerce Competence Center (ec3). He is a member of the Association for Computing Machinery (ACM), The Institute of Electrical and Electronics Engineers (IEEE), the Austrian Society for Artificial Intelligence (ÖGAI), and serves on the board of the IEEE Technical Committee on Digital Libraries (TCDL), as well as on the Board of the Austrian Computer Society (OCG).

Caitlin Sticco is a Systems Librarian at the National Library of Medicine, and a 2010 NLM Associate Fellow. She received her MLS in 2009, and a Specialist Certificate in Library and Information Studies in 2010, from the University of Wisconsin at Madison. She received her BS in Biomedical Science from Antioch College. She has previously worked as a health librarian, database administrator, molecular and cell biology technician, and an informatics assistant in LOCI (Laboratory for Optical Computation and Instrumentation). Her research interests include biocuration, data standards, and Natural Language Processing, focusing on topics such as automated indexing and summarization, curation policy, and evaluation. She is currently developing semi-automated systems for indexing MEDLINE and indexing quality evaluation methods.

Jamie Taylor – Google

Kristin M. Tolle, MS, Ph.D. is a Director in the Microsoft Research connections team and a Clinical Associate Professor at the University of Washington (College of Medicine). Since joining Microsoft, Dr. Tolle has acquired numerous patents and worked for several product teams including the Natural Language Group, Visual Studio, and the Microsoft Office Excel Team. Prior to joining Microsoft, Tolle was an Oak Ridge Science and Engineering Research Fellow for the National Library of Medicine and a Research Associate at the University of Arizona Artificial Intelligence Lab managing the group on medical information retrieval and natural language processing. She earned her Ph.D. in Management of Information Systems with a minor in Computational Linguistics. Her research interests include ubiquitous computing, global public health, contextual computing, natural language processing and machine translation, mobile computing, user intent modeling and information extraction from large heterogeneous data sources.

Douglas White - Doug has worked at NIST since 1987. His experience has covered distributed systems, distributed databases and telecommunication protocols. He has written programs in many areas, including real time biomonitoring, real time video processing, web site/database integration, system administration scripts and network monitoring scripts. He holds both a B.A and M.S. in computer science from Hood College, and is a member of the IEEE Computer Society and the Association for Computing Machinery. Doug has been involved with the National Software Reference Library (NSRL) since 2001, and is currently the project leader for the NSRL.

Jasmine Young is a Biocurator Team Leader working at RCSB PDB for more than 9 years.

The PDB is a single archival center of macromolecular structural data that is freely and publicly available to the global community.

She is responsible for managing the complex global services of the RCSB PDB Biocuration group, including: integration of curation and software development teams, setting standard procedures and guidelines in concert with wwPDB collaborators, maintaining format documentation, and creating new approaches to solve scientific data problems and to improve data quality.

Workshop Organizers

Gary Marchionini is the Dean and Cary C. Boshamer Professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. He teaches courses in human-information interaction, interface design and testing, and digital libraries. He has published over 200 articles, chapters and reports in a variety of books and journals. Professor Marchionini has had grants or research awards from the National Science Foundation, Council on Library Resources, the National Library of Medicine, the Library of Congress, Bureau of Labor Statistics, Kellogg Foundation, NASA, The National Cancer Institute, Microsoft, Google, and IBM among others. Professor Marchionini was Editor-in-Chief for the *ACM Transaction on Information Systems* (2002-2008) and is the editor for the Morgan-Claypool Lecture Series *on Information Concepts, Retrieval, and Services*. He has been program chair for ACM SIGIR (2005) and ACM/IEEE JCDL (2002) as well as general chair of ACM DL 96 and JCDL 2006. His current interests and projects are related to: interfaces that support information seeking and information retrieval and usability of personal health records. He currently is PI on a grant from NSF focused on a search results framework that supports searches over multiple sessions and in collaboration.

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches courses on archival administration; records management; digital curation; understanding information technology for managing digital collections; and acquiring information from digital storage media. He is a lead organizer and instructor for the DigCCurr Professional Institute, a week-long continuing education workshop on digital curation, and he teaches professional workshops on the application of digital forensics methods and principles to digital acquisitions.

Cal's primary area of research is the long-term curation of digital collections. He is particularly interested in the professionalization of this work and the diffusion of existing tools and methods into professional practice. Cal developed "A Framework for Contextual Information in Digital Collections" (Journal of Documentation), and edited and provided several chapters to I, Digital: Personal Collections in the Digital Era published by the Society of American Archivists. Cal is Principal Investigator of the BitCurator project, which is developing and disseminating open-source digital forensics tools for use by archivists and librarians. He was also Principal Investigator of the Digital Acquisition Learning Laboratory (DALL) project, which investigated and tested the incorporation of digital forensics tools and methods into digital curation education. Cal has served as Co-PI on several projects focused on preparing professionals for digital curation responsibilities. In a project called Curation of a Forensic Data Collection for Education, Cal investigated and developed resources to enhance access and use of disk images to support digital forensics education.

Heather Bowden is a Carolina Digital Curation Doctoral Fellow at the University of North Carolina at Chapel Hill. During her time at UNC, she has served as the project manager of the Closing the Digital Curation Gap and Digital Curation Curriculum (DigCCurr) II projects. As part of her work for these projects, she created the Digital Curation Exchange (DCE) website. Her doctoral research is centered on creating an actionable file format endangerment metric and establishing a baseline of data on file format endangerment levels from which continued risk monitoring may be conducted.

Appendix 3. Workshop Schedule

Sunday, September 9, 2012

Afternoon-evening	Travel and arrival in Washington DC

Monday, September 10, 2012

8:00-8:30am	BREAKFAST in meeting room
8:30-9:15 am	Introductions and overview
9:15-10:15am	Summaries of papers
10:15-10:30am	BREAK
10:30-11:30am	Breakout Session 1* – Pain points
11:30am-12:00pm	Breakout Session reports back in plenary
12:00-1:00pm	LUNCH in meeting room
1:00-2:00pm	Breakout Session 2* – Promising directions
2:00-2:45pm	Breakout Session reports back in plenary
2:45-3:00pm	BREAK
3:00-3:30pm	Plenary discussion
3:30-4:30pm	Brainstorm session - Project ideas
4:30-5:00pm	Grouping and ordering of projects
5:00-5:15pm	Wrap-up
6:00-8:30pm	Group dinner

** Assigned based on position paper subject areas*

Tuesday, September 11, 2012

8:00-8:30am	BREAKFAST in meeting room
8:30-9:00am	Review
9:00-10:30am	Project breakouts
10:30-11:00am	BREAK
11:00am-12:00pm	Reporting on project proposals
12:00-1:00pm	LUNCH in meeting room
1:30-2:00pm	Plenary discussion - Highlights and important takeaways
2:00-2:15pm	BREAK
2:15-3:00pm	Plenary discussion - Final report and task assignments
3:00-3:30pm	Wrap-up