A Brief Survey of Data Curation Literature

Lisa Palathingal Sergiu Dascalu Frederick C. Harris, Jr Yaakov Varol
Department of Computer Science and Engineering
University of Nevada, Reno
{lisa, dascalus, fredh, varol}@cse.unr.edu

Abstract

In today's scientific landscape, numerous fields of research are heavily dependent on data. The advent of big science has brought an impressive growth in the amount of data generated in many areas, which eventually led to a "data deluge". High quality data management is therefore vital in supporting high quality scientific research. In particular, if data pertaining to a current project is to be used in future research efforts, then proper data curation becomes essential. Data curation (DC) consists of a range of activities and processes aimed at preserving research data throughout its lifecycle. To gain a better understanding of DC and its challenges, we reviewed various scientific papers available in the literature as well as related case studies conducted by experts. This paper presents an overview of major data curation concepts, challenges, and approaches, proposes a taxonomy of major DC categories, and discusses main research trends in data curation.

Keywords: Data curation, data preservation, data management, models, frameworks.

1. Introduction

Our society is becoming more` and more dependent on data and most research processes are increasingly data driven.

Jahnke *et al* estimate that "Every day, we create 2.5 quintillion bytes of data – so much that 90% of the data in the world today has been created in the last two years alone" [1].

The advent of big science has brought a massive increase in the size and complexity of data across various fields of science [2]. High quality data management is therefore essential in order to ensure the sustainability of data. This points to the critical importance of data curation (DC) in the 21st century.

Data curation is generally defined as "the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use" [2, 3, 4]. This paper covers main concepts pertaining to data curation, the importance of data curation for effective data management, different models supporting data curation, and challenges faced by professionals involved in DC. The paper also proposes a classification of data curation approaches.

An example of a major initiative surveyed in this paper is based on the idea that it is essential to enrich graduate students with knowledge on data curation and its practices. The Data Curation Education in Research Centers (DCERC) program [5] was designed to establish a model for data curation graduate education where students are guided by both science and data mentors.

The rest of the paper is structured as follows. In Section 2 we briefly describe the main data curation concepts and several different models that support data curation. A proposed taxonomy of data curation approaches is presented in Section 3, followed by a discussion of such approaches in Section 4. In Section 5 we identify main research trends in DC. In Section 6, we summarize and conclude our survey, and outline future work we plan to undertake.

2. Data Curation Concepts, Challenges, and Approaches

The value of data is increasing quickly in our society because data is fundamental for supporting research. Data that are produced as part of research processes could be reused as the input to further research. Hence, data preservation is important [6].

Data curation is an effective way for addressing *data management*. Data curation consist of a range of activities and processes focused on maintaining, preserving, and adding value to research data throughout its lifecycle and all these activities are the responsibility of data curators. An example of an environment that heavily relies on data curation is the World Data System (WDS) [4]. WDS has a large number of data centers responsible for curation of research data sets for the scientific community.

978-1-880843-98-7 / copyright ISCA, CATA 2015
March 9-11, 2015, Honolulu, Hawaii, USA

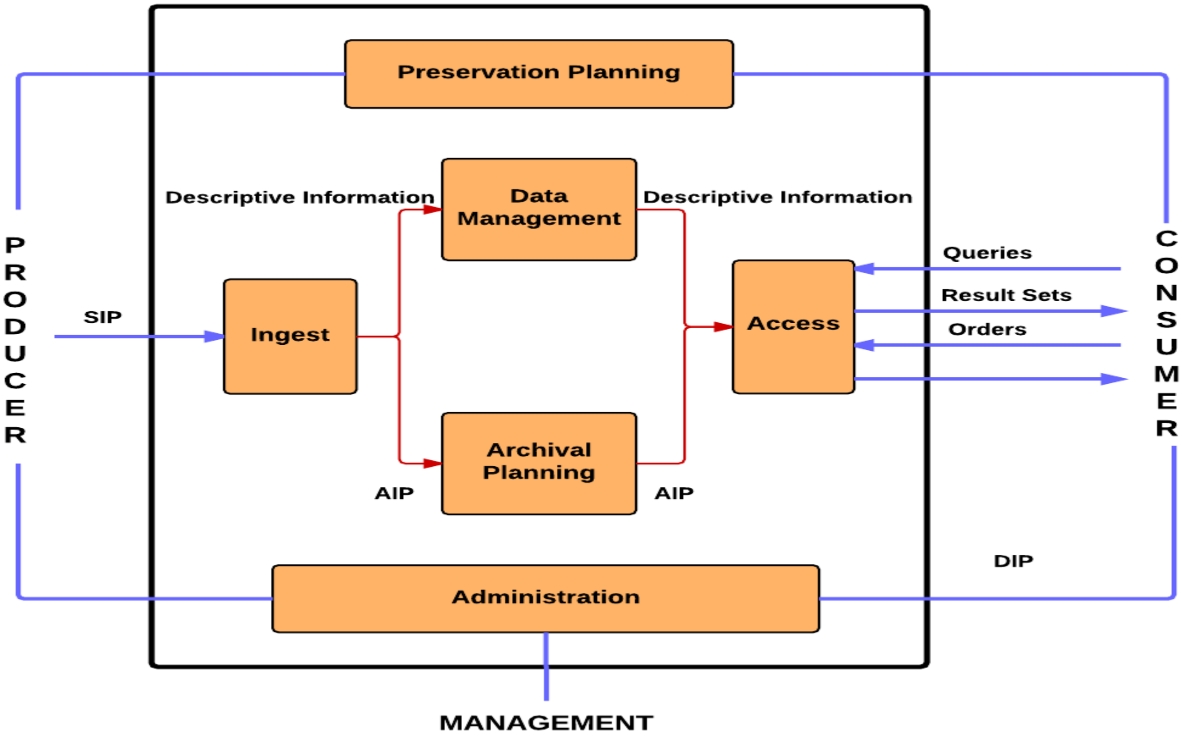

Figure 1: The OAIS Functional Model, adapted from [4]

A number of models and frameworks are being used for data curation. However, the only model approved by the International Organization for Standardization (ISO) is the Open Archival Information System (OAIS) Reference Model, and this model is used in the WDS.

The OAIS Reference model [4] consists of two sub-models: the information model and the functional model. The information model deals with the information objects and metadata that are used to preserve and access items in an archive. The functional model defines the six functions that are necessary for data curation, specifically Ingest, Archival Storage, Data Management, Administration, Preservation Planning and Access functions.

Ingest accepts data as Submission Information Packages (SIPs) and drives the data ready for archival storage and management. The Archival Storage function converts SIPs into Archival information Packages (AIPs). This conversion is essential to store, maintain and retrieve data. Data Management function aims at occupying, managing, and accessing Descriptive Information (DI) and administrative data. The Administration function is responsible for the overall operation of the archive. The Preservation Planning function monitors the environment in order to ensure that the data is accessible to the user group. The Access function supports the consumers to determine the existence, description, location and data availability within the archive. Figure 1 shows the OAIS functional model.

Even though OAIS is a well-designed, ISO approved approach, the model does have several shortcomings. According to Laughton and Du Plessis, the OAIS functional model would be more effective if it had included a phase prior to the Ingest function, Pre Ingest, to address the methods used for data collection [4]. Also, Pre Ingest would be very important as it aims to ensure quality, understanding, and accessibility of data. To overcome these challenges, a framework for data curation was proposed by WDS [4]. Framework includes two sections: digital and analogue. This two sided approach enables effective migration of analogue data to digital data.

Another lifecycle model that supports data curation is the Digital Curation Centre (DCC) Lifecycle model [4, 7]. Libraries, which play an important role in data curation, use the DCC lifecycle model [8]. The model provides an overview of the stages needed for successfully curating and preserving data [9]. The key element of the DCC Lifecycle model is the data. Data includes digital objects and databases. The DCC uses the model to help curators understand the processes involved in successful creation, and develop curation and preservation methodologies for their organizations. Figure 2 shows the well-known DCC Life-cycle model, often referred to in the related literature.

There are different types of actions in the DCC Lifecycle model [6]. They are *full lifecycle actions*, *sequential actions* and *occasional actions*. *Full lifecycle actions* take place at any time during the data curation lifecycle and are important to different sequential actions. Full lifecycle actions include description and representation information, preservation planning, community watch and participation, and curate and preserve.

Sequential actions are frequently considered to ensure that data curation is performed in its best form. Sequential actions include conceptualize, create and receive, appraise and select, ingest, preservation action, store, access, use and reuse, and transform.

Occasional actions interrupt or reorder sequential actions based on the decision. Occasional actions include dispose, reappraise and migrate [6].

Data curation that supports proper data management involves numerous challenges. In order to identify the barriers to data curation, the Council on Library and Information Resources (CLIR) [1] conducted interviews with faculty, graduate students and researchers in various fields. Some of the key findings were as follows:

- Most of the researchers working in the field did not receive any formal data management practices training, and they were not fully satisfied with their level of skills.
- There is a huge need for tools supporting large volume of generated data, and providing privacy and controls.

As pointed out by Jahnke *et al*, another issue pertaining to data management is the shortage of resources, as many organizations have limited financial and staff resources to support data curation. Furthermore, researchers working on the data could not always predict which data would be helpful in the future and therefore they are uncertain about which data should be preserved. As the protection of data privacy is becoming a bigger issue, it is essential to maintain control over who can access the data [1]. Some of the authors' recommendations to resolve these issues include [1]:

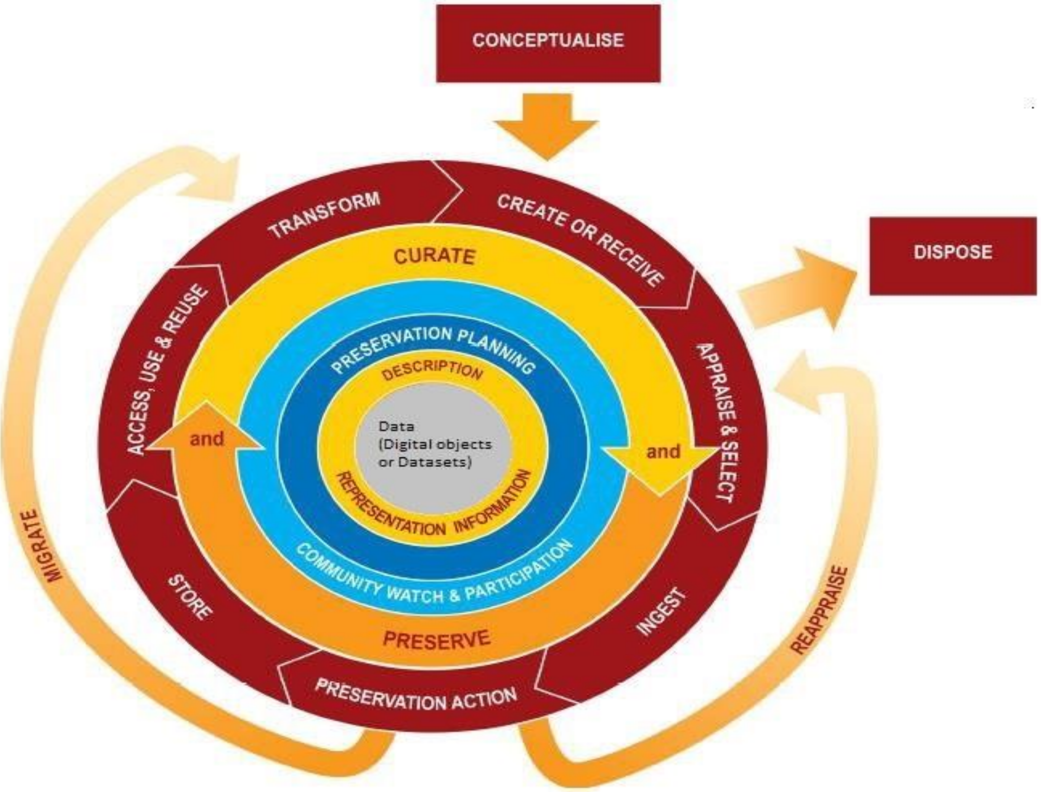

Figure 2: The Digital Curation Centre Lifecycle Model [4, 6, 8, 10]

- A single appropriate solution cannot be applied to the problem of data curation. An approach that highlights early engagement with the researchers and proper tools for data management and preservation would be most useful.
- Refined privacy and data access control are needed in order to avoid data loss.
- Help for researchers regarding the technological side of data management and preservation.

One approach to resolve the problem of data curation and its practices could be the establishment of dedicated institutions for data preservation and dissemination, such as survey data archives, and research data centers. To accomplish this for African countries, the University of Cape Town has established an data service called DataFirst [11]. This center aims at helping to improve data curation practices in Africa by providing data, promoting free curation tools, and undertaking data management training in African countries.

3. A Classification of Data Curation Approaches

We classify the data curation research and education into models, frameworks, case studies, and interviews. Figure 3 shows the proposed taxonomy of data curation approaches.

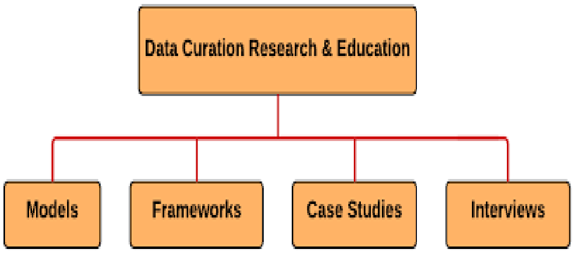

Figure 3: Proposed Taxonomy of Data Curation Approaches

For more details, Table 1 shows the different categories of data curation approaches along with examples of related references.

4. Discussion

Regarding *models*, two major examples of DC models are the Open Archival Information System (OAIS) model [4], and the Digital Curation Center (DCC) lifecycle model [4, 6, 7, 8, 9, 10].

Also, *frameworks* play an important role in supporting data curation. For example, the frame-work proposed in World Data System is based on the OAIS model and includes two sections: digital and analogue. This two-side approach enables effective migration of analogue data to digital data [4]. Another major framework used for mapping relationships and dependencies among scientific data practices, data types produced and used, and related curation activities has three categories: data, data practices, and curation [12]. According to Cragin *et al*, this framework is applied for identifying and representing curation requirements, supporting various descriptions, and assessing current or planned curation infrastructure and services.

Other categories of data curation approaches consist of the *case studies* and *interviews* conducted in order to understand the challenges of data curation as well as the practices that should be followed by researchers working in various fields. Amongst the challenges, one of the main issues is the lack of data management skills that characterizes many professionals working on research projects. In order to resolve this challenge, a proper training in data management practices is essential to the graduate students and researchers working in various fields. The Data Curation Education in Research Centers (DCERC) program [5] established a model to train the students by both science and data mentors. DCERC is also connected with research and development activities in the current NSF DataNet projects [15], which provides the students an opportunity to interact with working teams.

5. Main Research Trends

From the review of related literature, one of the main research trends in data curation is its application to the fields of library and information science [16, 17, 18, 19]. Important data curation aspects include developing interoperable standards for describing and interchanging datasets, the necessity for data curators to actively participate in data privacy and ownership policy development, the demand for a workforce skilled in data curation practices, and the need for different approaches for professional education required by a data-driven research agenda in sciences and humanities.

Data curation is not an activity that will be isolated to libraries or similar organizations—rather, it will be an activity that will be necessary in many other environments in which preserving data is important.

DC will require a collaborative effort that involves the application of a range of data management skills, beginning with the planning of research and encompassing the full data lifecycle. Researchers that specialize in data curation will need to be active in many kinds of organizations that generate and re-use data, including in more traditional environments such as libraries, archives, and data centers.

A global trend in data curation is the emergence of data-intensive science [13], which involves the collection of increasingly more data through automated systems such as sensor networks or large sets of instruments. This means that everyday new, huge amounts of data are becoming available for use by many different researchers.

Table 1: Data Curation Approaches

CATEGORY	REFERENCES
Models	
Open Archival Information System (OAIS) Reference Model	[4]
Digital Curation Center (DCC) Lifecycle Model	[4, 6, 7, 8, 9, 10]
Frameworks	
Data Curation in the World Data System: Proposed Framework	[4]
Relating Data Practices, Types, and Curation Functions: An Empirically Derived Framework	[12]
Case Studies	
Data Curation in the World Data System: Proposed Framework	[4]
Interviews	
The Problem of Data	[1]
Data Curation in Scientific Teams: An Exploratory Study of Condensed Matter Physics at a National Science Lab	[2]
Other (Education Oriented)	
The Problem of Data	[1]
Data Curation Education in Research Centers	[5]
Current Trends and Future Directions in Data Curation Research and Education	[13]
Education for Data Professionals: A Study of Current Courses and Programs	[14]

In the United States, national trends include the increased requirements from research funding agencies for enhanced data management, and the establishment of national centers for providing expert advice and services on data preservation, curation, collection, and reuse [13].

6. Conclusions and Future Work

Currently, numerous researchers are increasingly dependent on data, a reality reflected in the growing importance of effective data management. The sharing and reuse of current research data for future related or similar projects is becoming increasingly more important. Consequently, it is essential for researchers to understand data curation together with its main concepts, activities, challenges, and approaches.

Based on a review of related literature, in this paper we described the main data curation concepts and activities and emphasized the importance of data curation for effective data management. We also overviewed several major models that support data curation and identified some of the challenges faced by data curators. Further, we proposed a taxonomy of data curation approaches and explored existing solutions aimed at addressing data curation challenges.

As researchers in computer science and engineering, we believe that data curation also plays an important role in various computer science fields such as artificial intelligence, human computer interaction, software engineering, and computer vision. Hence, it is essential to enrich all the students and researchers working in these fields with knowledge on data curation and its practices.

Informed by the work on this paper, we plan to conduct further research on data curation challenges, techniques, practices, and tools, and develop a set of software tools that will assist the NSF-funded project [20] in which are involved with data curation activities.

Acknowledgement

This material is based upon work supported by the National Science Foundation under grant numbers IIA-1301726 and IIA-1329469.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] L. JAHNKE, A. ASHER, S. KERALIS, C. HENRY. 2012. The Problem of Data. *Council on Library and Information Resources.*

[2] C. HINNANT, B. STVILIA, S. WU, A. WORRALL, K. BURNETT, G. BURNETT, M. KAZMER, P. MARTY. 2012. Data Curation in Scientific Teams: an Exploratory Study of Condensed Matter Physics at a National Science Lab. *In Proceedings of the 2012 iConference*, 498-500.

[3] P. LORD, A. MACDONALD, L. LYON, D. GIARETTA. 2004. From Data Deluge to Data Curation. *In Proceedings of the 3rd UK e-Science All Hands Meeting*.

[4] P. LAUGHTON, T. DU PLESSIS, Data Curation in the World Data System: Proposed Framework. 2013. *In Data Science Journal*, vol. 12, 56-70.

[5] C. PALMER, S. ALLARD, M. MARLINO. 2011. Data Curation Education in Research Centers. *In Proceedings of the 2011 iConference*, 738-740.

[6] DCC Curation Lifecycle Model, (accessed November 26, 2014), http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[7] S. HIGGINS. 2008. The DCC Curation Lifecycle Model. *In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '08), 453-453.

[8] P. BRYAN HEIDORN. 2011. The Emerging Role of Libraries in Data Curation and E-science, *In Journal of Library Administration*, vol. 51, Issue 7-8, 662-672.

[9] S. HIGGINS. 2011. Digital Curation: The Emergence of a New Discipline. *The International Journal of Digital Curation*. Issue 1, vol. 3, No. 2.

[10] G. GOTH. 2012. Preserving Digital Data. *In Comm. of the ACM*, Issue 4, vol. 55, 11-13.

[11] H. WOOLFREY, Innovations for the curation and sharing of African social survey data. 2013. *In Data Science Journal*, vol. 12, 185-188.

[12] M. CRAGIN, C. PALMER, T. CHAO. 2010. Relating Data Practices, Types, and Curation Functions: an Empirically Derived Framework. *In Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (ASIS&T '10), vol. 47, No. 163, 738-740.

[13] Data-intensive science – trends in data reuse and management (accessed November 26, 2014), http://www.eresearch.org.nz/content/data-intensive-science-trends-in-data-reuse-and-management.

[14] V.VARVEL, E. BAMMERLIN, C. PALMER. 2012. Education for data professionals: a study of current courses and programs. *In Procs. of the 2012 iConference*, 527-529.

[15] C. LAGOZE, K. PATZKE. 2011. A Research Agenda for Data Curation Cyberinfrastructure, *In Proceedings of the 11th Annual International ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '11). 373-382.

[16] N. WEBER, C. PALMER, T. CHAO. 2012. Current Trends and Future Directions in Data Curation Research and Education. *In Journal of Web Librarianship*, vol. 6, Issue 4.

[17] J. CARLSON, D. LEITER. 2009. Addressing Researcher's Needs through the Data Curation Profile. *In Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '09), 365-366.

[18] C. THOMPSON, M. SENSENEY, K. BAKER, V. VARVEL, C. PALMER. 2013. Specialization in Data Curation: Preliminary Results from an Alumini Survey, 2008-2012. *In Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries* (ASIST '13), No. 152.

[19] Y. KIM, B. ADDOM, J. STANTON. 2011. Education for eScience Professionals: Integrating Data Curation and Cyberinfrastructure, *The International Journal of Digital Curation*, vol. 6, 125-138.

[20] The Nevada Research Data Center (accessed December 1, 2014). http://www.sensor.nevada.edu/NRDC/, Project funded by NSF EPSCOR.