Levels of Services and Curation for High Functioning Data

G. Sayeed Choudhury[1], Carole L. Palmer[2], Karen S. Baker[2], Timothy DiLauro[1]

The Sheridan Library ~ Johns Hopkins University Libraries

[1] Sheridan Libraries, Johns Hopkins University

[2] Center for Informatics Research in Science & Scholarship
Graduate School of Library & Information Science, University of Illinois, Urbana-Champaign

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois
CIRSS
Center for Informatics Research in Science & Scholarship

Introduction

The growing volume and variety of data brings new demands and opportunities. This conceptual model represents levels of data repository services and the cumulative nature of curation.

The Data Management Stack model integrates contributions from two groups within the Data Conservancy Initiative (http://dataconservancy.org):

- The Technical team and Data Management Services team at Johns Hopkins University, focused on designing and implementing systems (Choudhury & Hanisch, 2009; Mayernik et al, 2012)
- The Data Practices team at the University of Illinois, focused on social studies of data curation (Palmer et al., 2011; Weber et al, 2012).

The Model

The model represents four levels of activity and capacity shown in the center panel. It builds on definitions offered by Lord and Macdonald (2004). Today, the use of these terms, together with the notion of data stewardship (NAP, 2009), is fluid and inconsistent. Caution is advised in applying these concepts (BRTF, 2010).

Progress with Shared Vocabulary

The Stack Model has proven useful for communicating with researchers who often use terms such as **storage**, **archiving**, **preservation** and **curation** interchangeably.

The model contributes to building a shared vocabulary by making evident

- connections and dependencies among levels of services
- ramifications of repository choices made by researchers

Data Management Layers

Layers	Characteristics	Implication for PI	Implication relative to NSF
Curation	• Adding value throughout life-cycle	• Feature Extraction • New query capabilities • Cross-disciplinary	• Competitive advantage • New opportunities
Preservation	• Ensuring that data can be fully used and interpreted	• Ability to use own data in the future (e.g. 5 yrs) • Data sharing	• Satisfies NSF needs across directorates
Archiving	• Data protection including fixity, identifiers	• Provides identifiers for sharing, references, etc.	• Could satisfy most NSF requirements

Storage	• Bits on disk, tape, cloud, etc.
	• Backup and restore
		• Responsible for:
		  • Restore
		  • Sharing
		  • Staffing
			• Could be enough for now but not near-term future

The Stack

Increasing layers of support and functionality; each level depends on the level below. (Choudhury, 2009).

- **Storage**: lowest service; basic physical storage with backup and restore services.
- **Archive**: following BRTF, "activities that enable long-term retention of digital materials"; DC focus on data protection through replication, fixity, and identifiers.
- **Preservation**: providing enough representation information, context, metadata, fixity, etc. to support use and interpretation by agents other than the original data producer.
- **Curation**: processes that add value to foster discovery and reuse.

The curation level identifies a range of services, enabling use for purposes not necessarily envisioned by the data producers.

References

BRTF (2010). Blue Ribbon Task Force Report on Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Choudhury, S. and R. Hanisch (2009). The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation.

Lord, P., A. MacDonald, et al. (2004). From data deluge to data curation. Proceedings of the UK e-Science All Hands Meeting, Nottingham.

Mayernik, M.S., G.S. Choudhury, T. DiLauro, E. Metsger, B. Pralie, M. Rippin, R. Duerr, (2012). The Data Conservancy Instance: Infrastructure and Organizational Services for Research Data Curation. D-Lib 18(9/10).

Palmer, C.L., N.M. Weber, and M.H. Cragin (2011). The Analytic Potential of Scientific Data: Understanding Re-use Value Proceedings of the American Society of Information Science and Technology. ASIST 2011.

Weber, N., K.S. Baker, A. Thomer, T. Chao, and C. Palmer (2012). Value and Context in Data Use: Domain Analysis Revisted. Proceedings of the American Society of Information Science and Technology. ASIST 2012, Baltimore, Maryland.

National Science Foundation
DataConservancy

Acknowledgements

Thanks to other contributing team members Barbara Pralle, David Fearon, Betsy Gunia, Ruth Duerr, Tiffany Chao, Nicholas Weber, and Cheryl Thompson. This research was supported by the National Science Foundation DataNet award OCI-0830976 and IMLS award #RE-02-10-0004-10.

Institute of Museum and Library Services
Data Curation Education In Research Centers
DCERC