Journal of Information Management
ISSN: 2348-1765 (Print), ISSN: 2348-1773 (Online)
Volume 3, Issue 1 (January-June, 2016), pp. 71-79

http:// www.splpjim.org

DATA CURATION: THE PROCESSING OF DATA

Krishna Gopal
Librarian , Kendriya Vidyalaya NTPC Dadri, GB Nagar
eeshukrishna@gmail.com

ABSTRACT

Responsibility for data curation rests, in different ways, on a number of different professional roles. Increasingly, within the library, data curation responsibilities are being associated with specific jobs (with titles like "data curator" or "data curation specialist"), and the rise of specialized training programs within library schools has reinforced this process by providing a stream of qualified staff to fill these roles. At the same time, other kinds of library staff have responsibilities that may dovetail with (or even take the place of) these specific roles: for instance, metadata librarians are strongly engaged in curatorial processes, as are repository managers and subject librarians who work closely with data creators in specific fields. Different techniques and different organizations are engaged in data curation. Finally, it is increasingly being recognized that the data creators themselves (faculty researchers or library staff) have a very important responsibility at the outset: to follow relevant standards, to document their work and methods, and to work closely with data curation specialists so that their data is curatable over the long term.

Keywords:*Data Storage, Data Management, Data curation, Data annotation, Data preservation*

1. INTRODUCTION

Data Curation is a term used to indicate management activities related to organization and integration of data collected from various sources, annotation of the data, publication and presentation of the data such that the value of the data is maintained over time and the data remains available for reuse and preservation.

1.1 Definition by Whatls.com

Data curation is the management of data throughout its lifecycle, from creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted. The main purpose of data curation is to ensure that data is reliably retrievable for future research purposes or reuse. Within the enterprise, compliance is another primary purpose.

1.2 What process is the data curation?:- Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data".

1.3 Definition and practice by Council on Library and Information Resources

According to the 'University of Illinois' Graduate School of Library and Information Science, "Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.

Deep background on data libraries appeared in a 1982 issue of the Illinois journal, *Library Trends.* For historical background on the data archive movement, see "Social Scientific Information Needs for Numeric Data: The Evolution of the International Data Archive Infrastructure

This term is sometimes used in context of biological databases, where specific biological information is firstly obtained from a range of research articles and then stored within a specific category of database. For instance, information about anti-depressant drugs can be obtained from various sources and, after checking whether they are available as a database or not, they are saved under a drug's database's anti-depressive category. Enterprises are also utilizing data curation within their operational and strategic processes to ensure data quality and accuracy.

1.4 Project and studies

The Dissemination Information Packages (DIPS) for Information Reuse (DIPIR) project is studying research data produced and used by quantitative social scientists, archaeologists, and zoologists. The intended audience is researchers who use secondary data and the digital curators, digital repository managers, data center staff, and others who collect, manage, and store digital information

1.5 Data Curation Lifecycle

This Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualization or receipt through the iterative curation cycle.

The model enables granular functionality to be mapped against it: to define roles and responsibilities and build a framework of standards and technologies to implement.

It can be used to help identify additional steps that may be required – or actions not required by certain situations or disciplines – and to ensure that processes and policies are adequately documented

2. FULL LIFECYCLE ACTIONS

2.1 Description and Representation Information

Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.

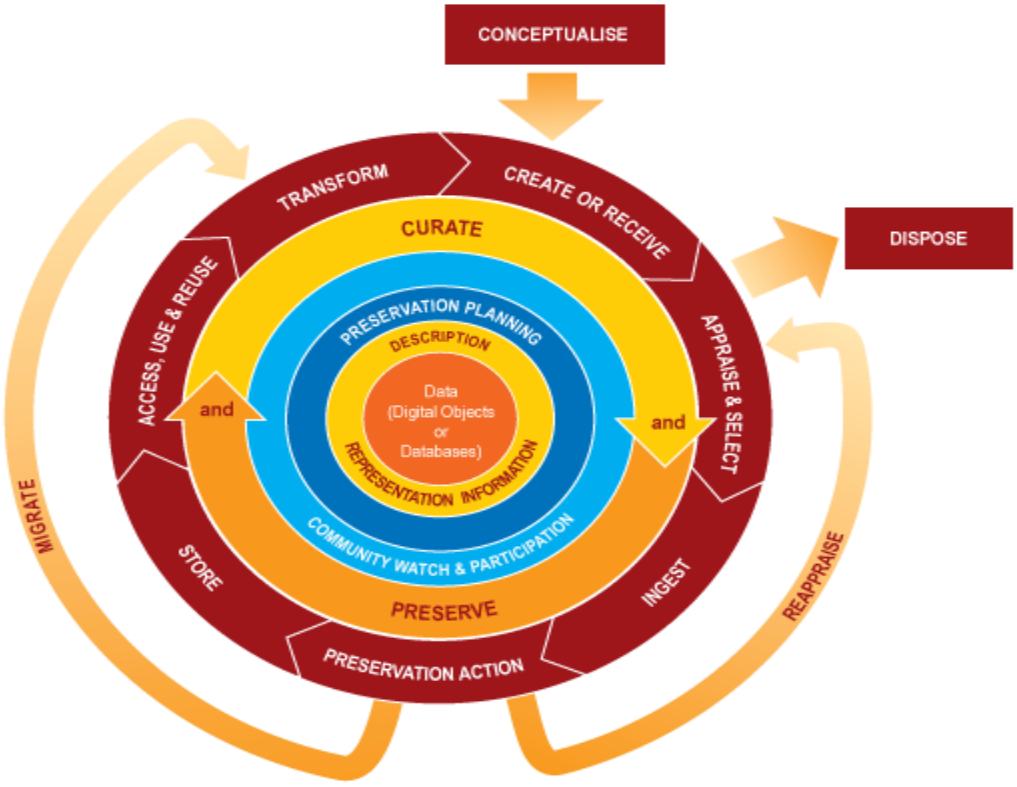

Fig. 1: Data Curation Lifecycle Model

2.2 Preservation Planning

Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

2.3 Community Watch and Participation

Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

2.4 Curate and Preserve

Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

3. ACTIVITIES INCLUDED IN THE PROCESS OF DATA CURATION.

3.1 Data Archaeology

It refers to the art and science of recovering computer data encoded and/or encrypted in now obsolete media or formats. Data archaeology can also refer to recovering information from damaged electronic formats after natural or man- made disasters.

The term originally appeared in 1993 as part of the Global Oceanographic Data Archaeology and Rescue Project (GODAR). The original impetus for data archaeology came from the need to recover computerized records of climatic conditions stored on old computer tape, which can provide valuable evidence for testing theories of climate change. These approaches allowed the reconstruction of an image of the Arctic that had been captured by the Nimbus 2 satellite on September 23, 1966, in higher resolution than ever seen before from this type of data.

3.2 Data degradation

It is also known as **data decay** or **data rot**, is a colloquial computing phrase for the gradual decay of storage media It should not be confused with “bit rot” defined in the *Jargon Files* as a jocular explanation for the degradation of a software program over time even if "nothing has changed"

3.3 Data format management (DFM)

It is the application of a systematic approach to the selection and use of the data formats used to encode information for storage on a computer. In practical terms, data format management is the analysis of data formats and their associated technical, legal or economic attributes which can either enhance or detract from the ability of a digital asset or a given information systems to meet

specified objectives. Data format management as an analytic tool or approach is data format neutral.

Historically individuals, organization and businesses have been categorized by their type of computer or their operating system.

3.4 Data governance

It is a control that ensures that the data entry by an operations team member or by an automated process meets precise standards, such as a business rule, a data definition and data integrity constraints in the data model. The data governor uses data quality monitoring against production data to communicate errors in data back to operational team members, or to the technical support team, for corrective action. Data governance is used by organizations to exercise control over processes and methods used by their data stewards and data custodians in order to improve data quality.

Data governance is a set of processes that ensures that important data assets are formally managed throughout the enterprise. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. It is about putting people in charge of fixing and preventing issues with data so that the enterprise can become more efficient.

3.5 Data stewardship

A **data steward** is a person responsible for the management and fitness of data elements (also known as critical data elements) - both the content and metadata. Data stewards have a specialist role that incorporates processes, policies, guidelines and responsibilities for administering organizations' entire data in compliance with policy and/or regulatory obligations. A data steward may share some responsibilities with a data custodian.

3.6 Data governance tools

Leaders of successful data governance programs declared in December 2006 at the Data Governance Conference in Orlando, Fl that data governance is between 80 and 95 percent communication. That stated, it is a given that many of the objectives of a Data Governance program must be accomplished with appropriate tools. Many vendors are now positioning their products as Data Governance tools; due to the different focus areas of various data governance initiatives, any given tool may or may not be appropriate, in addition, many tools that are not marketed as governance tools address governance needs.

4. DATA GOVERNANCE ORGANIZATIONS

4.1 DAMA International

DAMA (the Data Management Association) is a not-for-profit, vendor-independent, international association of technical and business professionals dedicated to advancing the concepts and practices of information resource management (IRM) and data resource management (DRM).

4.2 Data Governance Professionals Organization (DGPO)

The Data Governance Professionals Organization (DGPO) is a non-profit, vendor neutral, association of business, IT and data professionals dedicated to advancing the discipline of data governance. The objective of the DGPO is to provide a forum that fosters discussion and networking for members and to encourage, develop and advance the skills of members working in the data governance discipline.

4.3 The Data Governance Society

The Data Governance Society, Inc. is dedicated to fostering a new paradigm for the effective use and protection of information in which Data is governed and leveraged as a unique corporate asset.

4.4 The Data Governance Council

The Data Governance Council is an organization formed by IBM consisting of companies, institutions and technology solution providers with the stated objective to build consistency and quality control in governance, which will help companies better protect critical data."

4.5 IQ International -- the International Association for Information and Data Quality

IQ International is a not-for-profit, vendor neutral, professional association formed in 2004, dedicated to building the information and data quality profession.

4.6 Data management

The definition provided in the DAMA Data Management Body of Knowledge is: "Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

4.7 Other tools For Data Curation

1. Data governance
    - Data asset
    - Data governance
    - Data steward
2. Data Architecture, Analysis and Design

- Data analysis
- Data architecture
- Data modeling

3. Database Management
   - Data maintenance
   - Database administration
   - Database management system
4. Data Security Management
   - Data access
   - Data erasure
   - Data privacy
   - Data security
5. Data Quality Management
   - Data cleansing
   - Data integrity
   - Data enrichment
   - Data quality
   - Data quality assurance
6. Reference and Master Data Management
   - Data integration
   - Master data management
   - Reference data
7. Data Warehousing and Business Intelligence Management
   - Business intelligence
   - Data mart
   - Data mining
   - Data movement (Extract, transform, load)
   - Data warehouse
8. Document, Record and Content Management
   - Document management system
   - Records management
9. Meta Data Management
   - Meta-data management
   - Metadata
   - Metadata discovery
   - Metadata publishing
   - Metadata registry
10. Contact Data Management
    - Business continuity planning
    - Marketing operations
    - Customer data integration
    - Identity management
    - Identity theft
    - Data theft
    - ERP software
    - CRM software
    - Address (geography)
    - Postal code
    - Email address
    - Telephone number

REFERENCES

- Renée J. Miller, "Big Data Curation" in 20th International Conference on Management of Data (COMAD) 2014, Hyderabad, India, December 17-19, 2014

- Bio creative Glossary at http://biocreative.sourceforge.net/biocreative_glossary.html
- "An Introduction to Humanities Data Curation" by Julia Flanders and Trevor Muñoz http://guide.dhcuration.org/intro
- Pilin Glossary at http://www.pilin.net.au/Project_Documents/Glossary.html.
- https://en.wikipedia.org/wiki/Data_curation extracted on 24/6/2016.
- Dissemination Information Packages for Information Reuse (DIPIR) project http://www.oclc.org/research/themes/user-studies/dipir.html
- Kathleen M. Heim, "Social Scientific Information Needs for Numeric Data: The Evolution of the International Data Archive Infrastructure." in Collection Management 9 (Spring 1987): 1-53.
- Techno-archaeology rescues climate data from early satellites U.S. National Snow and Ice Data Center (NSIDC), January 2010 Archived.
- https://en.wikipedia.org/wiki/Data_archaeology extracted on 24/06/2016
- Raymond, Eric. "Bit rot". The Jargon File. Retrieved3 March 2013.
- https://en.wikipedia.org/wiki/Data_format_management extracted on 24/06/2016.
- Sarsfield, Steve (2009). "The Data Governance Imperative", IT Governance.
- https://en.wikipedia.org/wiki/Data_governance extracted on 24/06/2016.
- "DataGovernanceSoftware.com". The Data Governance Institute. Archived from the original on 2008-10-02. Retrieved 2008-10-02.
- DAMA International
- Data Governance Professionals Organization
- Data Governance Society
- Data Governance Council
- IQ International, the International Association for Information and Data Quality
- Data Governance and Information Quality Conference.
- DAMA-DMBOK Guide (Data Management Body of Knowledge) Introduction & Project Status" (Note: PDF no longer available online at https://www.dama.org.
- https://en.wikipedia.org/wiki/Data_management extracted on 24/06/2016.
- http://www.clir.org/initiatives-partnerships/data-curation extracted on 25/06/2016.
- http://whatis.techtarget.com/definition/data-curation extracted on 25/06/2016.

- Universal Meta Data Models, by David Marco and Michael Jennings, Wiley, 2004, page 93-94 ISBN 0-471-08177-9
- Metadata Solution by Adrinne Tannenbaum, Addison Wesley, 2002, page 412
- Building and Managing the Meta Data Repository, by David Marco, Wiley, 2000, pages 61–62
- *The Data Warehouse Lifecycle Toolkit*, by Ralph Kimball et. el., Wiley, 1998, also briefly mentions the role of data steward in the context of data warehouse project management on page 70.
- *Developing Geospatial Intelligence Stewardship for Multinational Operations*, by Jeff Thomas, US Army Command General Staff College, 2010, www.dtic.mil/dtic/tr/fulltext/u2/a524227.pdf.
- https://en.wikipedia.org/wiki/Data_steward extracted on 25/06/2016.
- http://www.dcc.ac.uk/resources/curation-lifecycle-model. Extracted on 27/06/2016.