IJDC | *Conference Pre-print*

Updating the DCC Curation Lifecycle Model

Sayeed Choudhury
Johns Hopkins University

Caihong Huang
University of Washington

Carole L. Palmer
University of Washington

Abstract

The DCC Curation Lifecycle Model has played a vital role in the field of data curation for over a decade. During that time, the scale and complexity of data have changed dramatically, along with the contexts of data production and use. This paper reports on a study examining factors impacting data curation practices and presents recommendations for updating the DCC Curation Lifecycle Model. The study was grounded in a review of other lifecycle models and informed by a site visit to the Digital Curation Centre and consultation with expert practitioners and researchers. Framed by contemporary conditions impacting the conduct of research and provision of data services, the analysis and proposed recommendations account for the prominence of machine-actionable data, the importance of machine learning for data processing and analytics, growth of integrated research workflows, and escalating concerns with fairness, accountability, and transparency of data and algorithms.

Submitted 16 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Sayeed Choudhury, 3400 N. Charles Street, Baltimore, MD 21218, USA, Email: Sayeed@jhu.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction and Background

For over a decade, the highly influential DCC Curation Lifecycle Model has played a vital role in the field of data curation, effectively communicating the activities involved in digital curation and guiding the conceptualization and development of curation services. Since its formal introduction in 2008, application contexts have changed dramatically, as have the scale and complexity of digital data and demands on data services offered by archives and repositories.

The DCC 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" (Digital Curation Centre, n.d.). Formally introduced by the Digital Curation Centre in 2008, the model is intentionally generic—indicative rather than exhaustive, adaptable to different domains, and applicable at different levels of granularity (Higgins, 2008). The lifecycle approach aims to ensure that essential curation stages are planned and implemented in sequence. However, as an "ideal", the model is necessarily modified in practice. Users "may enter at any stage of the lifecycle depending on their current area of need" and the context of use in an organization or domain (Digital Curation Centre, n.d.).

This research was motivated by changes in the conduct of research and current challenges in the profession not explicitly addressed by the DCC Curation Lifecycle Model (hereafter referred to as the DCC Model). The key research question is concerned with changes needed to update a fundamentally discrete, archival-based, and document-oriented model, given the rise of data science and new data intensive research methods. What elements of the DCC Model remain relevant? What new forms or modes of curation need to be accommodated given these changes? The research team made no a priori assumptions about the continued relevance of the DCC Model. We recognized, however, that other lifecycle models had been developed over time, while a systematic evaluation of the original DCC Model was lacking and would need to consider rapid changes in the context of curation, including the rise of data science, the growing connections between data archives and high-performance computing, and more automated methods for curation.

Based on our experience as curation professionals, researchers, and educators, we identified a set of contemporary conditions, which have become highly consequential since the development of the DCC Model, to serve as a framework for the study:

- increasing scope, complexity, and prominence of machine-actionable data
- importance of machine learning for data processing and analytics
- growth of integrated research workflows
- concerns with fairness, accountability, and transparency of data and algorithms

These framing conditions have significant implications for the future of data services. The scale and variety of data and the application of machine learning techniques have direct impacts, for example, on how and when data are packaged and described for deposit, access, and retrieval. Increasingly, curation and archiving will need to accommodate workflow technologies as they become more integrated into the production of research and its outputs, as evidenced by growing use of systems such as Taverna and Kepler for scientific workflow management, and Elsevier's acquisition of Mendely, SSRN, and BePress for publishing. Importantly, the first three conditions may have implications for the fourth, focused on combating bias and developing systems that promote fairness, accountability, and transparency (hereafter referred to as FAT).

These trends are also increasingly broad based as they escalate across disciplines, as seen in shifts in research approaches at Johns Hopkins University (JHU). Demands are increasing on

data services from new centers of research, such as the Institute for Data Intensive Engineering and Science. But, the scale and diversity of data are no longer associated primarily with the sciences and engineering. It is becoming commonplace for social scientists to work with vast collections of social media and mobile device data, and the humanities are evolving in similar ways. Notably, data-intensive humanities research and teaching have recently been identified as key priorities at JHU. In one exemplar case, Chris Cannon, Professor of Classics and English, worked with the University Libraries to apply machine learning techniques to the analysis of Chaucer concordances.

While there has been recognition that computing must be co-located with data ("bring the compute to the data rather than move the data to the compute"), our examination of the DCC Model was also informed by the need to bring data curation to the data within existing and evolving workflows. Choudhury's team at JHU has made progress in this direction through development of a high-level architecture that depicts various flows of articles and data across the institution. This complex set of workflows highlights the need to account for curation in a more iterative, continuous, and dynamic manner.

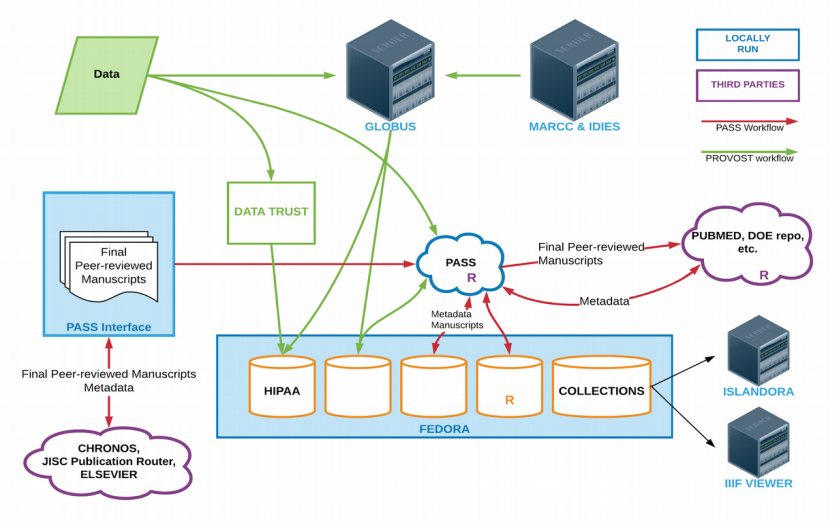

Figure 1. High-level architecture of research and publishing workflows at JHU. PASS is the Public Access Submission System; MARCC is the Maryland Advanced Research Computing Cluster; IDIES is the Institute for Data Intensive Engineering and Science; IIIF is the International Image Interoperability Framework.

In examining the implications of the framing conditions and advances in infrastructure in relation to the DCC Model, our aim has not been to design a new lifecycle model but rather to systematically consider factors impacting curation practices to develop explicit recommendations for updating the DCC Model.

Approach

The study was conducted in three phases: literature review, DCC site visit, and expert consultation. The literature review provided grounding in the broad range of related work on lifecycle models. The site visit to the Digital Curation Centre informed the team's understanding of the genesis and evolution of the original DCC Model. The expert consultants, representing practitioner and researcher constituencies, served as key informants on how the framing conditions are shaping data services and data science research and applications. These three phases of the study were supported by a grant from the Alfred P. Sloan Foundation. Additional

analysis was provided through a student capstone project at the University of Washington, dedicated to background research and synthesizing the sources of evidence. It should be noted that while Digital Curation Centre researchers and staff provided invaluable input, the study was conducted independently of the DCC.

Literature Review Scope

The literature review was essential in understanding the landscape of curation lifecycle models published since 2008, when the DCC Model was first introduced. To identify models with potential advances related to the framing conditions, the following criteria were used to select papers for review. Coverage of:

- multiple lifecycle models
- a curation lifecycle model applicable to big data challenges
- a research lifecycle model applicable to the data intensive research environment
- data curation in relation to FAT and issues of bias

For the first three criteria, 33 publications were identified representing more than fifty relevant lifecycle models. Five reviewed multiple curation lifecycle models; thirteen presented models related to big data; and fifteen reviewed or presented broader research lifecycle models. Analysis was based on descriptive synopses generated for each paper characterizing the models and their primary elements.

Selection of papers adhering to the fourth criteria was far more challenging. More than 700 papers were identified focusing on fairness, accountability, transparency, or bias in computational analyses. Due to the volume, the set of papers could not be systematically reviewed. We did confirm, however, that the preponderance of work focuses on algorithms and analytical dimensions with very limited attention to factors associated with the underlying data. One source, Rizvi, et al. (2017), provided a useful benchmark due to its explicit attention to data curation in the lifecycle of machine intelligence. Additionally, two continuum models are discussed as a type of model that, like lifecycle models, explicitly represents progressive stages of curation activity. The Records Continuum Model (RCM) was included based on feedback on a preliminary report on this research presented at IDCC 2019.

DCC Site Visit Activities

In the second phase of study, principal investigator Choudury conducted a site visit at the Digital Curation Centre (DCC) in October of 2018, to explore historical and internal organizational perspectives on the DCC Model. Interactions included four sessions with six key participants from the DCC, including Director Kevin Ashley and colleagues who addressed both past and current observations about the DCC Model. The first session included all participants, with subsequent sessions focused on selected topics with specific people. Some of the DCC participants had direct knowledge of the original motivations and evolution of the elements of model and had assessed its utility and applicability over time for a series of purposes, including research support and educational applications.

The sessions were conversational, guided by a set of semi-structured questions provided in advance. The questions were aligned with the framing conditions and informed by a review of existing DCC documentation, including *Disciplinary Approaches to Sharing, Curation, Reuse and Preservation: DCC SCARP Final Report to JISC* (Lyon et al., 2010). Summary outcomes of the conversations and follow-up email exchanges were documented in a site visit report that highlighted key themes and observations and included initial feedback from DCC participants on preliminary results from the research.

Expert Consultation

To expand ongoing engagement conducted with experts at JUH, Choudhury gathered external input through sessions with a range of experts at the University of Washington (UW) in September 2018. Nine informants were engaged, one-on-one or in pairs, in 7 discussion sessions covering the content and structure of the DCC Model in relation to the framing conditions. The experts included librarians representing data services, health sciences, and scholarly publishing; leadership of the eScience Institute; Information School faculty specializing in data science, data curation, and data ethics; representatives of a university group working on cloud and data solutions; and a data strategies group at a major medical research center that partners with UW.

Analysis

Landscape of Lifecycle Models

While our review concentrated on lifecycle models in relation to the framing conditions, it is important to acknowledge other contributions to the literature that have assessed the structural limitations of a lifecycle approach (Cox & Tam, 2018) and evaluated lifecycle models more generally (Weber & Kranzlmüller, 2019). Additionally, a number of established curation concepts continue to have broad relevance for best practices. Sheer curation or curation-at-source (Curry, Freitas, & O'Riáin, 2010) assumes a research lifecycle orientation, and agile curation (Baker & Duerr, 2017) has evolved to include a critical perspective aimed at "deconstructing" the lifecycle model (Young, et al., 2014). For the aims of this study, Freitas and Curry (2016) offer a comprehensive and sophisticated overview of curation in the big data environment, invoking a "value chain" rather than a lifecycle model. We discuss their contribution below, followed by three lifecycle models designed for big data with the highest degree of applicability to the framing conditions: the Data Life Cycles Laboratory Model (DLCL), the Big Data Lifecycle Model (BDLM), and the Comprehensive Scenario Agnostic Lifecycle Model (COSA-DLC). Each has different strengths that could contribute to updating the DCC Model.

In Freitas and Curry's (2016) big data value chain approach, data curation is the central segment in the chain, preceded by data acquisition and analysis and followed by storage and usage. The primary problems associated with curation—quality, scalability, and heterogeneity—are not unique to big data but quickly become untenable with large volumes or aggregations of data. The authors emphasize the particular challenge of complexity, introduced when big data are constructed from "a lot of small data put together" (p. 90), content is unstructured, and stores of data are decentralized. Most significantly, they highlight requirements not explicitly represented in the DCC Model, including incentives, economic models, curation at scale, human-data interaction, and trust. Two other primary requirements—standardization and interoperability—are productively advanced by drawing attention to important trends, including minimum information models and nanopublications. As discussed further below, minimum information standards respond to the need for domain specificity in data guidelines, another factor stressed by Freitas and Curry. In highlighting nanopublications, they draw attention to the challenges of provenance tracking in the linked data environment and the highly variable forms of research data objects, more generally.

The three big data models that qualified for further comparative analysis were strongly oriented to the framing conditions. All three—the Data Life Cycles Laboratory Model (DLCL), Big Data Lifecycle Model (BDLM), and the Comprehensive Scenario Agnostic Lifecycle Model (COSA-DLC)—explicitly addressed increases in the scope and complexity of data. The BDLM integrated requirements of data management, data curation, and the earlier stages of data activities within the research process (Pouchard, 2015). Additionally, it takes into account the range of infrastructural support for big data, including cloud infrastructure, institutional and disciplinary repositories, and high-performance computing. COSA-DLC drew on a comparison

and summarization of 17 data lifecycle models to explicitly address the escalating challenges associated with big data—value, volume, variety, velocity, variability, and veracity (Sinaeepourfard, et al., 2015; Sinaeepourfard, et al., 2016). It emphasizes temporal characteristics and data quality, encompassing the flow and interchangeability of real-time and historical data.

The DLCL was the most comprehensive in regard to the framing conditions and research context. The DLCL attends to conditions beyond scope and complexity, stressing management of machine actionable data, as well as processing activities and workflows (Jung, et al., 2014, 2015; van Wezel et al., 2012). As with many existing models, the graphical representation of the key elements is quite simple (see Figure 2), with the background and features of the model articulated in the associated publications. Elements of the scientific and data lifecycles are represented together in a circular process starting with project idea, followed by data acquisition, data management, data analysis, publishing, and teaching. The model also depicts the scope of data derived through experiment, simulation, and the more generic notion of "measurement". The advances of DLCL, compared to BDLM and COSA-DLC, stem from its contextualization of curation within the broader conduct of data intensive research, by adopting a research lifecycle framing and designing for the central aim of reconciling incompatibilities between generic versus domain-specific technologies and tools.

The authors of the DLCL drew on use cases from a range of data intensive research areas in energy, earth sciences, physics, and health research, all of which rely on evolving modes of instrumentation and emerging technologies for data collection, processing, and analysis. As a result, the model emphasizes seamless integration of data systems and data services, distributed data management, metadata and ontologies for data identification, as well as derivation of data over time, data security, and high-performance analysis. Since DLCL encompasses the full research lifecycle, its structure can adequately reflect significant processes and dependencies, such as the need to capture provenance information starting from the project idea phase, through publishing and teaching with data. Factors associated with scientific collaboration informed the conception as well, with recognition of the distinct roles and responsibilities of different groups and communities at different stages of the research lifecycle.

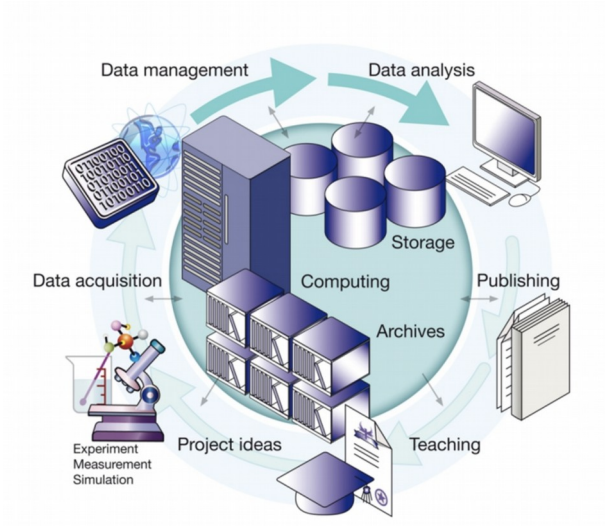

Figure 2. DLCL image blending the scientific and data life cycles (Jung et al., 2014)

Data continuum models are an interesting correlate to lifecycle models. One continuum model developed in archival science, the Records Continuum Model or RCM (Upward, 1996, 1997), predates the DCC Model and was assessed for its potential alignment with the current complexities of data reuse. While clearly not optimized for big data or high-performance computing, its concern with the temporal and mutable aspects of records is applicable to the

dynamics of data reuse and how the evidentiary, transactional, and contextual nature of data changes over time. More attuned with the contemporary data environment, the Data Curation Continuum was recently updated by Treloar and Klump (2019) and gives considerable attention to key factors of data volume and workflow processes.

Overall, significant progress has been made on lifecycle models that can directly inform adaptation of the DCC Model, as outlined in the first three framing conditions focused on the scope, complexity, and processing of machine actionable data and the workflows that generate a variety of research products. Little from that work, however, can be directly leveraged to expand the model for curation activities dedicated to promoting FAT data resources. One FAT relevant model, a machine intelligence lifecycle developed by Rizvi et al. (2017), encompasses curation aimed at decreasing discrimination risks and assuring selection of data congruent with the assumptions of algorithmic applications. The authors refer to the basic notion of bias-in-bias-out (BIBO), but ensuring "robustness of the overall life cycle" (p.68) requires documenting gaps in data, applying bias correction strategies, and determining resources for filling blind spots and deficiencies that emerge in application. But, curatorial strategies need to go further to address data provenance and documentation that capture specifications of data collection and critical assessment of underlying assumptions, especially to support reuse of data for new purposes over time.

As awareness grows on the serious threats of algorithmic bias, it is becoming clear that curation is a vital stage of the research lifecycle for combating risks to research integrity and promoting valid and transparent reuse. However, FAT data curation expertise and best practices are in their infancy. As emphasized in a recent report on responsible data science, machine learning, and AI in libraries (Padilla, 2019), managing bias and FAT will depend on workforce development, new data science services, and interprofessional collaboration.

DCC Site Visit

The Digital Curation Centre site visit provided an internal, historical perspective on the DCC Model. Multiple DCC representatives contributed invaluable context to the study, particularly on original motivations for creation and application of the model. Interestingly, while the initial development of the DCC Model was influenced by both internal and external stakeholders, the model's high level of exposure and adoption was unexpected. Work on the model was initiated in part by the DCC's need to organize and manage its own growing body of digital resources, as demand grew for DCC materials from outside audiences. The model was also intentionally designed to function as a framework for interaction, to foster conversations among data practitioners, researchers, and institutions. It was aimed at stimulating interrogation of *how*—the steps involved in metadata, preservation, rights, etc., as well as *who*—those responsible for the different activities. As a pathway to practical implementation across constituencies, the model aimed to guide structured interactions with researchers about their data practices, development of curation strategies and policies by archiving organizations, and identification of gaps in resources by funders.

The original theoretical context reflected an archival science perspective, which is consistent with the model's object-oriented approach and emphasis on tracing how digital content flows through a series of curation steps. The model was also informed by the ISO records management standard and the Open Archival Information Systems (OAIS) Reference Model. OAIS had been a topic of interest due to a DCC led review of the standard, in response to observations by the repository community on its limitations as an implementation plan. Similarly, the DCC Model was not designed as an implementation plan. Moreover, it was not initially focused on the curation of research data. The DCC's main funder was JISC, which had a broader remit encompassing library collections, leading to a model that successfully accommodated a wide range of digital content, including databases. An emphasis on research data evolved with the interest and support of other funders with data-focused programs.

DCC researchers and staff could not have forecast the varied and diverse applications of the model, and its impact continues. The canonical publication on the model (Higgins, 2008)

remains the most downloaded paper from the *International Journal of Digital Curation*, and the model has proven to be a widely used resource in LIS education.

Expert Consultation

Consultation sessions were designed to extend previous input from JHU experts. UW contributors added important views from computing and library professionals, and academic researchers and educators specializing in data science, data curation, digital preservation, and data ethics. Strong themes resonated across the JHU and UW perspectives. Both institutions have experienced profound changes in the nature of research since the late 2000s, when the DCC Model was developed. Librarians emphasized the dynamic nature of research that deviates from the discrete, clearly defined steps represented in the DCC Model. From the academic perspective, use cases are needed to demonstrate successful and unsuccessful applications of the model, especially for educational purposes. Data scientists emphasized the primacy of machine learning techniques that will increasingly defer or automate some curation activities.

From an infrastructure perspective, researchers and data curation specialists at both JHU and UW affirmed the growing presence and impact of third-party storage and computing services such as Amazon Web Services or Microsoft Azure. Researchers are attracted to the combination of storage and computation, including a suite of machine learning capabilities. These environments are often characterized by continuous processing of data and generation of metadata in a dynamic and iterative manner not accommodated by the DCC Model. Significant from a preservation perspective, tools such as Amazon Machine Images or Kubernetes offer new possibilities for addressing both data and computation in containers.

The research consultants cautioned that while such capabilities raise exciting new prospects, progress on bias in data resources and algorithms is critical. It will require qualitative as well as quantitative approaches and interventions, including interpolation and weighting techniques that assess and address limitations of data samples or specific properties. The consultants were encouraged by recent work on FAT that highlights the value of documenting assumptions and methods for data selection, processing, and choice of algorithms (e.g., Gebru et al., 2018), a key function of data curation documentation and provenance activities.

Recommendations

Our analysis confirmed the continuing value of the DCC Model. All of its elements remain relevant, but adjustments are needed in relation to the four framing conditions and trends in the diversity of data types, research questions, and computational processes across disciplines. Systematic review of progress on curation models and factors impacting curation practices set the foundation for the following recommendations for an updated model:

1. Adopt a process-oriented approach that accounts for workflows, sheer curation, and agile software development processes, to acknowledge the blurring boundaries between data and computing and integration of workflows into overarching research frameworks.

2. Organize the model in modules for case-by-case application to specific disciplines, data types, and infrastructure (e.g., cloud computing), since not all components are relevant to all situations. In practice, module use can advance context-based curation knowledge.

3. Develop a decision tree approach for navigating modules in a dynamic, non-linear, iterative manner. Each node should support identification of curation functions and activities, roles, cost factors, and associated policies.

4. Identify explicit mechanisms for highlighting and managing potential bias across phases of research—data sampling and selection, processing, choice of algorithms, assumptions from results, etc. These should be mapped to decision nodes to make implications of choices more transparent.

Below we elaborate on these recommendations, highlighting three particular facets, each of which poses challenges for evolving the DCC Model. At the same time, they can also serve as partial criteria for what a successful new model can encompass and achieve.

- overarching research lifecycle framing;
- lifecycle of curation costs;
- explicit representation of FAT curation.

As an overarching framework, a highly articulated research lifecycle would accommodate the complexity represented in Recommendation #1. For example, a structure similar to the Idealized Scientific Research Activity Lifecycle Model (I2S2) (Patel, 2011) could offer a viable base, if adapted to represent computational and scale dimensions, as well as extended lifecycle stages associated with data publication, reuse, and archiving. This kind of comprehensive framing would support additional layers of modeling prioritizing disciplinary practices, as emphasized in Recommendation #2.

The disciplinary practices emphasis is consistent with important trends in data standards. Minimum information standards and guidelines (Taylor, et al., 2008), for example, are created by communities of cross-disciplinary specialists with a focus on specific methods, to facilitate application and interpretation of data by a wider scientific community. Their rapid growth and adoption indicate the need to customize around methods and data types that align with disciplinary research processes and technologies. To do so, data services need to partner and collaborate closely with researchers (Sesartic et al., 2016; Wittenberg et al., 2017), as well as campus level research computing and other research support units, to be responsive as technology applications change within research communities.

As suggested in Recommendation #3, an updated lifecycle model should guide best practices on estimation of curation costs. A number of useful tools and protocols exist for development of business models and prediction of costs for curation services (e.g., Beagrie, 2011; Kilbride and Norris, 2014; Palaiologk, 2012). Existing practice-based curatorial approaches have also been optimized for costing applications (Chao, Cragin, & Palmer, 2015) and are well suited for integration into a lifecycle framework. However, it should be noted these tools often focus on static data. Further work is needed to support costing activities with dynamic data, structured within a broader lifecycle of curation costs. It is worth noting that this research has informed Choudhury's work on the U.S. National Academies Committee on Forecasting Costs for Preserving, Archiving, and Promoting Access to Biomedical Data, which will issue its final report in the spring of 2020.

Detailed articulation of disciplinary research processes can also support integration of FAT data assessment and enhancement into data curation services. The broad base of research on algorithmic bias provides an indirect starting point for extrapolating data curation principles and practices for FAT, but the field needs to develop a robust research agenda, with data resources as the object of study, to fully grasp bias risks and develop principles and practices for FAT curation services. Previous work focused on identifying points of intervention for curation suggest some methodological directions for such an agenda. For example, Thomer et al. (2018) show how research process modeling can be applied leveraging the PROV ontology to document the movement and transformation of data, while also capturing relationships among data, agents, and processes. While their approach was demonstrated through a non-computational case in geobiology, it is illustrative of techniques that could be tailored to the study of points of interventions for managing data bias and FAT enhancement of data resources.

Conclusion

A full implementation of the above recommendations in an update of the DCC Model will require a structure that integrates elements of both data lifecycles and research lifecycles, drawing from models designed for the data intensive environment. While this strategy would introduce a high degree of complexity compared to the rendering of the original DCC Model and most other lifecycle models, that complexity affords an opportunity to support critical functional areas, such as costing and proactive FAT curation, as a correlate of a new general lifecycle model. Most importantly, the general value of the lifecycle approach remains paramount. While this research was motivated by technological advances in data intensive settings, any successful update to the DCC Model will need to offer abstraction and clear representation of the interrelated systems of research processes and their broader context, conceiving of technologies in the service of the conduct of research and in support of the practices embedded in the cultures of knowledge production.

Acknowledgements

This research was supported by an award from the Alfred P. Sloan Foundation, Grant #G-2018-11214.

References

Baker, K. S., & Duerr, R. E. (2017). Research and the changing nature of data repositories. In Johnston, L. R. (Ed.). *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* (pp. 33-59). Chicago, IL: Association of College & Research Libraries. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

Beagrie, Charles. (2011). *User guide for keeping research data safe: Assessing costs/benefits of research data management, preservation, and re-use* (Version 2). Retrieved from https://beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science & Technology*, 66(3), 616–633. doi:10.1002/asi.23184

Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2), 142-157. doi:10.1108/AJIM-11-2017-0251

Curry E., Freitas A., & O'Riáin S. (2010). The role of community-driven data curation for enterprises. In D. Wood. (Ed.), *Linking enterprise data* (pp. 25–47). Boston, MA: Springer. doi:10.1007/978-1-4419-7665-9_2

Digital Curation Center. (n.d.). DCC Curation Lifecycle Model. Retrieved from: http://www.dcc.ac.uk/resources/curation-lifecycle-model

Freitas, A., & Curry, E. (2016). Big data curation. In Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.), *New horizons for a data-driven economy* (pp. 87-118). Cham, Switzerland: Springer. doi:10.1007/978-3-319-21569-3

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford. K. (2018). Data sheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, FAT/ML 2018*. Retrieved from: https://www.fatml.org/media/documents/datasheets_for_datasets.pdf

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

Jung, C., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Prabhune, A., Rigoll, F., Schwarz, K., & Streit, A. (2015). Progress in multi-disciplinary data life cycle management. *Journal of Physics: Conference Series, 664*(3). 032018. doi:10.1088/1742-6596/664/3/032018

Jung, C., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Rigoll, F., Schwarz, K. Stotzka, R., & Streit, A. (2014). Optimization of data life cycles. *Journal of Physics: Conference Series, 513*(3), 032047. doi:10.1088/1742-6596/513/3/032047

Kilbride, W., & Norris, S. (2014). Collaborating to clarify the cost of curation. *New Review of Information Networking*, 19(1), 44-48. doi:10.1080/13614576.2014.898543

Lyon, L., Rusbridge, C., Neilson, C., & Whyte, A. (2010). *Disciplinary approaches to sharing, curation, reuse and preservation: DCC SCARP final report to JISC.* Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

Padilla, T. (2019). *Responsible operations: Data science, machine learning, and AI in libraries.* Dublin, OH: OCLC Research. doi:10.25333/xk7z-9g97

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. *International Journal on Digital Libraries*, 12(4), 195–214. doi:10.1007/s00799-012-0092-1

Patel, M. (2011). I2S2 Idealised Scientific Research Activity Lifecycle Model. Retrieved from https://researchportal.bath.ac.uk/en/publications/i2s2-idealised-scientific-research-activity-lifecycle-model

Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation, 10*(2), 176-192. doi:10.2218/ijdc.v10i2.342

Rizvi, S. A. A., Heerden, E. Van, Salas, A., Nyikosa, F., Osborne, M. A., Roberts, S. J., & Rodriguez, E. (2017). Identifying sources of risk in the life cycle of machine intelligence applications. *AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good (Technical Report SS-17-01)*, 64–70. Retrieved from https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15259/14516

Sesartic, A., & Töwe, M. (2016). Research data services at ETH-Bibliothek. *IFLA Journal*, *42*(4), 284-291. doi:10.1177/0340035216674971

Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., & Marín-Torder, E. (2016). Towards a comprehensive data lifecycle model for big data environments. In Anjum, A., & Zhao, X. (Eds.) *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT 16* (pp.100-106). doi:10.1145/3006299.3006311

Sinaeepourfard, A., Masip-Bruin, X., Garcia, J., & Marín-Tordera, E. (2015). *A survey on data lifecycle models: Discussions toward the 6Vs challenges* (Technical Report UPC-DAC-RR-2015–18). Retrieved from https://www.ac.upc.edu/app/research-reports/html/RR/2015/18.pdf

Taylor, C. F., Field, D., Sansone, S. A., Aerts, J., Apweiler, R., Ashburner, M., ... & Brazma, A. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology*, *26*(8), 889-896. doi:10.1038/nbt.1411

Thomer, A., Wickett, K., Baker, K. S., Fouke, B., & Palmer, C. L. (2018). Documenting provenance in non-computational workflows: Development of Research Process Models through a case study of geobiology research in Yellowstone National Park. *Journal of the Association for Information Science & Technology, 69*(10), 1234-1245. doi:10.1002/asi.24039

Treloar, A., & Klump, J. (2019). Updating the Data Curation Continuum: Not just data, still focussed on curation, more domain-oriented. *International Journal of Digital Curation, 14*(1), 87-101. doi:10.2218/ijdc.v14i1.643

Upward, F. (1996). Structuring the records continuum, Part 1: Post custodial principles and properties. *Archives and Manuscripts, 24*(2), 268-285.

Upward, F. (1997). Structuring the records continuum, Part 2: Structuration theory and recordkeeping. *Archives and Manuscripts*, *25*(1), 10-35.

van Wezel, J., Streit, A., Jung, C., Stotzka, R., Halstenberg, S., Rigoll, F., Garcia, A., Heiss, A., Schwarz, K., Gasthuber, M., & Giesler, A. (2012). Data life cycle labs: A new concept to support data-intensive science. arXiv:1212.5596v1 [cs.DL]

Weber, T., & Kranzlmüller, D. (2019). Methods to evaluate lifecycle models for research data management. *Bibliothek Forschung und Praxis*, *43*(1), 75-81. doi:10.18452/19691

Wittenberg, J., Elings, M., Witt, M., & Horstmann, W. (2017). Building a research data management service at the University of California, Berkeley: A tale of collaboration. *IFLA Journal*, 43(1), 89-97. doi:10.1177/0340035216686982

Young, J. W., Lenhardt, W. C., Parsons, M. A. & Benedict, K. K. (2014). Taking another look at the data management life cycle: Deconstruction, agile, and community. American Geophysical Union (AGU) Fall Meeting, December 2014, abstract id. IN51B-3779.