2013

Infrastructure, Standards, and Policies for Research Data Management

Jian Qin
Syracuse University, jqin@syr.edu

Infrastructure, Standards, and Policies for Research Data Management

Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY, USA
Email: jqin@syr.edu

Abstract—This paper discusses the needs and importance of research data management and introduces the concept of research data management as an infrastructure service. Although many resources have been made available for research data management, most of them are developed as "islands" and lack linking mechanisms. The lack of integrated and interconnected resources has contributed to high cost and duplicated efforts in data management operations. The vision of research data management as an infrastructure service is not only to improve the efficiency of research data management but also the productivity of the research enterprise. Each of the three dimensions—infrastructure, standards, and policies—addresses a critical aspect of research data managementto make the data infrastructure services work.

Keywords—Data infrastructure; Research data management; Data services; Metadata standards; Data policies

I. INTRODUCTION

Research data management has gained increasing recognition for its value and importance among funding agencies and research institutions, as evidenced by the fast growth of data repositories at disciplinary community and institutional levels. Examples of these repositories include the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/), Dryad (http://datadryad.org/), and GenBank (http://www.ncbi.nlm.nih.gov/genbank/), among others. While these disciplinary repositories are important venues for data curation and sharing, they targeted on the end product of a research lifecycle. The large amounts of work necessary for data to reach the submission point are left to researchers to deal with.

Two years ago the *Science* magazine conducted a survey to their peer reviewers from the previous year on the availability and use of data. The 1,700 responses represented input from an international and interdisciplinary group of scientific leaders. As the *Science* editorial reported, "About 20% of the respondents regularly use or analyze data sets exceeding 100 gigabytes, and 7% use data sets exceeding 1 terabyte. About half of those polled store their data only in their laboratories—not an ideal long-term solution. Many bemoaned the lack of common metadata and archives as a main impediment to using and storing data, and most of the respondents have no funding to support archiving" [1].

The *Science* magazine survey presents two major problems in the current state of scientific data management in a research lifecycle: there is a lack of funding and staff support for managing active data and a lack of metadata standards and tools for managing active data in research lifecycle. What does it take to solve these problems? In other words, what needs to be done to provide the support necessary for improving research productivity through effective data management? The answers lie in a good understanding of research and data lifecycle and their implications to data management and support needed for managing scientific data.

This paper will first discuss what a research and data lifecycle is and its relations and requirements to data management, and then go on to describe the three pillars in data management: institutionalization, standards, and infrastructure. As these three concepts may be interpreted differently in other contexts, each of them will be articulated with examples. The goal of this position paper is to raise the awareness of the data management issues and advocate for research data infrastructure services.

II. RESEARCH LIFECYCLE AND DATA LIFECYCLE

Lifecycle is a term frequently used in our technology-driven society. Examples include information systems lifecycle, information transfer lifecycle, and many other variations depending on for which domain the term lifecycle is used. In the science data management domain, this term is used in several contexts: research lifecycle, data lifecycle, data curation lifecycle, and data management lifecycle. Each version has a different emphasis but they are often related or overlap in one way or the other. A research lifecycle generally includes study concept and design, data collection, data processing, data access and dissemination, and analysis [3]. As a research project progresses along the stages, different data will be collected, processed, calibrated, transformed, segmented or merged. Data at these stages go through one state to the next after certain processing or condition is performed on them. Some of these data are in the active state and may be changed frequently while others such as raw data and analysis-ready datasets will be tagged with metadata for discovery and reuse.At each stage of this lifecycle, the context and type of research (Fig. 1) can directly affect the types of data generated and requirements for how the data will be processed, stored, managed, and preserved.

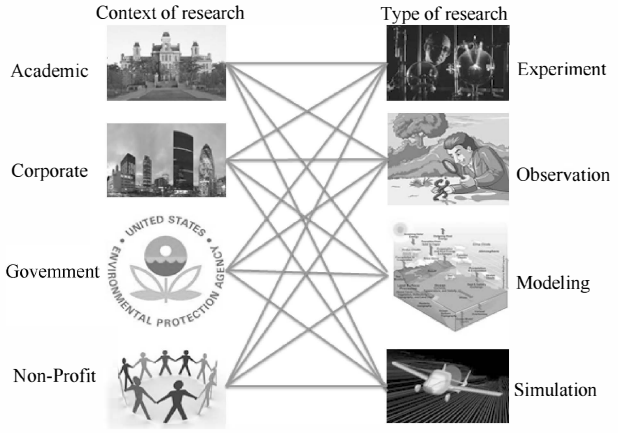

Fig.1 The types and contexts of research

For example, in the United States, national research centers such as NCAR (National Center for Atmospheric Research) and NOAA (National Oceanic and Atmospheric Administration) regularly collect data about the global ecosystems and process them into data products for scientific research and learning. The research lifecycle and data lifecycle at this level will be different from those at the individual project level where teams of scientists have specific goals to solve specific problems. The scale of data and requirements for data management will vary along the stages of the whole lifecycle. National research centers are publically funded agencies and have the obligation of preserving and providing access to ecosystems data they collected. Hence generating data products and providing ways to discover and obtain data is crucial for them. Another example is the type of research projects carried out at academic institutions. These research projects can be collaborative among institutions or within a department/college within the same institution. The data collected and generated from these projects are specialized and subject to the control and regulation of different data policies and compliances, which creates a different set of issues and requirements for data management and use from those generated by the national research centers.

Regardless of the context and nature of research, scientific data need to be stored, organized, documented, preserved (or discarded), and made discoverable and usable. The amount of work and time involved in these processes is daunting and intellectually intensive as well as costly. The personnel performing these tasks must be highly trained in technology and subject fields and able to effectively communicate between different stakeholders. In this sense, the lifecycle of research and data is not only a technical domain but also a domain requiring management and communication skills. To be able to manage scientific data at community, institution, and project levels without reinventing-the-wheel, a data infrastructure is necessary to provide the efficiency and services for scientific research as well as data management.

III. RESEARCH DATA MANAGEMENT AS AN INFRASTRUCTURE SERVICE

The data-centric research lifecycle no doubt relies heavily on effective research data management. But what is research data management? In a nutshell, research data management is essentially a series of services that an organization develops and implements through institutionalized data policies, technological infrastructures, and information standards. The concept of data infrastructure adopts the principle of “Infrastructure as a Service (IaaS),” which is “a standardized, highly

automated offering, where compute resources, complemented by storage and networking capabilities are owned and hosted by a service provider and offered to customers on-demand" [4].In the context of a data infrastructure, stakeholders will be able to carry out data management functions through a Web-based user interface.

Infrastructure is a notion of modern society. Being modern is to live within and by means of infrastructures: basic systems and services that are reliable, standardized, and widely accessible, at least within a community.Susan Leigh Star and Karen Ruehdler [5] neatly summarized the features of infrastructures:

• Embeddedness. Infrastructure is sunk into, inside of, other structures, social arrangements, and technologies.

• Transparency. Infrastructure does not have to be reinvented each time of assembled for each task, but invisibly supports those tasks.

• Reach or scope beyond a single event or a local practice.

• Learned as part of membership.

• Links with conventions of practice.

• Embodiment of standards.

• Built on an installed base.

• Becomes visible upon breakdown.

• Is fixed in modular increments, not all at once or globally [5].

These characteristics can also well describe the one that supports science data management. For example, a service that ingests a large number of small data files to build a searchable and filterable database can be scaled up for any disciplines that have the same data management need.

Although so far there is no single agreed-upon definition for the concept of data infrastructure, scientific research powerhouses such as UK and US have consistently invested in building it. In a recent program solicitation, the U.S. National Science Foundation (NSF) delineates that a data cyberinfrastructure has the functions of storing digital data, applying new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams. Also required of the data infrastructure are data services to support acquisition, documentation, security and integrity, storage, access, analysis and dissemination, migration, and de-accession of data archives and repositories [6]. A report by the DG Information Society and Media in United Kingdom uses the term "e-infrastructure" to refer to the technologies of various kinds for creating, collecting, annotating, manipulating, storing, finding and re-using information and services such as those to provide user support, training, and preservation. Included in this e-infrastructure are also information resources and associated tools such as vocabularies, ontologies, rights management and privacy protection systems, and curation [7]. In summary, a data infrastructure is an orchestration of technologies, data and metadata standards, and policies embedded in the research enterprise. Such an infrastructure may exist within an institution, a research community, or at national and international scales.

IV. THREE DIMENSIONS OF DATA INFRASTRUCTURE SERVICES

The concept of a data infrastructure implies three dimensions: technologies, data and metadata standards, and policies that govern the management, sharing, and use of data.

A. The Technology Dimension

The technology infrastructure covers a wide range of technologies for collecting, storing, processing, organizing, transmitting, and preserving data as well as platforms for communication and collaboration. Included in this dimension of the data infrastructure are networks, databases, authentication systems, and software applications. Scientific data and databases are different from conventional ones used for business transactions or employee records due to the idiosyncrasies of scientific data. Not only are scientific data collected from various sources such as observations, experiments, crowd-contributions (e.g., data generated from citizen science projects), or computer modeling /simulations, but also come with a wide variety of types and formats as well as varying levels of processing. Raw data collected from observations, experiments, modeling, or simulations often need to go through a series processing, transformation, and quality check before the data can be used for analysis. Differences in data types and formats

cross disciplines or even within the same discipline field can become barriers for data sharing and reuse[8]. The technological dimension of data infrastructure, therefore, is not just a simple technical issue but rather, is closely tied with the policies and standards.

B. The Dimension of Data and Metadata Standards

Another important dimension of a data infrastructure is data and metadata standards. Scientific data can be grouped into three large blocks based on discipline and type:

• Physical and chemical data: include element data, chemical data, isotope data, and particle data;

• Earth and astronomical data: range from weather and climate data, geodesy data to astronomical data for static and dynamic properties of stars, planets, and other objects; and

• Life sciences data: this group contains a long list of varieties, including genome data, flora and fauna data, protein data, nucleotide sequences, biomedical and clinical data, and the list can go on.

What complicates the diverse types of scientific data is the large number of data format standards that were developed since the introduction of computer into research. Fig. 2 shows data formats from the very basic physical level to metaformats tospecialized scientific data formats. As data formats move from basic level to more specialized formats, the diversity and complexity increases drastically. The Common Data Format (CDF), for example, is a data format standard developed in 1985 by the National Space Science Data Center (NSSDC) and contains self-describing metadata for the storage and manipulation of scalar and multidimensional data in a platform- and discipline-independent fashion [9]. Another example is the biomedical data that appear in a large number of formats and each of them serves a specific type of data. . Protein Data Bank (PDB) format, for instance, is designed for recording macromolecular data for the PDB archive, including atomic coordinates, crystallographic structure factors and NMR (Nuclear Magnetic Resonance) experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

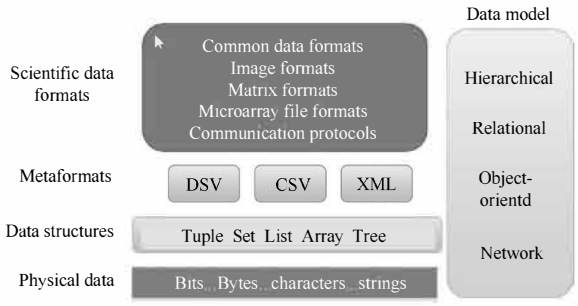

Fig. 2 Data formats at different technical layers

The complexity of scientific data types and formats requires specialized tools to process and analyze, which results in a large number of standards for both data and metadata. Data format standards specify how data are stored in computer and read by application software. In a case scenario of large data archives, portable and self-describing data formats are critical for data archiving to allow data sets to be read not only by current software in use but also by future technologies. The transmission of data also requires data in formats that can be delivered across hardware and software. Standardized data formats allow data to be 1) convertible—get in and out of storage easily, 2) portable—readable anywhere, and 3) extensible—can add types and structures later. Data formats are also critical for data to work with all software so that researchers can minimize the time spent converting between formats.

Metadata standards for scientific data define the elements and their structures used to describe data sets. Each disciplinary field has its own metadata standard to describe data sets while sharing some common elements with other standards. The metadata standards shown in Fig. 3, for example, all need to use geospatial elements to describe the geographical region or location related to species, natural phenomena such as precipitation, temperature, and wind speed, and landscape features.

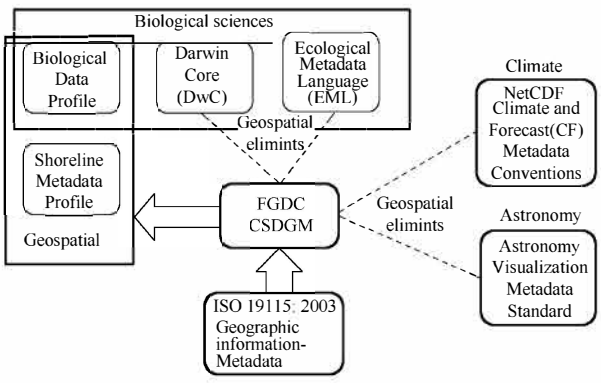

Fig. 3 Major metadata standards and their relations through the geospatial description

Metadata standards for scientific data are designed to document details about who collected the data at where, what the data content is about, and how the data were collected. All these are critical for effective data discovery and use. The complexity of scientific data mentioned earlier in this paperhas led to complex metadata standards. It is not uncommon for metadata standards in the scientific data domain to have hundreds of elements with deep layers of structures. While complex, large metadata standards do provide a comprehensive description for data sets and satisfy the requirements for data discovery and use, their sizes can become barriers for metadata description because large standards make automatic metadata creation almost impossible and at the same time, manual metadata creation is time consuming and expensive and can never keep up with the pace of scientific data growth. At present, each metadata standard has its own tool(s) and most of them are standalone, that is, names of entities and controlled vocabularies are not automatically linked and relationships between data and publications need to be manually added.A data infrastructure will be able to tackle these problems by making metadata schema, entity instances, and controlled vocabularies into infrastructural services.

C. The Policy Dimension

Policies for scientific data cover a wide range of topics. From national and global perspectives, data policies are mostly related to data sharing, intellectual property protection, ethical issues, and open access [8]. At this level, the role of data policies is to guide the practices of data management, sharing, and use.The National Institute for Health (NIH) has implemented guidelines on data sharingas early as in 2003, which require projects exceeding $500,000 "in direct cost in any year" to include plans for data sharing [10]. NSF also made it mandatory in 2011 that research grant proposals submitted to NSF must include a supplementary document with a label "Data Management Plan" (DMP). This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results [11].

The DMP mandate by NSF sprung a flurry of training tutorials, workshops, and studies on how DMP should be prepared to address the key questions on data archiving and sharing, data citation, copyright and privacy/confidentiality of data, data documentation and management, file formats and data types, data organization, security, and storage and backups of data. Research libraries in the U.S. developed DMP template tools and consulting services to help researchers prepare their DMP document in proposal writing process. The Data Management Consulting Group at the University of Virginia Library (http://dmconsult.library.virginia.edu/) and the Research Data Management Service Group at Cornell University (https://confluence.cornell.edu/display/rdmsgweb/Home) are two well know exemplar data management services offered by research libraries.

The NSF mandate for data management plan also brought up many issues that many institutions have not well thought out before. For example, DMP requires research proposals to specify the types and formats of data to be produced and how they will be stored, shared, and managed. To address these requirements, researchers must make their DMP compliant with their institutional data policies in addition to the federal mandate. Researchers need to know what institutional policies are regarding which data types and formats should be archived, whether the institution has a data repository for storing their data files, and what procedures they should establish when sharing data with colleagues and community. In a content analysis of institutional data policies, Bohémier et al. identified six aspects of data policies that should be addressed: data curation, management, use, access, publishing, and sharing. They discovered that data policies are implanted unevenly across institutions: only 15% of all policies applied to the institutions as a whole while most applied only to specific disciplines, collections, or projects [12].

Data policies at national and institutional levels establish the framework for individual researchers and projects to make their policies in day-to-day operation. Different policies address different areas of questions. For data archiving purpose, an understanding of the nature of data can directly affect the policy. For example, data generated from observing nature phenomena such as volcano eruptions, hurricanes, earthquakes, and precipitations cannot be reproduced or replicated, hence will be preserved indefinitely, while clinical trials on a drug's effects on certain medical conditions can be and should be able to reproduced and replicated and hence may need to be archived for regulation and compliance

purposes. When data are being actively worked on, they can change frequently. Creating comprehensive metadata descriptions for active data files may not be practical. A data policy at an individual project level can help researchers comply with funding agency and institutional requirements and at the same time establish best practices in managing data and preparing them for submitting to institutional and disciplinary data repositories for arching and sharing.

In many ways, the process of developing data policies is also a process of institutionalization. "To institutionalizing something means to establish a standard practice or custom within a human system" [13]. Data management in many institutions and disciplinary fields is still an area to be studied. The survey findings mentioned at the beginning of this paper demonstrate the importance of institutionalization of data management, which includes establishing data policies, administrative support that will ensure the funding and personnel for data management operations, and best practice guidelines.

CONCLUSION

This paper discussed the needs and importance of research data management and introduced the concept of research data management as an infrastructure service. Although many resources are available for research data management, many of them are developed as "islands" and lack linking mechanisms. This has contributed to high cost and duplicated efforts in data management operations. The vision of research data management as an infrastructure service is not only to improve the efficiency of research data management but also the productivity of the research enterprise. Each of the three dimensions—infrastructure, standards, and policies—addresses a critical aspect to make the data infrastructure services work. In-depth studies will be needed to understand what user and architectural requirements there are for a data infrastructure in scientific domains, what resources have been made available and how can they be connected to support the research lifecycle and data management lifecycle, as well as what tools are needed and how various resources can be incorporated into the tools to make data management more effective and productive.

REFERENCES

[1] Science Staff, "Challenges and opportunities," Introdutionto special sectionDealing with Data. Science, 11 February 2011: Vol. 331, pp. 692-693.

[2] P. N. Edwards, A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, Mass.: The MIT Press, 2010, pp. 3-8.

[3] C. Humphrey, "e-Science and the life cycle of research," unpublished, 2006. http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc.

[4] Gartner, "IT glossary", http://www.gartner.com/it-glossary/infrastructure-as-a-service-iaas/.

[5] S.L. Star & K. Ruhleder,"Steps toward an ecology of infrastructure: Design and access for large information space." Information Systems Research, Vol. 7, pp. 111-134, 1996.

[6] NSF, "Data infrastructure building blocks (DIBBs)," Program Solicitation NSF-12-557. Arlinton, VA: NSF, 2012. http://www.nsf.gov/pubs/2012/nsf12557/nsf12557.htm

[7] The Digital Archiving Consultancy Limited, "Towards a European e-infrastructure for e-science digital repositories," 2006 S88-092641, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

[8] W. L. Anderson, "Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data."Data Science Journal, Vol. 3, pp. 191–202. http://www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf

[9] Space Physics Data Facility (SPDF), "CDF frequently asked questions," January 15, 2013, http://cdf.gsfc.nasa.gov/html/Version/V3500/FAQ.html.

[10] National Institutes of Health, "NIH Data sharing policy and implementation guidance," Office of extramural research., March 5, 2003,http://grants.nih.gov/grants/policy/data_sharing/data_sharing _guidance.htm

[11] National Science Foundation, "NSF Data Management PlanRequirements." December 8, 2010,. http://www.nsf.gov/eng/general/dmp.jsp

[12] K.T. Bohémier, A. Atwood, A. Kuehn, & J. Qin, "A content analysis of institutional data policies," Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL'11), June 13-17, 2011, Ottawa, Canada, pp. 409-410.

[13] M. Kramer, "Make it last forever: The institutionalization of service learning in America," 2000, pp. 14., http://www.national-serviceresources.org/filemanager/download/NatlServFellows/kramer.pdf