American Chemical Society

Customer Advisory Panel

Data Curation Subcommittee Report

September 2011

Table of Contents

ACS Customer Advisory Panel – Data Curation Subcommittee Report

Overview

Data curation is a set of activities that facilitate the long-term reuse of data, as well as its preservation to enable this reuse. This set of activities encompasses nearly every aspect of the life cycle of scientific data including standards for formatting and citing data as well as policies and procedures for access, sharing, archiving and preservation. In the context of this report "data" encompasses digital scientific datasets including text, numbers, images, video, audio, software, algorithms, equations, animations, models, simulations, etc.

The international scholarly community has begun to identify the issues surrounding data curation and to develop approaches for handling the vast amount of scientific data that currently exists and is exponentially growing. As it exists today, data curation is still an emergent set of policies, roles and activities. The data curation arena is evolving.

Executive Summary and Recommendations

The American Chemical Society (ACS) Customer Advisory Panel (CAP) Data Curation Subcommittee has been tasked with answering three questions concerning the role of ACS in the data curation arena. The subcommittee has responded by performing an environmental scan covering the issues in question.

Question 1: What should the ACS Publications (or ACS) position, if any, be on the emerging data policies?

The concept of data management and data management planning has been around for many years, but became much more high profile when, in January of 2011, the National Science Foundation (NSF) began to require a data management plan (DMP) as part of all grant applications. Prior to this, major US federal funding agencies had been requiring DMPs only for grant funding of $500K and greater.

However, in the last several years academic institutions have already been, both collaboratively and independently, working to address data curation issues. Collaborations such as DataCite[1], Digital Curation Centre[2], DMP Partnership[3], Distributed Data Curation Center at Purdue[4], University of California Curation Center[5] and DataONE[6] have been developing policies and

standards to support access and reuse, providing resources to develop knowledge and skills and creating the infrastructure needed to support long term preservation of data.

Publishers have also been actively participating in the issues surrounding data curation. Publishers such as American Association for the Advancement of Science, American Geophysical Union and the National Academy of Sciences have longstanding policies requiring the availability or permanent archiving of data, unique and persistent identifiers for datasets or have official position statements on data curation. Others, such as Royal Society of Chemistry, recognize there are opportunities for learned societies in the arena of data curation, but also realize the effort involved is potentially substantial.

Q1 Recommendations:

- *ACS should fully support the principle of open sharing and preservation of digital research data and its metadata.*
- *ACS should create a formal position statement on data sharing, re-use and preservation.*

Question 2: Should ACS Publications take an active role in establishing standards for data formats and data preservation, in the chemistry domain?

Relevant standards for data formats and data preservation would include:
1) Standards related to the reporting of data
2) The structure and nature of data reporting standards (e.g., the nature of derivative data sets when primary data sets would be too large to maintain)
3) Data format standards
4) Data citation standards
5) Metadata standards
6) Additional data related standards not included above

Q2 Recommendations:

It would be recommended that ACS engage in standards development. This recommendation is based upon the vast expertise of ACS in this domain, the impact of developing chemical data standards on the corporate and academic community, and ACS' role in serving the communities and disciplines impacted by chemistry.

The nature of engagement is a critical element of this recommendation. At a time that we see a proliferation of standards in this space, it is vital that ACS engage broadly in the data community and help to develop a common set of international standards related to chemical data. ACS could play a pivotal role in helping to harmonize data standards, and in the process, help save the community time and money.

The American Chemical Society has long provided a vital societal role in progressing science through scholarly communications. As scientific communications evolves to encompass data communications and exchange, it is critical that ACS takes up this challenge and remains engaged at the bulwarks of the profession.

Question 3: Is there a role for ACS or ACS Publications in helping authors, or more globally, chemists, in meeting the emerging requirements?

An environmental scan of the roles various publishers play in the process of data curation reveals a wide spectrum of services provided. They range from passive activities such as maintaining, or pointing to, a list of current locations of datasets or providing recommendation services for appropriate data centers, to much more robust services like providing links between papers and associated data along with persistent identifiers or supporting standard formats for data, metadata and citation of data.

Publisher roles generally support the theme of providing context for the reader. There is value in connecting data and articles. 85% of researchers believe that it is useful to link underlying research data with the formal literature[7]. Linking could be at the article and/or entity level.[8] There is also value for publishers in collaborating with data repositories to ensure long term availability of data and to coordinate processes. Overall, it will be valuable for publishers to begin to un-silo the literature and the underlying data and to introduce an integrated user experience.

Q3 Recommendations:

- *ACS should maintain active awareness of and support standards for formats of data, metadata and citation of data as well as standards for peer review of data.*
- *There are opportunities for services ACS can provide to its authors, members and readership to introduce an integrated user experience. These services can include lists or directories of existing datasets associated with ACS papers along with provision of links and persistent identifiers for such data.*
- *Recognizing that data management is an immense undertaking, ACS should investigate opportunities for collaborative activity with existing or new subject-appropriate data centers, other professional societies or existing collaborations.*

Summary of Findings

Data sharing is integral to translation of research results into knowledge. The report of the National Science Board (NSB) on digital data collections says "Long-lived digital data collections are powerful catalysts for progress and for democratization of science and education."[9] This meshes well with the ACS vision of "improving people's lives through the transforming power of chemistry." The importance of data collections in support of research and education is growing.

ACS is timely with its questions about the role of a publisher in the digital data curation arena. Education leads to informed decisions. In an evolving landscape, it is especially true that it is important to continually monitor events that are integral to business concerns. Overall, it is incumbent upon ACS to further analyze its internal policy issues relevant to digital data curation and to begin to actively participate in this arena.

Recommendations – Near Term

- *ACS should fully support the principle of open sharing and preservation of digital research data and its metadata.*
- *ACS should create a formal position statement on data sharing, re-use and preservation.*
- *ACS should maintain active awareness of and support for standards regarding formats of data, metadata and citation of data as well as standards for peer review of data.*

Recommendations – Mid Term

- *Provide recommendation/referral services, for its authors and members, to appropriate existing data centers.*
- *Provide links between ACS papers and associated data (both directions) along with support for persistent identifiers for datasets.*

Recommendations – Long Term

- *ACS should investigate opportunities for collaborative activity with existing or new subject-appropriate data centers, other professional societies or existing collaborations.*

Overview of Potential Outcomes

Data curation issues of quality assurance, protection and security, heterogeneity of data, interoperability and intellectual property rights must be managed. However, these are issues for virtually any data management process and are not new. Open sharing and preservation of digital research data and its associated metadata is a powerful force for inclusion. It removes barriers, increases reach and broadens impact of research outcomes. ACS can support these positive outcomes through adoption of some or all of the various recommendations made in this report.

Conclusion

There is no doubt there are costs associated with the above mentioned recommendations. Some costs are low (creating a list of existing data centers) and some are higher (collaborative projects). However, the costs of doing nothing are greater. The NSB report on digital data collections says "The need for digital collections is increasing rapidly, driven by the exponential increase in the volume of digital information. There is a need to rationalize action and investment …". All participants in the research life cycle have a stake in this arena.

The scope of issues and activities surrounding the data curation arena is significant in size and complexity. Participation in this arena is of substantial importance to the progress and democratization of science. ACS is strongly urged to participate actively, collaboratively and early in the interest of not only its own goals, but those of ACS authors, members and the "benefit of Earth and its people"[10].

Bibliography/Footnotes/Links

1. DataCite - http://datacite.org/
2. Distributed Data Curation Center (D2C2) - http://d2c2.lib.purdue.edu/
3. DMP Partnership (US) –
4. Digital Curation Centre (UK) - http://www.dcc.ac.uk/
5. University of California Curation Center (UC3) - http://www.cdlib.org/services/uc3/
6. DataONE (Data Observation Network for Earth) - https://www.dataone.org/
7. Parse.Insight survey - http://www.parse-insight.eu/
8. Cambridge Crystallographic Data Centre - http://www.ccdc.cam.ac.uk/
9. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century, 2005, http://www.nsf.gov/pubs/2005/nsb0540/start.jsp
10. American Chemical Society mission statement - http://portal.acs.org/portal/PublicWebSite/about/aboutacs/index.htm

Author Contacts

Carol Hoover (Team Lead)

Los Alamos National Laboratory, Digital Information Resources Manager, Research Library

Email: hoover@lanl.gov

Bob Schwarzwalder

Stanford University, Director, Library Technologies Division

Email: rns@stanford.edu