Data Curation: An opportunity for the libraries

Conference Paper · December 2018

CITATIONS
0

READS
1,507

1 author:

Amit Tiwari
Indian Statistical Institute
15 PUBLICATIONS **22** CITATIONS

SEE PROFILE

Data Curation: An opportunity for the libraries

Amit Tiwari
Junior Research Fellow, Documentation Research and Training Centre,
Indian Statistical Institute, Bangalore
amittiwari@drtc.isibang.ac.in

Abstract

Data is a new factor of production in this era. A huge volume of data production is leading to the situation of data deluge. The challenge is how to manage such data? Data curation is the data management activity which aims to enable the accessibility and reusability of the data. Data is managed in the layers of curation, preservation, archiving and storage. Data curation deals with the extraction, transformation, and loading of the data. The expertise in traditional library activities has given the library professionals an opportunity to lead from the front as a data curator.

Keywords : Data Curation, Data Management, Metadata, DCC Life Cycle Model

Data Curation: An opportunity for the libraries

Introduction

Data are the new fuel in the 21st century. Petabytes of data are generated every day due to the various human and machine activities. It is expected, the total volume of useful data will reach up to 16 zettabytes i.e., 16 trillion GB (Turner et al. 2014). Some of the major phenomena which generate an immense amount of data are social media, search engines, medical devices, sensors etc. The data generated due to this phenomenon can be extremely useful to make the future inferences.

Data curation deals with the acquisition, describing and providing access to the users. It involves the various data management activities e.g., organization, publishing, and preservation of the data. It is a data management activity which facilitates the further researchers for reuse of the research data and also warns them about the redundancy.

The major areas where library professionals are working for the centuries are the acquisition, information organization, and management, information dissemination and preservation. Though library professionals are in a better position to do major work in the curation activity, they require the support of the domain experts, statisticians and technology experts. A single domain expert will be inefficient in the curation of the data, so the collaboration is a better way to perform the curation activity.

Definition and Meaning of Data Curation

In a generic sense, data curation is an active data management activity which harnesses the data for further use in the analysis and to derive the inferences.
Graduate school of library and information science, University of Illinois defines data curation as:

"The active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation."

According to *Cragin et al. (2007),*

"Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness; … curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time"

Here, the discovery, reuse, and retrieval are the major goals, which is being achieved using various data curation activities.

Literature review

In a generic sense the data curation in the continuous process to manage the data by adding or removing different phenomena in the data. Its very purpose is to enhance the reusability of the data.

Data curation is a part of data management lifecycle. Different scholars and agencies have explained the data management lifecycle using different workflow models. Lenhardt and others explain their model in 8 steps namely, “plan, collect, describe, preserve, discover, integrate and analyze”(Lenhardt, Ahalt, Blanton, Christopherson & Idaszak, 2013). United States geological survey is more specific in its workflow model as they include publishing, metadata, quality management, and security also in their model. These lifecycle workflow models are the step by step guides of data curation activities.

Data curation activity involves data collection, selection, erroneous data identification and correction, heterogeneity treatment, categorization and classification, metadata creation, and preservation.

As per the data management lifecycle, the data management steps are planning, data creation, relevance, classification, data description, analysis, publish, archival or preservation, and reuse. In above-mentioned steps data creation, relevance classification, data description, and preservation are the steps which a data curator uses for the curation.

Need for Data Curation

Data is the fifth factor of production after land, labor, capital, and entrepreneurship. However, the volume, velocity, value, and veracity of the data are the challenges. Data analytics is an impressive tool to deal with such big data. But as the nature of data is nonhomogeneous i.e., scattered in the form of structured, unstructured and semistructured nature. The analytics on all kind of data is neither effective nor efficient, therefore to make the analysis worth, data curation is one of the extremely essential components in data management lifecycle. Some of the purposes of data curation are as follows:

- The sole purpose of data curation is to make digital data reusable in the future.
- Data curation intends to minimize the risk of data loss as the nature of digital data is fragile.
- The data formats are changing regularly, due to this change the available data gets corrupted, data curation is helpful to cope with this issue.
- Data curation leads to easy acquisition and dissemination of the data.
- Another purpose of data curation is to provide easy access and discovery of data to the users.

Data Curation models and approaches

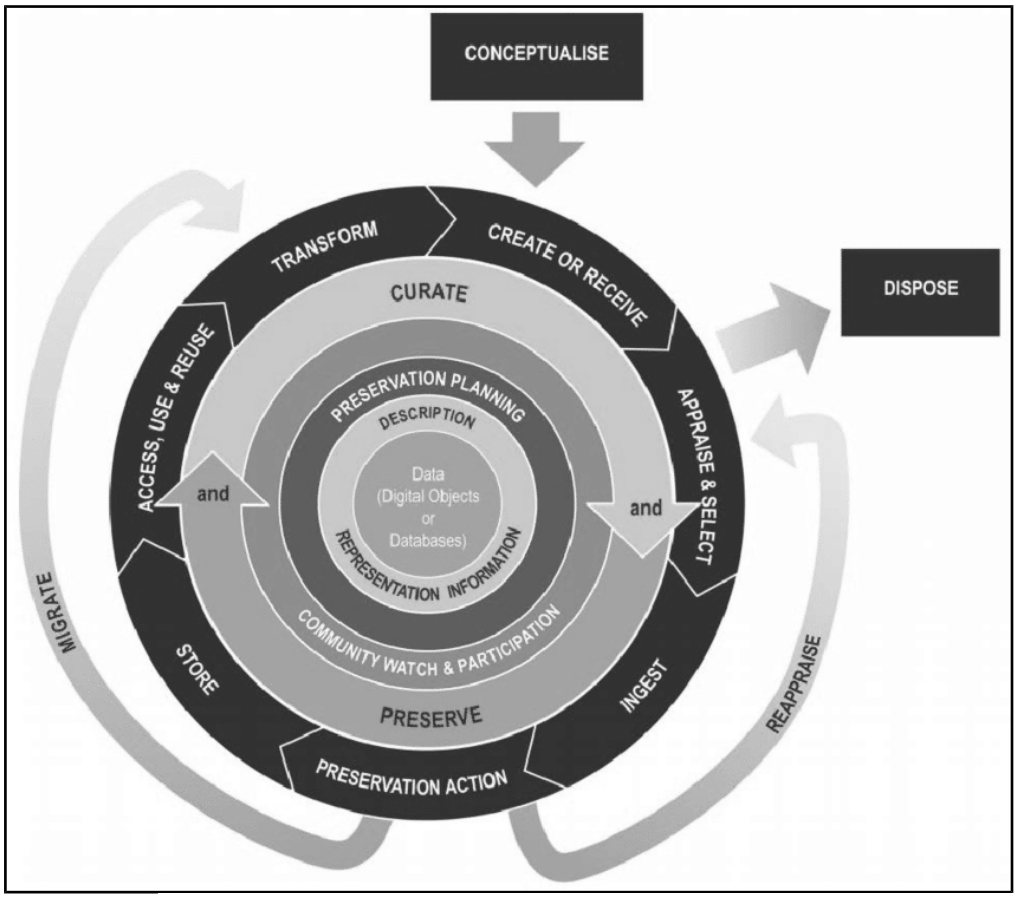
Since the emergence of the data curation concept, many workflow models and approaches evolved. As the curation depends on the data and the data is dependent on

the domain, so the workflow models are also domain dependent. Some of the domain-specific cases and workflow models are as follows:

- **ChemSpider:** for search engine data curation,
- **Protein data banks' workflow model:** For biological data curation
- **FoldIt:** For human computation and gamification data curation
- **The New York Times workflow and press association workflow model:** For the curation of media data
- **Ebay's model:** For the curation of e-Commerce data
- **DCC lifecycle model:**

Digital curation centre (DCC) has adopted a powerful lifecycle model which is more or less domain independent. It is the most accepted data curation model. It provides a graphical overview of the steps involved in data curation and digital preservation. The DCC curation lifecycle is shown in (figure 1). Entire DCC graph can be divided into three kinds of actions:

- Full lifecycle action
- Sequential actions
- Occasional actions



[Figure 1] (**Source:**https://www.researchgate.net/figure/The-DCC-digital-asset-lifecycle-model-Full-lifecycle-actions-are-activities-that-need-to_fig2_315861215)

Full lifecycle actions: Innermost circle indicates about the data which is taken from some database or is a digital object. First, four inner circles are denoted by the full lifecycle action. In this part the performed actions are:

Description and representation information: In this step metadata is assigned to the data to ensure the long-term preservation and discovery.
Preservation planning: This step is about the planning for the management of whole lifecycle actions
Community watch and participation: This step is to keep the eyes of the actions of the related community and participation in shared activities
Curate and preserve: THis step is to take the action for management and to promote curation and preservation.

Sequential Actions: In DCC lifecycle approach of data curation The just outer circle of full lifecycle action is called a sequential action. This action consists of the following processes:

Conceptualize: Conceive and plan the creation of data.
Create or receive: In this step, the data is created with its respective metadata. In another case, if the data is received from other sources assign it the metadata.
Appraise and select: Data is evaluated and selected for long-term curation and preservation
Ingest: The selected data is transferred to the archive or repository
Preservation action: To ensure long-term preservation the various preservation actions has been done
Store, access: Store the data in a relevant format
Use and reuse: Make the data publicly accessible for the further use and reuse i.e., publish the data on open platforms.
Transform: The final action is to transform the earlier data into new data which will lead to a continuous life cycle.

Occasional actions: This action makes connections between the processes of sequential actions It consist of the following processes:

Dispose of: Sometimes data is disposed of due to some policy, guidelines and legal reasons. The data which are not selected for long-term preservation also disposed of.
Reappraise: Return the data which fails validation procedure.
Migrate: After transformation, the data could be migrated in some other format and preservation action is done.

Technologies For the Data Curation

There are many established and widely adopted technologies are used in data curation activity. Some of them are as follows:

Sheer Curation: It is also called as curation at the source. In this approach, the lightweight curation activities are integrated with the data creating sources itself. It can do the lightweight normalization and categorization of the data

Collaboration Spaces: Like the CMS(Content management systems) and wiki platforms, it allows the users to curate the data collaboratively. It involves the people to curate both Structured and unstructured data.

Master Data Management (MDM): Freitas, A. and Edward, C. explain MDM as follows *"Master Data Management (MDM) tools can be used to remove duplicates and standardize data syntax, as an authoritative source of master data"*. There are 163 organizations which use MDM approach for the business impact (Rowe, 2012). The following processes are included in the MDM approach:

"source identification, data transformation, normalization, rule administration, error detection and correction, data consolidation, data storage, classification, taxonomy services, schema mapping, and semantic enrichment" (Freitas & Edward, 2016).

Data Curation Platforms

Data curation platforms are listed in Table 1

Sr. No.	Features	ZenCrowd	CrowdDB	Data Tamer	Qurk	Wikipedia Bots
1.	**Primary Aim**	To deal with the problem of linking named entities in text	To give the sql results not given by RDBMS	To introduce the automated data integration in place of the developer centric (ETL) process	To improve the efficiency of CrowdDB	To get the quality of text articles, defined as Bots
2.	**Major Functions**	Improve the result of automated linking	Applies fuzzy operations with the human help	Schemas are Automatically mapped and entities are de-duplicated	Automatic sorting & joining	Wikipedia editor is recommended for the flagged articles
3.	**Based on**	Knowledge base	Database	Database	Database	Database
4.	**Nature**	Manual & Automated	Manual & Automated	Automated	Automated	Manual & Automated
5.	**Deals With**	Both Structured & Unstructured data	Both Structured & Unstructured data	Both Structured & Unstructured data	Both Structured & Unstructured data	Both Structured & Unstructured data

[Table 1]

Other than these platforms there are other tools which are used for data curation e.g., OpenRefine, Tableau, Data-Driven Documents(D3.js).

OpenRefine, Tableau, and D3.js are open source tools and very popular in big data curation.

Issues and challenges

There are several challenges involved in data curation. Some of them are as follows:

- Data curation is a lengthy process which requires proper funding.
- Data curation involves both machine and human intellect and skill. Trained and skilled curators are relatively less than the requirement
- Which standard should be followed for long-term preservation of the data is another challenge.
- The selection of a better storage device is another issue for data curation.
- Technology is changing day by day, on which technology a data curator should rely?
- Licencing, intellectual property and distributed ownership of the data are some of the other important issues for the data curators
- Data curation is domain dependent activity. The selection for the different domain metadata is also a big challenge

Data Curation and Metadata

Metadata is the information about the data and as data are not self-describing the metadata is required to describe the data. The very purpose of metadata is resource discovery, therefore it is required for the better accessibility of the data or any other resource.

The sole aim of data curation is to facilitate the reusability of the data. In the era of big data, it is not possible to use the data without having its proper metadata. All kind of metadata (descriptive, administrative, structural, preservation and rights) are used in data lifecycle model.

Data curation is domain dependent therefore for the different kind of data the metadata standard also varies. For example, for the description of web resources, OAI-ORE *(Open Archive Initiative Object Reuse and Exchange)* and DCAT *(Data Catalog Vocabulary)* are used. For statistical and social science data DDI *(Data Documentation Initiative)* is useful while AGRIS *(International system for Agriculture Science and Technology)* is used for agricultural data.

Data curation and libraries

Researchers are the data creators and these data can be reused if it is curated properly. Once the data is created the job of data creators is over. Libraries are the experts into storing and managing the data in repositories. The data stored in the repository becomes junk if some value is not added to it. The curation activity adds value to the raw research data which can be further processed and reuse.

Libraries are the centre of information management. For the management, organization, and dissemination the of information, libraries and information center use the classification, cataloging and indexing techniques from the decades. Still, the advanced techniques like ontologies and metadata are the derivative of those initial classification and cataloging tools respectively. In data and information management these techniques are immensely used. Data curation deals with most of the activities which library professional are doing since ages. This expertise put library professional far ahead than the other managers in case of data management i.e., data curation. Data integration is another aspect of data curation. The application of linked data and ontologies are inevitable for effective data integration. The understanding and expertise of the linked data, ontologies and metadata have brought a huge opportunity in the area of data curation for the library professionals.

The concept of data librarian is evident to state the importance of library professionals in data management. Some of the American universities (e.g. University of New Mexico) having the data curation librarians, whose job is to manage whole data lifecycle.

Universities like *University of Illinois, Toronto University, the University of North Carolina at Chapel Hill* and many other institutions have already started the course on the data curation for their master's curriculum of Library and Information Management.

Conclusion

Data curation is the emerging need for the scholars and research communities. Data curation involves various activities which are not possible to do alone. The libraries are required to choose the coordinating approach to make the curation activity effective. As Weber and others state in the *Data Curation Research Summit Report 2010 "LIS will need to develop stronger partnerships with domain researchers, informaticists, and other stakeholders in the research enterprise, to succeed at making research data an integral and enduring part of the information assets retained for science and scholarship over the long term" (Weber et al. 2011)*

There are so many organizations and libraries who provide data curation services and training but data curation is still in the primary stage of its development. In developing countries like India, this concept is very new. There are various challenges and problems in data curation activity but these challenges should be accepted as data curation has more opportunities than the threats. On the data curation the remark of P. B. Heidorn is very relevant, he says:

"If libraries do not actively engage in the task, then society may choose to create a new type of institution to curate digital data" (Heidorn, 2011).

References

Cavanillas, J. M., Curry, E., & Wahlster, W. (2018). *New Horizons for a Data-Driven Economy A Roadmap for Usage and Exploitation of Big Data in Europe*. Retrieved October 12, 2018. (Book)

Freitas, A., & Curry, E. (2016). Big Data Curation. *New Horizons for a Data-Driven Economy,*87-118. doi:10.1007/978-3-319-21569-3_6

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration,51*(7-8), 662-672. doi:10.1080/01930826.2011.601269

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation,3*(1), 134-140. doi:10.2218/ijdc.v3i1.48

Jahnke, L., Asher, A., & Keralis, S. (n.d.). *The Problem of Data*. Retrieved October 15, 2018, from http://www.clir.org/wp-content/uploads/sites/6/pub154.pdf

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication,6*(1), 2198. doi:10.7710/2162-3309.2198

Martinez-Uribe, L., & Macdonald, S. (n.d.). *A new role for academic librarians: Data curation*.

Neuroth, H., Strahmann, S., Oßwald, A., & Ludwig, J. (2013). *Digital curation of research data experiences of a baseline study in Germany*. Retrieved October 18, 2018.

Rusbridge, C., Buneman, P., Burnhill, P., Giaretta, D., Ross, S., Lyon, L., & Atkinson, M. (n.d.). The Digital Curation Centre: A Vision for Digital Curation. *2005 IEEE International Symposium on Mass Storage Systems and Technology*. doi:10.1109/lgdi.2005.1612461

Data Curation Lifecycle Models. (2013,December 05). Retrieved September 21,2018, from https://ethicsofdatacuration.wordpress.com/data-curation-lifecycle-models/ You searched for data curation. (n.d.). Retrieved September 21,2018, from https://ropercenter.cornell.edu/?s=data%2Bcuration

Appraisal and Selection. (n.d.). Retrieved September 22,2018, from https://www.era.lib.ed.ac.uk/bitstream/handle/1842/3325Appraisal%20and%20Selection%20_%20Digital%20Curation%20Centre%239FEB.html?sequence=1&isAllowed=y

Data Life Cycle, Research Data Management | Boston University. (n.d.). Retrieved September 20, 2018, from https:www.bu.edu/datamanagement/background/data-life-cycle/

Eynden, V. V., & Corti, L. (2011). *Managing and sharing data: Best practice for researchers*. Colchester: UK Data Archive.

Charlesworth, A. (2008). Digital Curation, Copyright, and Academic Research. *International Journal of Digital Curation, 1*, 17-32. doi:10.2218/ijdc.v1i1.3

Chao, T. (2015). Mapping Methods Metadata for Research Data. *International Journal of Digital Curation, 10*(1), 82-94. doi:10.2218/ijdc.v10i1.347

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration, 51*(7-8), 662-672. doi:10.1080/01930826.2011.601269

W., M., N., C., P., L., C., V., . . . Tiffany C. (2011, November 17). Report on the Data Curation Research Summit. Retrieved from http://hdl.handle.net/2142/28355

Lenhardt, W. C., Ahalt, S., Blanton, B., Christopherson, L. & Idaszak, R. (2013). Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science. 10.6084/M9.FIGSHARE.791561.