Data Curation in Academic Libraries as Part of the Digital Revolution[1]

Tibor Koltay

ORCID 0000-0003-0548-7527

Institute of Learning Technologies, Eszterházy Károly University, Hungary

Abstract

Purpose/Thesis: This paper outlines the role of data curation in the context of Research 2.0 and Research Data Management.
Approach/Methods: The argument is based on a non-exhaustive review of the literature.
Results and conclusions: Despite the relative vagueness and variety of definitions of data curation, academic libraries should engage in it.
Research limitations: The study focused mainly on theoretical writings.
Practical implications: The worldwide challenge associated with Research Data Management and data curation. Several countries and institutions have already answered the challenge, but the overall level of its recognition is low, and thus there is a need to raise awareness of its importance.
Originality/Value: The premise of the argument is based on the assumption that views on data are changing.

Keywords:

Data curation. Digital curation. Research 2.0. Research data management.

1. Introduction

As asserted by Michael Buckland,

> the potentially useful record of science comes in the form of (mainly non-textual) digital data sets, but the infrastructure behind it is not yet comparable to the one of print-on-paper materials and their digital versions (Buckland, 2011, 35).

The expression *Research 2.0* indicates this growing importance of data, especially in regard to research data that requires new approaches to scholarly research that influences the natural sciences, the social sciences and the humanities. It is clearly associated with the digital revolution, as without easy access to a wealth of information and data, enabled by a well-developed information infrastructure, openness of science would not be possible. Neither would have new disciplines, such as the digital humanities, been born.

If we agree that Research 2.0 is characterized by the recognition that sharing and re-using research data is inevitable, we also agree with the views, promulgated by the proponents of

---

information ecology, who are committed to providing a new framework for interpretation of relationships between social actors, information technologies, and information objects (Steinerová, 2012).

This review paper is in many regards a continuation of this author's earlier paper in this journal (Koltay, 2017). It focuses on data curation, which is part of Research Data Management (RDM) that needs to be cared for by researchers, as well as by academic and special libraries across the world In order to contribute to a better understanding of the nature of data curation, the conceptual analysis covers the effect of new views on data, some features of the changing academic library environment and varied approaches to data curation itself.

2. The environment

Today, we can clearly see an increase in the attention towards data literacy, which shares several features with other literacies and is especially closely connected to information literacy (Koltay, 2015). The nature of this interest is influenced by the changing ways of seeing data, compared to information. As a result of this and owing to the importance of research data, never seen before, working climate of academic library services has begun to change.

2.1. New views on data

Describing the aforementioned development would be impossible without contemplating the changes in the ways of seeing data. Library and Information Science (LIS) has traditionally perceived data as somewhat secondary (or even inferior) to information, because it was regarded as occupying the bottom of the data–information–knowledge–wisdom pyramid (Rowley, 2007). However, the growing importance of data may change this view (Cox, 2018), inducing views that are far from seeing the relationship between these fundamental concepts as simple as it has been presented by earlier views (Makani, 2015), especially as data is not only a concept that appears with much higher frequency than ever before in the history of LIS but is "heavily loaded with epistemological problems" (Hjørland, 2019, 686). Nonetheless, some definitions clearly emphasize the closeness of data and information. For instance, Liangzhi (2015) affirms that from the ontological point of view, data and information are close to each other as both exist as signs. Stating that data is anything recordable in a semantically and pragmatically sound way, Frické (2008) indicates the potential of comprehensive approaches to defining data.

As this paper focuses on academic libraries' role mission, it is necessary to address the more specific idea of research data, i.e. data collected as part of a research project that might be contrasted to data collected or digitized for curation and preservation, without forgetting that data collected for similar reasons can become research data (Robinson, 2016). Research data can be seen not just as the result of empirical work or the raw material for statistical analysis, but also as a research object in its own right (Golub & Hansson, 2017). Furthermore, in the digital humanities there seems to be an implicit agreement that data can be interpreted as texts, and – conversely – that texts can be interpreted as data (Koltay, 2016).

Even new and utterly simple views on data can find their legitimate use and appropriate place. For instance, the Digital Curation Centre (2018) defines data as "any information in

binary digital form". To this definition, we can add that documents not born digital may become digital at some point (Robinson, 2016). If we take another, principally practice-oriented view, we can say that data constitutes a primary intellectual asset that can be subjected to peer review and other forms of quality assessment, and is open to reuse (Heidorn, 2011).

Approaches that conceive data differently are also influenced by a new definition of information literacy (IL), which states that IL is related not only to print, but to data, images and the spoken word (CILIP, 2018). This definition recognizes the abovementioned close relationship between data and information, and indicates information literacy's overlap with other literacies.

2.2. *The changing academic library environment*

There is a host of varied and often synonymous or vaguely defined data-related terms, such as research data management, data curation, preservation, and data stewardship, used in library environments. Having chosen one of these in its policy document, the European Open Science Cloud sets the requirement for European science to be

> grounded in a common culture of data stewardship, so that research data is recognized as a significant output of research and is appropriately curated throughout and after the period conducting the research (EOSC, 2017,1).

This statement reminds us of the fact that – differing form the case of Open Access to journal articles – the worldwide challenge to give the appropriate attention to research data is far from being properly recognized in some countries, and that its recognition has not been evenly spread yet. However, when acknowledged, this challenge is often answered by academic libraries in many countries by providing a wide array of RDM services. Notwithstanding, there is still a need to raise awareness of its importance.

Another issue that we should not forget about is that our information environment is not only dominated by the importance of data, underlined in the above document. We must also see that the quantity of data available is overwhelming, and therefore we cannot be mistaken if accepting that – beside information overload – we witness the phenomenon of data overload or data deluge (Little, 2012). This means that we are overloaded by different kinds of data. Such a situation calls for efficient data curation and a high level of data literacy education (Robinson, 2016).

Data curation has been identified as both a challenge and an opportunity for

> finding new ways to communicate the value of the skills librarians already possess and in developing roles that were previously not associated with librarians (ACRL 2012, 312).

The issue becomes especially important if we consider that one of the academic librarians' roles is to support researchers (Mierzecka, 2019).

Data curation activities should be guided by the understanding that researchers do not need more data, but the right data (Borgman, 2015) and that data is of no use if not is not analyzed. Accordingly, librarians have to acquire analytical skills and abilities (Kirkwood, 2016). They also should set up appropriate and innovative new services, recognized by several stakeholders, including research funders, research organizations, the communities of researchers, as well as academic and special libraries. The latter two are crucial actors,

representing the human element in data curation, declared of uttermost importance by Poole (2015). Librarians' participation is of real value as libraries are integral part of the necessary institutional structure and only a few librarians are qualified to perform the related tasks (Griffin, 2013). Obviously, data management initiatives must acquire backing from the whole staff, but the role of the libraries' managements should not be underestimated either (Burton & Lyon, 2017).

3. Approaches to data curation

The place of data curation in academic libraries' workflow is not defined precisely. Nonetheless, it is one of the potential RDM activities, which ideally begin with data literacy instruction, followed by RDM itself, then by data curation and data preservation (Thomas & Urban, 2018). Digital curation can be seen as an umbrella term for data curation and digital preservation. Its main goals are preventing data loss, and adding value to trusted data assets for current and future use (Poole, 2016).

By subsuming digital preservation and supplying different kinds of documentation to provide context, as well as including metadata, data curation aims to make selected data accessible, usable, and useful throughout its lifecycle (Giarlo, 2013). As Burgi, Blumer, and Makhlouf-Shabou (2017) put it, the preservation of digital information is a complex and costly process that cannot be circumvented. Consequently, we can say – in general terms – that curation involves planning, acquiring, preparing, analyzing, preserving and discovering the data (Pouchard, 2016).

Curation in analogue or digital form is not only a domain of librarians, who manage libraries as organized collections within a certain place and space. Curators work in a variety of institutions, who care for a host of artefacts. Somewhat differently, archivists care for the permanency of records and their historical value. The definitions of digital curation and data curation converge as they emphasize the processes related to taking care of resources, including organization and description (Kouper, 2016). Therefore, if we regard data (as we did above) as any information in digital form (sic!), then data curation is identical with digital curation, i.e. they both consist of the management and preservation of digital material (Robinson, 2016).

While researchers inevitably need to curate their own data, most of them do not receive formal training in data management practices, and thus they are learning data management and curation on the job. Moreover they do it in an ad hoc fashion, because it is secondary to their main occupation, i.e. doing research (Jahnke et al., 2012). In a broader context, mandating open access to data may pressure researchers and research teams to provide data, even if intellectual property or data protection rights would specifically and explicitly limit their ability to do so, or it may aggravate ethical concerns about sharing research data in general (RECODE, 2014). Helping researchers to solve such dilemmas matches the librarians' skill portfolio that traditionally contains open access and copyright (Southall & Scutt, 2017).

Furthermore, data curation can be seen from different angles. If we emphasize the role of data repositories, data curation can be defined as the

> encompassing work and actions taken by curators of a data repository, in order to provide meaningful and enduring access to data (Johnston et al., 2018, 5).

As said above, data curation can also be seen as a part of RDM, which might be defined as a set of general activities, not specifically attached to the library, but potentially performed by it, involving organizing, collecting, describing, and storing data for communication among scholars. Accordingly, the emphasis of data curation is not on internal storage, but gives priority to extracting data for general purposes (public use) is stress, in which it is similar to curating and exhibiting collections in museums (Thomas & Urban, 2018). In general, knowledge and skills that may prove useful for data curation are among the ones that librarians are qualified to have and can be required to make use of (Burgi et al., 2017).

In an online survey of professionals, who identify themselves as digital curators Kouper (2016) found preservation as the most frequent theme. Slightly less frequently mentioned terms were access and dissemination, and the lifecycle approach. If we subscribe to the idea that digital curation and data curation are closely related to each other, we can suppose that the tasks of both types of curation are identical to a considerable degree.

While data curation, similarly to data literacy and RDM has to focus on data quality (Ridsdale et al., 2018), it goes beyond RDM, because it additionally comprises of preservation and enables reuse (Partlo et al., 2015).

In order to be successful, data curation requires management of digital data over its entire lifecycle, including the use of varied services, tools, and infrastructure (Schmidt & Shearer, 2016) to

> ensure the maintenance of authenticity, reliability, integrity and usability of digital material, which in turn ensures maximization of the investment in their creation (Higgins, 2008, 135.)

There is a close connection between technical aspects of RDM, such as preparing datasets for deposit into a repository and deaccessioning them from depositories, providing persistent storage, assign unique identifiers and data curation (Cox & Pinfield, 2014). However, the latter activities appear less frequently than informational (consultative) services, such as offering guidance on copyright, helping researchers or students in setting up Data Management Plans, or consulting on metadata standards (Tenopir et al., 2016; Tenopir et al., 2017).

An analysis of several data curation and digital curation curricula in the United States shows the presence of varied topics and broad scope of issues that include data types, standards, lifecycles, provenance, metadata, collection development, as well as digital preservation. Digital scholarship in the humanities also appears in the list of subjects (Harris-Pierce & Quan Liu, 2012). This reminds us of the decisive impact of data-intensive research on the digital humanities, mentioned at the beginning of this paper.

4. Conclusion

Data curation can be seen as a new label, conceived to name already existing activities that have been practiced in libraries, museums, and archives (Ray, 2012). In libraries, this can be observed if we understand that the roles of librarians in general and the duties of subject librarians in particular already entail activities designated as data curation (Bracke, 2011). No need to say that this is also true for data librarians. Still, it can be expected that data curation will become more than a label and will catch up with research data management

in meeting the demand of serving researchers, teaching staff members and students, as soon the need for such services arises.

Offering data-related services, including data curation, may prove useful for transcending the – otherwise respectable – role of the library in relieving researchers' technical and administrative burdens (LERU, 2013). In other words, data curation is not only one of the professional tasks for libraries on the to-do list, but it might serve to demonstrate that they play an essential role in research processes (Scaramozzino et al., 2012).

We should know that supporting researchers involves much more than data curation and RDM. Such activity might be, for instance Data Science Support, which is an emerging field of activities, provided by academic libraries (Oliver et al., 2019).

In order to keep the views and library practice related to research data up to date, librarians should constantly look out for new approaches and methods brought in to managing research data by data science, which – despite of its different origin and goals – may turn out to be applicable to data librarianship (Cox, 2018).

If we said that academic libraries' involvement in curating research data may be looked at and appreciated differently, it is worth to add that already in 2012 a research team at Colorado State University Libraries found that researchers require quality assistance and

> do not necessarily have preconceived perceptions that the libraries are poorly equipped to offer assistance (McLure et al., 2014, 157).

This shows that data-related activities may involve not only new tasks, but enhanced recognition of their work as well.

Even though it may sound strange, data curation-as-publishing is a promising new way for libraries to cooperate with digital humanists as it consists of curation not only in the sense of registering and making it public, but making it available and suitable for re-use (Muñoz, 2013).

Last, but not least, librarians should not forget that researchers often – and in many cases rightfully – may perceive themselves as owners of research data. Therefore librarians have to ensure them that libraries will be responsible stewards of their information, without taking ownership of this data, while original owners will be able to access their access content or take it out of library systems at any point and in-perpetuity (Lucky & Harkema, 2018).

References

ACRL (2012). 2012 Top Ten Trends in Academic Libraries: A Review of the Trends and Issues Affecting Academic Libraries in Higher Education. *College & Research Libraries News*, 73(6), 311–320.

Borgman, Ch. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Bracke, M. S. (2011). Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data. *Journal of Agricultural & Food Information*, 12(1), 65–74.

Buckland, M. (2011). Data Management as Bibliography. *Bulletin of the American Society for Information Science and Technology*, 37(6), 34–37.

Burgi, P. Y., Blumer, E., Makhlouf-Shabou, B. (2017). Research Data Management in Switzerland: National Efforts to Guarantee the Sustainability of Research Outputs. *IFLA Journal*, 43(1), 5–21.

Burton, M., Lyon, L. (2017). Data Science in Libraries. *Bulletin of the Association for Information Science and Technology*, 43(4), 33–35.

CILIP (2018). *CILIP Definition of Information Literacy 2018* [online]. CILIP. The Library and Information Association, [16.05.2019], https://infolit.org.uk/ILdefinitionCILIP2018.pdf

Cox, A. M. (2018). Academic Librarianship as a Data Profession: The Familiar and Unfamiliar in the Data Role Spectrum. *eLucidate*, 15(1/2), 7–10.

Cox, A. M., Pinfield, S. (2014). Research Data Management and Libraries: Current Activities and Future Priorities. *Journal of Librarianship and Information Science*, 46(4), 299–316.

DCC (2018). DCC Curation Lifecycle Model [online]. Digital Curation Centre, [16.05.2019] http://www.dcc.ac.uk/resources/curation-lifecycle-model

EOSC (2017). *EOSC Declaration* [online]. European Open Science Cloud, [16.05.2019], http://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none

Frické, M. (2008). The Knowledge Pyramid: a Critique of the DIKW Hierarchy. *Journal of Information Science*, 35(2), 131–142.

Giarlo, M. (2013). Academic Libraries as Quality Hubs. *Journal of Librarianship and Scholarly Communication*, 1(3): 1–10.

Golub, K., Hansson, J. (2017). (Big) Data in Library and Information Science: a Brief Overview of Some Important Problem Areas. *Journal of Universal Computer Science*, 23(11), 1098–1108.

Griffin, S. (2013). New Roles for Libraries in Supporting Data-Intensive Research and Advancing Scholarly Communication. *International Journal of Humanities and Arts Computing*, 7(Suppl.), 59–71.

Harris-Pierce, R. L., Quan Liu, Y. (2012). Is Data Curation Education at Library and Information Science Schools in North America Adequate? *New Library World*, 113(11/12), 598–613.

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and e-Science. *Journal of Library Administration*, 51(7–8), 662–672.

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, 3(1), 134–140.

Hjørland, B. (2019). Data (with Big Data and Database Semantics). *Knowledge Organization,* 45(8), 685–708.

Jahnke, L., Asher, A., Keralis, S. D. (2012). *The Problem of Data*. Washington, DC: Council on Library and Information Resources.

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C. (2018). How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 6(1), p.eP2198.

Kirkwood, R. J. (2016). Collection Development or Data-Driven Content Curation? *Library Management*, 37(4–5), 275–284.

Koltay, T. (2015). Data Literacy: in Search of a Name and Identity. *Journal of Documentation*, 71(2), 401–415.

Koltay, T. (2016). Library and Information Science and the Digital Humanities: Perceived and Real Strengths and Weaknesses. *Journal of Documentation*, 72(4), 781–792.

Koltay, T. (2017). Beyond Literacies: The Evolving Landscape of Library Support to Research 2.0. *Zagadnienia Informacji Naukowej. Studia Informacyjne*, 55(1), 7–18.

Kouper, I. (2016). Professional Participation in Digital Curation. *Library & Information Science Research*, 38(3), 212–223.

LERU (2013). *LERU Roadmap for Research Data* [online]. League of European Research Universities Research Data Working Group, [16.05.2019], https://www.leru.org/publications/leru-roadmap-for-research-data

Liangzhi, Y. (2015). Back to the Fundamentals Again. *Journal of Documentation*, 71(4), 795–816.

Little, G. (2012). Managing the Data Deluge. *Journal of Academic Librarianship*, 38(5), 263–264.

Lucky, S., Harkema, C. (2018). Back to Basics: Supporting Digital Humanities and Community Collaboration Using the Core Strength of the Academic Library. *Digital Library Perspectives*, 34(3), 188–199.

Makani, J. (2015). Knowledge Management, Research Data Management, and University Scholarship: Towards an Integrated Institutional Research Data Management Support-System Framework. *VINE*, 45(3), 344–359.

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., Culbertson, M. (2014). Data Curation: A Study of Researcher Practices and Needs. *Portal: Libraries and the Academy*, 14(2), 139–164.

Mierzecka, A. (2019). The Role of Academic Libraries in Scholarly Communication. A Meta-Analysis of Research. *Studia Medioznawcze*, 19(1), 42–55.

Muñoz, T. (2013). Data Curation as Publishing for the Digital Humanities. *Journal of Digital Humanities*, 2(3), 14–22.

Oliver, J. C., Kollen, C., Hickson, B., Rios, F. (2019). Data Science Support at the Academic Library. *Journal of Library Administration*, 59(3), 241–257.

Partlo, K., Symons, D., Carlson, J.D. (2015). Revolutionary or Evolutionary? Making Research Data Management Manageable. In: B.L. Eden (ed.). *Creating Research Infrastructures in the 21st Century Academic Library: Conceiving, Funding, and Building New Facilities and Staff* (pp. 175–201). Lanham, MD: Rowman and Littlefield.

Poole, A. H. (2015). How Has Your Science Data Grown? Digital Curation and the Human Factor: a Critical Literature Review. *Archival Science*, 15(2), 101–139.

Poole, A. H. (2016). The Conceptual Landscape of Digital Curation. *Journal of Documentation*, 72(5), 961–986.

Pouchard, L. (2016). Revisiting the Data Lifecycle with Big Data Curation. *International Journal of Digital Curation*, 10(2), 176–192.

Ray, J. (2012). The Rise of Digital Curation and Cyberinfrastructure: From Experimentation to Implementation and Maybe Integration. *Library Hi Tech*, 30(4), 604–622.

RECODE (2014) *RECODE Policy Recommendations for Open Access to Research Data* [online]. RECODE Project Consortium, [16.05.2019], https://trilateralresearch.co.uk/wp-content/uploads/2018/09/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D. (2015). *Strategies and Best Practices for Data Literacy Education. Knowledge Synthesis Report.* Halifax, NS: Dalhousie University.

Robinson, L. (2016). Between the Deluge and the Dark Age: Perspectives on Data Curation. *Alexandria*, 26(2), 73–76.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, 33(2), 163–180.

Scaramozzino, J. M., Ramírez, M. L., McGaughey, K. J. (2012). A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. *College & Research Libraries*, 73(4), 349–365.

Schmidt, B., Shearer, K. (2016). *Librarians' Competencies Profile for Research Data Management* [online]. Joint Task Force on Librarians' Competencies in Support of e-Research and Scholarly Communication, [15.07.2019], https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf

Southall, J., Scutt, C. (2017). Training for Research Data Management at the Bodleian Libraries: National Contexts and Local Implementation for Researchers and Librarians. *New Review of Academic Librarianship*, 23(2–3), 303–322.

Steinerová, J. (2012). Information Ecology – Emerging Framework for Digital Scholarship. *Libraries in the Digital Age (LIDA) Proceedings* [online], 12, [16.05.2019] http://ozk.unizd.hr/proceedings/index.php/lida/article/view/66/37

Tenopir, C., Pollock, D., Allard, S., Hughes, D. (2016). Research Data Services in European and North American Libraries: Current Offerings and Plans for the Future. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–6.

Tenopir, C., Talja, S., Horstmann, W., Late, E. Hughes, D., Pollock, D. Schmidt, B., Baird, L., Sandusky, R. J., Allard, S. (2017). Research Data Services in European Academic Research Libraries. *LIBER Quarterly*. 27(1), 23–44.

Thomas, C. V., Urban, R. J. (2018). What Do Data Librarians Think of the MLIS? Professionals' Perceptions of Knowledge Transfer, Trends, and Challenges. *College & Research Libraries*, 79(3), 401–423.

Data curation w bibliotekach akademickich jako część rewolucji cyfrowej

Abstrakt

Cel/Teza: W artykule przedstawiono rolę *data curation* w kontekście Nauki 2.0 i zarządzania danymi badawczymi.

Koncepcja/Metody badań: Wykorzystano krytyczną analizę wybranego piśmiennictwa przedmiotu.

Wyniki i wnioski: Pomimo względnej niejasności i wielości definicji terminu "data curation", biblioteki akademickie powinny zaangażować się w związane z nim działania.

Ograniczenia badań: W artykule skupiono uwagę głównie na publikacjach opisujących podejście teoretyczne.

Zastosowanie praktyczne: Obserwowane na całym świecie przedsięwzięcia związane z zarządzaniem danymi badawczymi i *data curation*. Wiele krajów i instytucji podjęło już działania na tym polu, ale ogólny poziom rozpoznania tematu nadal pozostaje niewysoki, dlatego potrzebne jest zwiększenie świadomości dotyczącej znaczenia tego zagadnienia.

Oryginalność/Wartość poznawcza: Przesłanki przedstawione w artykule oparte są na założeniu, że zmienia się podejście w postrzeganiu danych.

Słowa kluczowe:

Data curation. Digital curation. Nauka 2.0. Zarządzanie danymi badawczymi.

Dr. habil. TIBOR KOLTAY, PhD is Professor, Chair of the Institute of Learning Technologies and Principal Researcher of the Information Society research group at Eszterházy Károly University, Hungary. His research focuses on information literacy and other literacies, as well as Research Data Management. His most recent publication is Accepted and Emerging Roles of Academic Libraries in Supporting Research 2.0 *(2019, Journal of Academic Librarianship, https://doi.org/10.1016/j.acalib.2019.01.001).*

Contact to the Author:
koltay.tibor@uni-eszterhazy.hu
Institute of Learning Technologies
Eszterházy Károly University, Hungary
Rákóczi út 53,
Jászberény,
H-5100, Hungary