DCC Curation Lifecycle Model 2.0: Literature Review and Comparative Analysis

Caihong Huang, Jian-Sin Lee, and Carole L. Palmer

Information School
University of Washington
March, 2020

Information School
UNIVERSITY of WASHINGTON

Table of Contents

1. Introduction

The literature covered in this report was gathered and reviewed in 2019 as background for the DCC Curation Lifecycle Model 2.0 project. Led by Sayeed Choudhury at Johns Hopkins University and funded by the Alfred P. Sloan Foundation, the project was designed to provide a foundation for updating the original DCC model, originally published in 2008.

A revised model is needed to accommodate the dramatic changes in the context of data production and use, and the associated impact on data archives and repositories. Four contemporary conditions, related to the "big data" environment, were identified as initial framing for the project.

Contemporary Conditions

1. growing scope, complexity, and prominence of machine-actionable data
2. importance of machine learning for data processing and analytics
3. growth of integrated research workflows
4. concerns with fairness, accountability, and transparency (hereafter: FAT) of data and algorithms

The first step in the project was to identify recent lifecycle models that account for advances made in relation to the four contemporary conditions. The papers identified met one of the following relevance criteria:

- reviewed multiple lifecycle models
- presented curation models applicable to big data challenges
- presented research lifecycle models applicable to the data-intensive environment

Thirty-three publications were selected covering more than fifty relevant lifecycle models related to the first three contemporary conditions, listed above. Five reviewed multiple lifecycle models; thirteen introduced models related to big data; and fifteen presented models that encompassed data in a broader research lifecycle.

In relation to the fourth condition on FAT issues, more than 700 papers were identified. While the large number of candidate papers could not be systematically reviewed, initial screening revealed a strong focus on algorithms with limited attention to the underlying data. One paper that explicitly addresses the role of data curation is included to represent the important implications of FAT for a revised data curation lifecycle model.

After in-depth review and assessment by the research team, 19 publications most aligned with the criteria and the aims of the study were selected for coverage in this report. We begin by introducing publications associated with the original DCC curation lifecycle model, followed by three sections that review papers that cover multiple models, big data models, and research lifecycle models. As mentioned above, one benchmark paper is included on FAT. We conclude with a comparative analysis that highlights the three models determined to be the most applicable to the first three framing conditions.

2. DCC Curation Lifecycle Model

Of the three publications discussed in this section, Higgins (2008) is the primary paper associated with the original DCC Curation Lifecycle Model (hereafter: the DCC Model). While it is the most referenced publication, two related papers contributed to the discourse around the time the original DCC Model was introduced: Pennock (2007) and Constantopoulos et al. (2009). The earliest publication, Pennock (2007), outlines six stages that stress the necessity of sequential activities throughout the lifecycle of digital information to ensure continuity of service. Referencing Pennock, Higgins (2008) followed with the canonical DCC Model that includes four layers of actions and eight sequential stages, providing a more comprehensive cycle and in-depth descriptions of each action. In 2009, the DCC Model was extended by the Constantopoulos et al. group from the Digital Curation Unit of the Athena Research Center in Greece. Their DCC&U Curation Lifecycle Model modifies the original DCC Model to include "user experience" among other sequential actions. Pennock's (2007) model, while not as elaborate as the Higgins, includes stages that encompass creation and active use of digital information. This emphasis on upstream point-of-creation activities re-emerged as a major theme in later work, as seen in Curry, Freitas, & O'Riain's (2010) notion of "sheer curation" and elsewhere, promoting a broader view of curation that more fully integrates the activities of data creators and users, in addition to the roles of archives and repositories.

2.1. Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library & Archives*, *1*, 34-45.

Digital curation is defined as maintaining and adding value to a trusted body of digital information for both current and future use, emphasizing the active management and appraisal of digital information over its entire lifecycle. Lifecycle stages include creation, active use, appraisal and selection, transfer, storage and preservation, and access and re-use. The lifecycle approach is necessary because:

1. Digital materials are fragile and susceptible to change from technological advances throughout their lifecycle.
2. Activities at each stage in the lifecycle directly influence our ability to manage and preserve digital materials in subsequent stages.
3. Reliable reuse of digital materials is only possible if materials are curated to retain their authenticity and integrity.

Benefits of the lifecycle approach include facilitating the continuity of services, supporting verification of the provenance of digital data despite the technological and organizational changes in their context, and maximizing the initial investment made in creating or gathering them.

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

The DCC Curation Lifecycle Model offers a graphical high-level overview of the lifecycle stages required for successful curation. The model is indicative rather than exhaustive, and it complements, rather than restates, a number of standards and can be used in conjunction with relevant models, such as the Reference Model for an Open Archival Information System (OAIS). It was conceived as a training tool to help curators understand the processes involved in successful curation and develop curation and preservation methodologies for their organizations. The intent was for future work to focus on the development of domain-specific variations to contextualize and tailor applications.

2.3. Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y., & Papatheodorou, C. (2009). DCC&U: An extended digital curation lifecycle model. *International Journal of Digital Curation*, *4*(1).

Constantopoulos et al. (2009) propose a fusion of the DCC Model with the Digital Curation Unit (DCU) Model developed by the Athena Research Centre, consisting of the following processes.

Digital resources lifecycle management:

a) appraisal
b) ingest
c) classification, indexing, and cataloguing
d) knowledge enhancement
e) presentation, publication, and dissemination
f) user experience
g) repository management, and preservation

Context management:

h) goal and usage models
i) domain models
j) authority management

The revised model adds user experience as a sequential action; augments description and representation information to also cover maintaining authorities—information regarding the main entities, concepts and relations; extends the curate and preserve cycle to include knowledge enhancement—new knowledge generated through research and practice encoded and organized in annotations, rules, or ontologies.

3. Reviews of Multiple Data Lifecycle Models

The eight papers in this section cover a range of data lifecycle models. Each provides a systematic analysis of the foci and differences of multiple models, except for the report by the Committee on Earth Observation Satellites (CEOS) Working Group [3.2] that offers a broader inventory with limited analysis.

3.1. Ball, A. (2012). Review of data management lifecycle models. Bath, UK: University of Bath.

Ball (2012) covers the eight lifecycle models, listed below with the associated organization or authors. The CEOS Working Group report [3.2], discussed below, includes all but the ANDS and Research360 models.

1. DCC Curation Lifecycle Model (Higgins, 2008)
2. I2S2 Idealized Scientific Research Activity Lifecycle Model (Patel, 2011)
3. DDI Combined Life Cycle Model (Structural Reform Group, 2004)
4. ANDS Data Sharing Verbs (Burton & Treloar, 2009)
5. DataONE Data Lifecycle (Michener & Jones, 2012)
6. UK Data Archive Data Lifecycle (UK Data Archive, n.d.)
7. Research360 Institutional Research Lifecycle (Jones, 2011)
8. Capability Maturity Model for Scientific Data Management (Crowston & Qin, 2011)

3.2. Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (2012). *Data life cycle models and concepts*, CEOS 1.2, Version 13.0.

This white paper was developed to fulfill the data management best practices needs of the CEOS Working Group on Information Systems and Services (WGISS) and the U.S. Geological Survey (USGS) Community for Data Integration (CDI). It is a compilation of 55 data lifecycle models, designed as a reference source to provide descriptive information rather than analysis.

Approximately 20% of the models come from U.S. government agencies, with several associated with the USGS, the National Oceanic and Atmospheric Administration (NOAA), and the Environmental Protection Agency (EPA). Social science models include the ICPSR Data Lifecycle Model, which guides the data preparation and archiving of the ICPSR repository, and the Combined Lifecycle Model from the Data Documentation Initiative (DDI), creator of the metadata standard adopted by ICPSR. The DataONE lifecycle model represents the organization's perspective as an aggregator of earth science data. Nine models were produced by academic libraries, with several others from professional organizations (e.g., JISC, ARL, and ARMA International) and two based in industry (IBM and Dell).

The inventory includes linear and non-linear models, as well as some expressed in plain text. Complexity varies, ranging from six basic stages without elaboration to highly detailed models with multiple layers, loops, and descriptions that account perspectives of different stakeholders.

3.3. Faundeen, J., Burley, T., Carlino, J., Govoni, D., Henkel, H., Holl, S., Hutchison, V., Martín, E., Montgomery, E., Ladino, C., Tessler, S., & Zolly, L. (2014). *The United States Geological Survey Science Data Lifecycle Model* (No. 2013-1265). US Geological Survey.

3.4. Faundeen, J., & Hutchison, V. (2017). The evolution, approval, and implementation of the US Geological Survey Science Data Lifecycle Model. *Journal of eScience Librarianship, 6*(2).

The CEOS white paper was followed by a 2014 report and 2017 journal publication on the Science Data Lifecycle Model, developed by the US Geological Survey (USGS), Community for Data Integration (CDI), Data Management Working Group (DMWG). The Science Data Lifecycle Model draws on the DCC Model, the Federal Geographic Data Committee (FGDC) stages of the Geospatial Data Lifecycle, and other sources reviewed in the CEOS report, resulting in a six step model: plan, acquire, process, analyze, preserve, and publish/share; and three continuous activities: describe (metadata, documentation), manage quality, and backup and secure.

The model has been applied as a framework for workflows and the following policies: *Scientific Data Management*, based on the "plan" segment; *Metadata for USGS Scientific Information Products*, based on "describe"; *Review, Approval, and Release of Information Products*, based on "preserve"; and *Preservation Requirements for Digital Scientific Data*, based on "preserve".

The following two papers are covered in greater detail, since they are more fully aligned with this report's framing contemporary conditions on changes in the growth and complexity of the data environment. The data lifecycle models introduced are based on "the Vs", a widely adopted conceptualization of the characteristics of big data. An early discussion of three Vs—volume, velocity, and variety—appeared in the META Group Research Note authored by Laney (2001). Since then, other V terms have been added, including veracity, value, variability, and visualization. The "value" dimension, discussed in Sinaeepourfard et al. (2015) below [3.6], has received somewhat less attention than the other V terms but is a critical factor in the curation of data over its full lifecycle.

3.5. Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation*, *10*(2), 176-192. [Included in comparative analysis in Section VII.]

Pouchard (2015) introduces the following four characteristics of big data:

1. Volume: size of the data
2. Variety/complexity: vast amount of unstructured data, variety of formats and sources
3. Velocity: the speed at which data accumulates, including rate of change
4. Veracity: quality of big data, understood in terms of accuracy (reliable methods of data acquisition), completeness (missing data or duplicates), consistency (accuracy of measurements and unit conversions), and uncertainty (sources and model approximations).

As background, the eight models reviewed by Ball (2012) are considered, with emphasis on how models can obscure the complexity and variety of research and the importance of early stages in the process. Carlson (2014) is noted for attention to additional limitations of lifecycle models, including their inability to adequately represent the diversity of research approaches, over emphasize linearity in data-related activities, and lack of adaptability to new organizations. The DCC Model and the DataONE model are contrasted with a cross-disciplinary approach taken by van Wezel et al. (2012) in the Data Life Cycles Laboratory Model (DLCL) model [4.1], which attends to the volume and complexity of big data contexts. DLCL is derived from five community-based initiatives, each with domain-specific data analysis tools and cross-cutting data management services, optimized for the community served.

Pouchard proposes an eight-step Big Data Lifecycle Model (hereafter: BDLM).

Steps 1 and 2 are continual, taking place as part of each of the subsequent steps:

1. **Describe** data and processes, capturing the provenance trace.
   Amount and granularity of metadata generated by big data explorations require more than metadata standards: scalable tools for the automatic generation and extraction of metadata are needed to capture the relationships between raw data and to preserve analysis (e.g., semantic tools that derive metadata from annotations and ontologies, automated scientific workflow tools, open source tools, etc.).
2. **Assure** data quality and veracity of big data.
   Acquiring data from various sources may lead to outdated or conflicting data. Different models of data representation may lead to metadata errors. Assuring quality may also occur once a dataset is analyzed.

Steps 3-8 are independent:

3. **Plan** – includes selection of data due to the potential volume of data to be preserved.
4. **Acquire** – how data is produced, generated, and ingested in research processes.
5. **Prepare** – preparing and staging data for analysis; data wrangling, through iterative data exploration and transformation to enable analysis, including integration of various sources and building pipelines of data processing.
6. **Analyze** – recording and preserving parameters of experiments, simulation scripts, and the entire computational environment for reproducibility of results.
7. **Preserve** – creation of pipelines or workflows that track dependencies between data and processes, allowing linking raw data to the results in a publication.
8. **Discover** – ensure datasets relevant to particular analysis or collection can be found by others, including decisions on what data is made discoverable.

The image illustrating the model includes central cogs to infrastructure: cloud infrastructure, institutional repository (IR), disciplinary repository (DR), and high performance computing (HPC) center data facility.

- Institutional repositories have challenges with volume and variety of data or high bandwidth.
- High performance computing centers offer high capacity storage and fast access beyond the life of a project.
- Strategic partnerships between libraries and HPC centers are emerging, where HPC centers provide infrastructure and IRs provide data curation expertise.

Stages 3-8 are applied to 4 Vs (i.e., volume, variety, velocity, and veracity) to illustrate the use of BDLM. Pouchard includes a table that lists questions arising when the 4 Vs are mapped to each stage of the 8 stages of BDLM.

Below we extract the questions associated with Stage 3 (Plan) and Stage 6 (Analyze) to demonstrate complications with volume, variety, velocity, and veracity in HPC in relation to factors such as sensitive data, deaccessioning of data, and trust.

	volume	**variety**	**velocity**	**veracity**
Plan	What is an estimate of data volume and growth rate?	How do data policies from different sources combine? What provisions are made to accommodate sensitive data?	Are bandwidth and planned storage sufficient to accommodate input rates?	What are the data sources? What allows a researcher to trust them? Who will own derived data, and data resulting from aggregation?
Analyze	Are adequate computing power and analysis methods available?	Are the various analytical methods compatible with the different datasets?	Do some data need to be discarded due to accumulation?	What level of code sharing is needed to ensure transparency and reproducibility? Is the chosen type of analysis appropriate for the collected data?

3.6. **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).***

Applying six Vs to assess the comprehensiveness of 17 lifecycle models, the suvey focuses on the characteristics of big data in relation to the 5 Vs—volume, variety, velocity, variability, and veracity, and then proceeds to add "value" to create a new 6Vs model. Importantly, the background discussion takes into account open data and open government data, in addition to big data. Open data provides a supportive public space for data stakeholders to share information, with economic benefits for businesses and free resources and new services for the public. Weaknesses of open data include lack of data quality, incompatible formats and access methods, and semantic variability. Open government data, in particular, promotes transparency, participation, and collaboration, and supports access to valid data sources and standardization of data formats. Risks include potential for violating legislation, difficulties with data ownership, misinterpretation and misuse of raw data, and negative consequences of transparency.

In articulating the 6Vs covered in the model, "value" is presented as the highest priority, with recognition throughout of the challenges related to new techniques and tools.

1. *Value* – asserting the goal of data analysis and management as enriched information. Challenge lies in developing smart approaches to discovering and recognizing hidden value of information within data resources.
2. *Volume* – handling massive data.
3. *Variety* – handling heterogeneity of data from different sources and in structured, semi-structured, and unstructured formats.
4. *Velocity* – speed of data streams, demands on data collection, processing, and analysis, with particular issues for business value.
5. *Variability* – ever-changing data meanings that are updated over time, with need to define specific algorithms and approaches (e.g., sentiment analysis, opinion mining, etc.) to make text deeply and globally understandable.
6. *Veracity* – involving quality assurance and control, as well as security concepts to guarantee that access is secure, authorized, and protected against changes and attacks during the whole data lifecycle.

3.7. **Gupta, S., & Müller-Birn, C. (2018). A study of e-Research and its relation with research data life cycle: A literature perspective. *Benchmarking: An International Journal, 25*(6), 1656-1680.**

This study reviews data lifecycle models to inform the e-Research environment, characterized by an increase in collaborative research and the need to support the quality of research and growth of scientific knowledge. Seven models frequently mentioned in the e-Research literature were compared, including ICPSR, DDI, USGS, DCC, and DataONE, to develop nine lifecycle stages:

1. *Concept and design* – conceptualization and development of scientific activities or research projects
2. *Data collection* – concerned with the type of data and where they come from
3. *Data processing* – aimed at meeting the quality assurance metrics for the ease of use or reuse by researchers
4. *Sharing and distribution of data* – ensures that collaborative research can be conducted efficiently through effective mechanisms for data sharing and distribution
5. *Data discovery* – process of extracting actionable patterns from data

6. *Data analysis* – crucial stage of determining meaning from raw data
7. *Data repurposing* – creation of new data based on processing or analysis of existing data
8. *Data archiving* – work of repositories based on parameters set by organizational policies or guidelines
9. *Data publishing* – sharing of research output in the form of publications, conference seminars, patents, other tangible media of communication, etc.

The model emphasizes the centrality of the data repurposing and archiving stages and how they interact with the other seven stages. Data archiving is recognized as an essential platform for knowledge transfer and effective management of huge volumes of data to enable collaborative research.

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

Covering 90 lifecycle models, this recent review is unique in its focus on approaches for systematically assessing lifecycle models. Two approaches are introduced: 1) comparison to determine common quality indicators, and 2) classification based purpose or use. In addition, a metamodel is introduced consisting of five lifecycle model characteristics, with the 90 models each exhibiting at least one characteristic:

1. set of data **states** during scientific processing (e.g., creation, analysis, etc.)
2. **connection** between these states (e.g., edges in a directed graph)
3. set of **roles** in the context of research data management (researchers, data stewards, etc.)
4. set of **actions** regarding research data management (e.g., collecting, documenting, etc.)
5. **mapping** of roles, actions, and states to each other (e.g. "in the state of creation, researchers describe their methods")

The first approach to assessment revealed a high degree of heterogeneity among the lifecycle models and the terminology applied to the practices of research data management. In the second approach to assessment, five classes were identified related to the purposes of lifecycle models: documentation, explanation, design, assessment, and instruction. The second approach was determined to be more promising and readily applicable to measuring lifecycle model quality.

4. Big Data Lifecycle Models

This section outlines three papers that make important contributions to understanding data lifecycle models for data curation in the big data environment. Two lifecycle models are presented—the Data Life Cycle Labs (DLCL) model and the COmprehensive, Scenario Agonistic, Data LifeCycle (COSA-DLC) model. Both of these are considered further in the comparison presented in Section 7. In this section, we also include an overview of a sophisticated treatment of curation in the big data environment that proposes a "value chain" approach, rather than a lifecycle model.

4.1. **van Wezel, J., Streit, A., Jung, C., Stotzka, R., Halstenberg, S., Rigoll, F., ... & Giesler, A. (2012). Data life cycle labs, a new concept to support data-intensive science. *arXiv preprint: 1212.5596*.** [See also comparative analysis in Section 7.]

The Data Life Cycle Laboratories (DLCL) model, discussed above in relation to Pouchard (2015) [3.5], was developed by the Large Scale Data Management and Analysis (LSDMA) project of the Helmholtz Association of German Research Centres. LSDMA activities are concerned with the incompatibilities between generic technologies and tools and those designed for use by specific communities, as well as the different levels of awareness and adoption of data management practices across different domains. The model was created through the collaborative efforts among data experts and domain scientists concerned with data and analysis in five areas:

1. Atmospheric research and climate modeling
2. Energy research
3. Medicine
4. Imaging in biology and material research
5. Large-scale physics

The model identifies six steps: (1) project ideas, (2) data acquisition (conducted by teams of scientists), (3) data management (conducted by the Data Services and Integration Team), (4) data analysis, (5) publishing, and (6) teaching. "The general goals of the DLCLs are supporting scientists in: organizing their data and metadata, establishing easy access and use of local, national and international infrastructures for data storage, data processing and data archiving, and standardizing data management techniques in the scientific communities and their data life cycles" (Heßling, 2013). In addition to Heßling (2013), see the following related articles on the DLCL initiative:

- 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).
- Jung, C., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Prabhune, A., Schwarz, K., Stotzka, R., & Streit, A. (2015). Progress in multi-disciplinary data life cycle management. *Journal of Physics: Conference Series, 664*(3).
- Fischer, M., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Prabhune, A., Rigoll, F., Schwarz, K., & Streit, A. (2017). Advancing data management and analysis in different scientific disciplines. *Journal of Physics: Conference Series, 898*(8).

4.2. **Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., & Marín-Torder, E. (2016). Towards a comprehensive data lifecycle model for big data environments. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, 100-106.** [See also comparative analysis in Section 7.]

Following on Sinaeepourfard et al. (2015) [3.6], discussed above in Section 3, this paper proposes a COmprehensive, Scenario Agonistic, Data LifeCycle (COSA-DLC) model, drawing on two Barcelona use cases: smart city and the libraries of Universitat Politècnica de Catalunya (UPC), Barcelona Tech. COSA-DLC advances a model that can be adjusted to new scenarios or environments, covers all stages of the data life cycle, and addresses the challenges associated with the 6Vs framework.

COSA-DLC is organized in three components:

1. Data acquisition: collection, data filtering, data quality, and data description
2. Data processing: process, data quality, and data analysis
3. Data preservation: classification, quality, archiving, and data

In addition to Sinaeepourfard, A., et al. (2015) [3.6], see also:

- Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., Marin-Tordera, E., Yin, X., & Wang, C. (2016). A data lifecycle model for smart cities. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 400-405.

4.3. Freitas, A., & Curry, E. (2016). Big data curation. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy*, pp. 87-118. Springer, Cham.

While not presented as a lifecycle model per se, this chapter offers the most comprehensive treatment of data curation in the big data environment, framed by a discussion of the "data value chain". Based on an extensive literature review, survey, interviews with data curation experts, questionnaires, and case studies, the authors identify challenges, requirements, and approaches to data curation in the big data era. The curation segment of the value chain is preceded by data acquisition and analysis, followed by data storage and usage. Curation is broken down into activities, including content creation, selection, classification, transformation, validation, and preservation.

Primary problems associated with big data curation include quality, scalability, and heterogeneity. Curation models are recognized as an important requirement, as are standardization and interoperability and a range of other technical and social factors: incentives, economic models, curation at scale, human-data interaction, trust, social engagement mechanisms, and unstructured-structured integration. Several exemplar case studies illustrate big data curation processes in action in domains such as health and life sciences, media and entertainment, and retail and workflows to cope with the growth and decentralized data generation in the big data environment.

5. Research Lifecycle Models

Four resources are covered in this section to provide a brief introduction to research lifecycle models--three single models and one publication with an overview of seven models. Numerous other research lifecycle models have been as visual guides for representing phases of the research process. Those developed by university libraries often map research activities to data management expectations and the data services they offer.

5.1. JISC (2014). How JISC is helping researchers. Retrieved from

Similar to the DCC Model, the JISC research lifecycle has been widely used in practice. It is included here as an example of a general purpose model and a point of comparison with the more detailed examples that follow.

Figure 1. The JISC Research Lifecycle. Reproduced from “How JISC is helping researchers”, by JISC, 2014.

5.2. Lyon, L., Heery, R., Duke, M., Coles, S. J., Frey, J. G., Hursthouse, M. B., Carr, L., & Gutteridge, C. J. (2004). eBank UK: Linking research data, scholarly communication and learning. In *eScience All Hands Meeting*. Engineering and Physical Sciences Research Council.

Lyon et al. (2004) introduce a model integrating research processes into a scholarly knowledge cycle. The representation includes roles of digital repositories and aggregator services that link datasets, e-prints, and peer-reviewed articles as resources in learning management systems. Some important elements of data intensive environments are represented, including distributed information architecture, ontologies, data models, metadata schema, open linking technologies, provenance, and workflows.

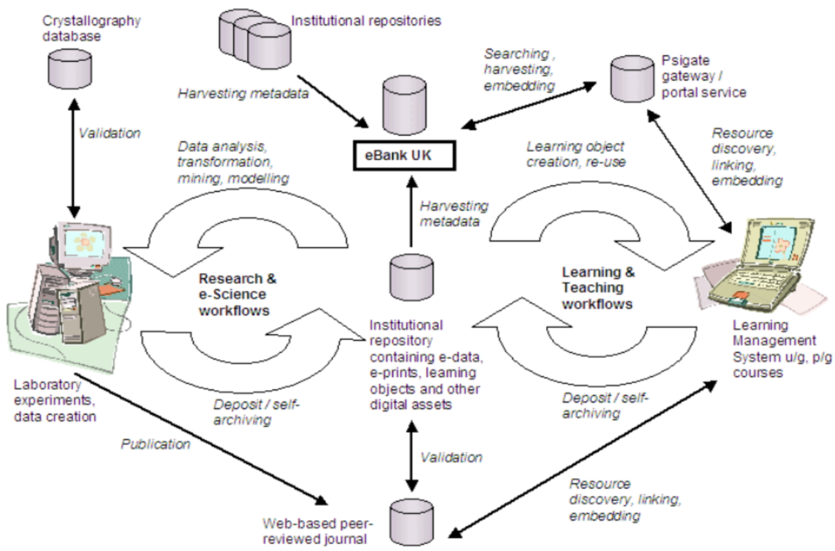

Figure 2. The Scholarly Knowledge Cycle. Reproduced from “eBank UK: Linking research data, scholarly communication and learning”, by L. Lyon et al., 2004.

5.3. Patel, M. (2011). I2S2 Idealised Scientific Research Activity Lifecycle Model.

Based on the case of experimental physical science, the Idealised Scientific Research Activity Lifecycle (I2S2) model incorporates important elements of research data and technology lacking in many other models. The experimental process, represented in the blue boxes, includes the transformation of raw data to derived data, as well as the generation of software. Stages related to the long-term management and availability of scientific data are represented in the green boxes. Considered "idealized" stages, these make explicit key activities related to intellectual property rights, embargo, access control. Importantly, archiving activities are associated with standards (OAIS), and documentation of provenance, context, and details such as calibration is also articulated.

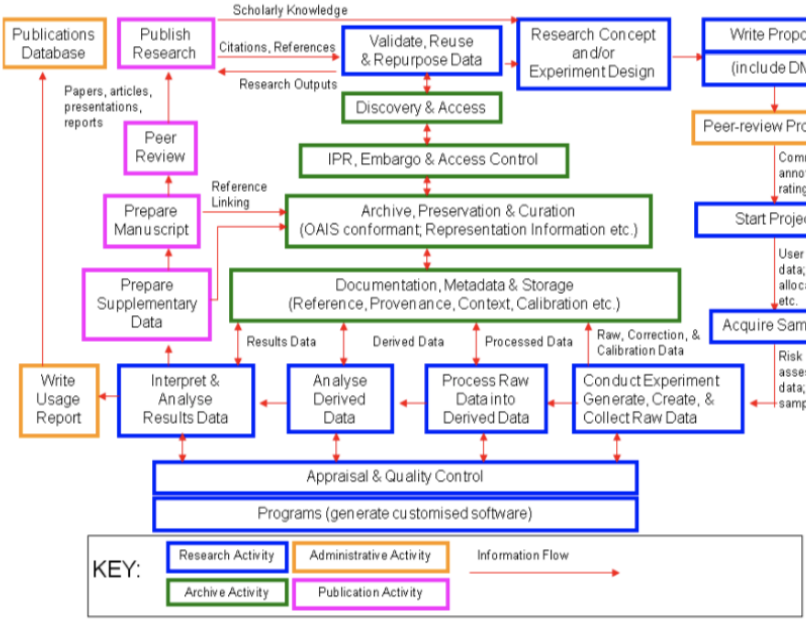

Figure 3. An Idealised Scientific Research Activity Lifecycle Model. Reproduced from "I2S2 Idealised Scientific Research Activity Lifecycle Model", by M. Patel, 2011.

5.4. 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.

This analysis examines the following seven research lifecycle models, (including 5.2 and 5.3), and two data lifecycle models.

1. Research Lifecycle (Research Information Network (RIN)/National Endowment for Science Technology and the Arts (NESTA), 2010)
2. Research Lifecycle at UCF (University of Central Florida Libraries' Research Lifecycle Committee, 2012)
3. Scholarly Knowledge Cycle (Lyon et al., 2004)
4. Research Lifecycle enhanced by an "Open Science by Default" Workflow (Grigorov et al., 2014)
5. Idealized Scientific Activity Lifecycle Model (Patel, 2011)
6. Integrated Scientific Lifecycle of Embedded Networked Sensor Research (Pepe et al., 2010)
7. E-science and the Lifecycle of Research (Humphrey, 2006)

Data lifecycles:

1. Create and manage data (Corti et al., 2014)
2. DCC Curation Lifecycle Model (Higgins, 2008)

The analysis considers three dimensions and associated elements:

1. Scope and point of view (e.g., the subject matter of the lifecycle, whether it is project-, organization-, or community-based, etc.)
2. Elements and processes (e.g., level of abstraction, homogeneous/heterogeneous, closed/open, centralized/distributed infrastructure, etc.)
3. Visualization (e.g., types of lifecycles, use of color, visuality/textuality, etc.)

To conclude this short section, it is important to note that many of the data curation lifecycle models presented in Sections 2-4 could benefit from better representation and integration of the research lifecycle. However work is still needed on research process models that adequately capture the conduct of data driven research and other aspects of the contemporary research environment, including the influence of disciplines and methods, and the iterative, nonlinear, heterogeneous, open, distributed, and messy nature of research.

6. Fairness, Accountability, and Transparency

As noted in the Introduction, the literature related to the fourth contemporary condition outlined in the introduction—fairness, accountability, and transparency (FAT) of data and algorithms—is extensive, and the preponderance of work focuses on the algorithms and analytical processes rather than the underlying data. In this section, we discuss one publication, Rizvi et al. (2017), that explicitly represents the role of data curation in relation to FAT and the analytical processes performed by algorithms. In their analysis of discrimination risks, the authors include data curation as a fundamental stage in the lifecycle of machine intelligence, articulating its importance in research integrity and in encouraging transparent reuse.

6.1. Rizvi, S., Heerden, E., Salas, A., Nyikosa, F., Roberts, S. J., Osborne, M.A., & Rodriguez, E., (2017). Identifying sources of discrimination risk in the life cycle of machine intelligence applications under new European Union regulations. In *AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01.*

A life cycle approach to Machine Intelligence Systems (MIS) is presented in the context of the European Union General Data Protection Regulation (GDPR) on personal data, emphasizing risk assessment and quality standardization. Risk assessment is challenging, due in part to lack of guidelines, but also because of compounding effects from use of multiple MIS products, and the inapplicability of traditional methods. A two-part Lifecycle of Machine Intelligence Systems is conceptualized as an inner data cycle and an outer algorithm cycle, each with vulnerabilities to bias. Data collection is a source of bias and curation is understood to be an essential point of intervention for bias correction.

Stages	Algorithm cycle	Data cycle
1	Objective definition	Data collection - data origins are a key source of discrimination bias
2	Algorithm selection	**Data curation** - bias correction strategies should be deployed in algorithm selection to fill in gaps of data collection
3	Algorithm deployment	Data manipulation
4	Machine intelligence insights	Data insights - blind spots in data may require inbuilt checks from the MIS
5	Algorithm maintenance	Data management

Discrimination risks are universal to all data processing systems and strongly associated with the origins of data, or the bias-in-bias-out (BIBO) risk. Incorrect and poor-quality input with inherent biases produces faulty and discriminating outputs in data analysis processes. Large-scale data tend to contain inherent biases due to the missing data points on some segments of the public. Managing bias risk in data collection requires representation of the provenance of data and application of data quality standards. Technical complexity combined with scale introduces a deep opacity in terms of understanding how algorithms work—the black-box risk.

7. Data Lifecycle Model Comparison

Three of the lifecycle models discussed above were selected for comparison with the DCC Model: the Data Life Cycles Laboratory model (DLCL) [4.1], the Big Data Lifecycle model (BDLM) [3.5], and the Comprehensive Scenario Agnostic Data LifeCycle model (COSA-DLC) [4.2]. They explicitly address the increase in the scope and complexity of data and have direct applicability to the contemporary conditions framing this review. Each model is reproduced below followed by an outline of seven dimensions: 1) organization and purpose, 2) audience, 3) relationships to other lifecycle models, 4) data type, 5) lifecycle structure and stages, 6) alignment with contemporary research environment, and 7) distinct elements. We conclude with further reflection on the framing conditions.

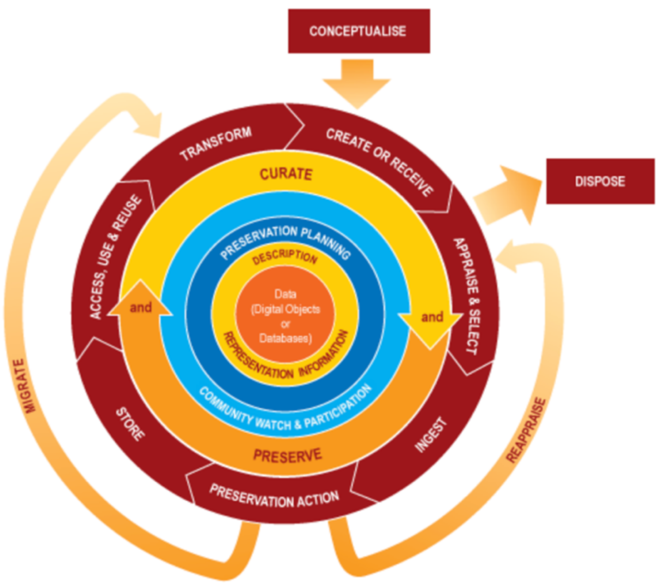

Figure 4. The DCC Curation Lifecycle Model. Reproduced from “The DCC Curation Lifecycle Model”, by S. Higgins, 2008, *International Journal of Digital Curation, 3*, 134-140.

Figure 5. The Big Data Life Cycle Model. Reproduced from "Revisiting the data lifecycle with big data curation", by L. Pouchard, 2015, *International Journal of Digital Curation, 10*, 176-192.

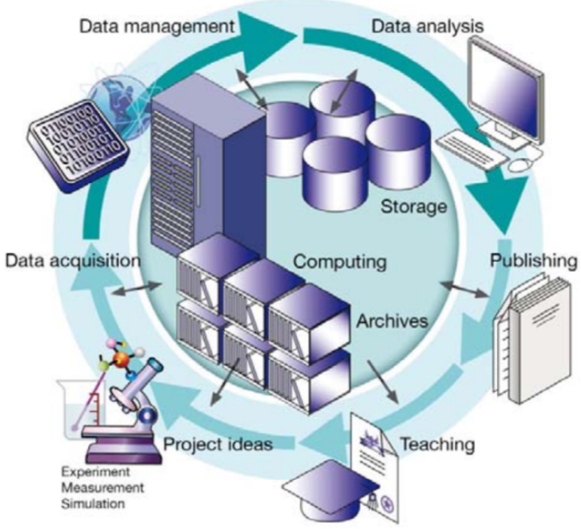

Figure 6. The Scientific Data Life Cycle (i.e., the DLCL model). Reproduced from "Data life cycle labs, a new concept to support data-intensive science", by J. van Wezel et al., 2012, *arXiv preprint:1212.5596*.

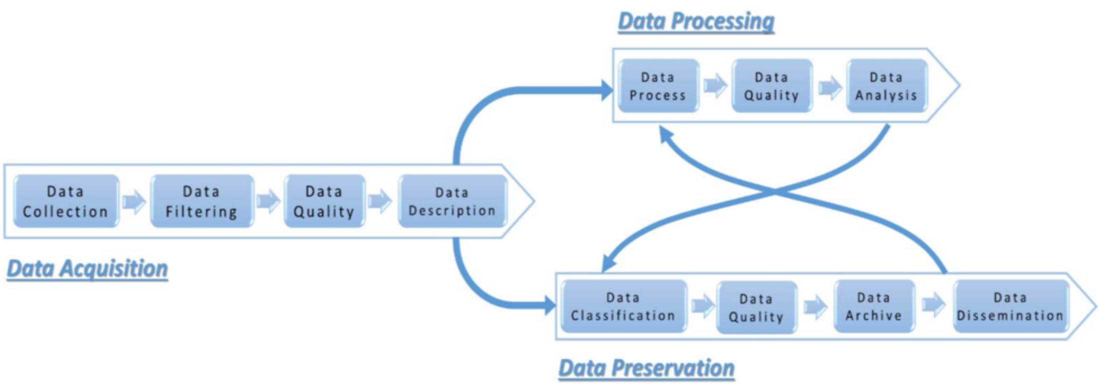

Figure 7. The COmprehensive Scenario Agonistic Data LifeCycle (COSA-DLC) Model. Reproduced from "Towards a comprehensive data lifecycle model for big data environments", by A. Sinaeepourfard et al., 2016, *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, 100-106.

(1) Organization and purpose

The DCC Model was designed by the Digital Curation Center (DCC) in 2008 to provide a high-level overview of the lifecycle stages required for successful curation (Higgins, 2008). Its development was motivated by the DCC's growing body of digital resources, but it was also intentionally designed to foster conversations among data practitioners, researchers, and institutions (Choudhury, Huang, & Palmer, 2020).

The goals of the DLCL initiative were to support scientists in organizing data and metadata; provide access and use of local, national, and international infrastructures for data storage, data processing and data archiving; and standardize data management techniques in the scientific communities and their data life cycles (Heßling, 2013).

BDLM was based on comparison of the DCC, DLCL, and DataONE models. Observing that none of the models covered data management, data curation, and the early stages of data planning and production, BDLM aimed to fill the gap with a more comprehensive model that encompassed the three previous models (Pouchard 2015).

COSA-DLC aimed to develop a model that could be readily adjusted to new scenarios, cover all lifecycle stages, and address the 6V challenges of big data: value, volume, variety, velocity, variability, and veracity (Sinaeepourfard et al., 2016).

(2) Audience

The DCC Model focuses on data curation and preservation to help curators "understand the processes involved in successful curation and develop curation and preservation methodologies for their organizations" (Higgins, 2008, p.136). In application it serves as a guide for "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" Choudhury, Huang, & Palmer, 2020).

DLCL was developed for the benefit of scientists and researchers, to support data management and data analysis tools across multiple domains. BDLM was designated to "be used by researchers, data managers, and librarians to plan the workflow of research data management and data curation activities in their projects or organizations" (Pouchard, 2015). COSA-DLC targeted practitioners from scientific communities with the aim of adaptability to a range of environments and applications, including smart cities.

(3) Relationship to other lifecycle models

The DCC Model is complementary to a number of standards for managing digital materials, and it can be used in conjunction with relevant reference models such as the OAIS model and ISO 15489 (Higgins, 2008).

Relationships to other models are not specified for DLCL. The development of BDLM was based on analysis and comparison of the DCC, DLCL, and DataONE models to serve as a general model. COSA-DLC drew on analysis of 17 data lifecycle models.

(4) Data types

The DCC Model has been specified for digital objects or databases:

- *Digital objects*
  Simple digital objects (e.g., text files, image files or sound files, along with related identifiers and metadata) and complex digital objects (e.g., combined digital objects, such as websites).
- *Databases*
  Structured collections of records or data stored in a computer system.

DLCL indicates application to raw datasets, intermediate results, and high-resolution images. BDLM is concerned with "big data" more generally. COSA-DLC emphasizes "real-time" data that proceeds to data processing and archivable data that is channeled to preservation.

(5) Lifecycle structure and stages

DCC is an object-based circular model, with activities divided into three categories:

- *Full lifecycle actions*: description and representation, preservation planning, community watch and participation, and curate and preserve
- *Sequential actions*: conceptualize, create or receive, appraise and select, ingest, preservation action, store, access, use and reuse, and transform
- *Occasional actions*: dispose, reappraise, and migrate

DLCL is also circular, encompassing a broader scientific lifecycle. The data lifecycle is a component that represents data acquisition, management, analysis, and archiving. Data management includes distributed data management, storage and access, metadata and ontologies for data identification and derivation over time, standardization of formats, data security, and high performance analysis (Jung et al., 2014).

BDLM presents two activities outside a circular cycle, Describe and Assure, which occur at every stage. The data lifecycle stages are plan, acquire, prepare, analyze, preserve, and discover. Supporting big data infrastructures are also represented at the core: cloud, institutional repository (IR) or disciplinary repository (DR), and high performance computing (HPC) data facilities.

COSA-DLC has a linear layout representing the temporal flow and interchangeability of real-time and historical data. Three primary stages begin with data acquisition, which branches into data processing and data preservation. Data quality is specified in each stage with connections specified between the between steps within the processing and preservation stages.

(6) Alignment with contemporary research environment

DLCL is the most systematic in attention to the elements expressed in the contemporary conditions framing this study, accounting for scope, complexity, machine-actionability, and some aspects of integrated workflows. The model was derived from requirements for data intensive sciences, including climate modeling and large-scale physics. It acknowledges challenges of data heterogeneity, structure data (including metadata), and semi-structured information. Scale was considered in relation to imaging and other types of data that demands large-scale storage and file sharing, and high-speed data access. Attention to standards and interoperability support the machine actionable element.

By contrast, BDLM and COSA-DLC directly addressed questions raised by the Vs (volume, variety, velocity, veracity, variability, and value) and imply support for and use of machine actionable data. With BDLM, Pouchard (2015) traces the first 4 Vs across every stage of the model, from planning to discovering, to characterize the big data lifecycle (p. 13). Estimated growth rates of data are considered in relation to selection for preservation, analysis, and storage. Data variety is also highlighted in terms of sources and formats that impact preparation, analysis, discovery, integration, distribution, and associated policies. COSA-DLC also focuses on data volume and data variety, noting researchers' use of multiple data sources and complex devices (e.g., sensors and smart devices), databases, web-generated data, and third-party applications. BDLM addresses big data workflows and process management, specifying the need for preservation activity pipelines that track dependencies between data and processes and link data to results in publications.

(7) Distinct elements

Key distinctions across the models can be seen in the representation of preservation, quality, planning and data preparation activities, and scope of the overall lifecycle. Preservation is explicit in BDLM and more prominent in COSA-DLC, but neither is as fully elaborated as the DCC Model, where preservation actions are distinguished from ingest and storage data, and reappraisal, migration, and disposal specified around the perimeter of the model. BDLM highlights the important early planning and data preparation work within the lifecycle. COSA-DLC has an emphasis on quality throughout the stages.

DLCL is the most comprehensive overall. It represents both the research process and research outcomes, including relationships to publishing and teaching, with the aim of seamless integration of data systems and data services. "The overarching goal of all data efforts must be to start managing the scientist's data right after the data [has] left the data acquisition device as this is the only way to capture gapless provenance information (Jejkal et al., 2017)." DLCL also recognizes the prevalence of collaboration in science and the distinct contributions made by domain experts and data experts, distinguishing roles and responsibilities of scientists who are situated within their research communities from those of data experts who work within a data services context.

References

Ball, A. (2012). Review of data management lifecycle models. Bath, UK: University of Bath.

Burton, A., & Treloar, A. (2009). Designing for discovery and re-use: The 'ANDS Data Sharing Verbs' approach to service decomposition. *International Journal of Digital Curation, 4*(3), 44-56.

Carlson, J. (2014). The use of life cycle models in developing and supporting data services. In J. Ray (Eds.), *Research data management: Practical strategies for information professionals*, pp. 63-86. West Lafayette, Indiana: Purdue University Press.

Choudhury, S., Huang, C., & Palmer, C. (2020, February). Updating the DCC Curation Lifecycle Model. 15th International Digital Curation Conference (IDCC), February 17-20, 2020, Digital Repository of Ireland, Dublin, Ireland.

Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (2012). *Data life cycle models and concepts*, CEOS 1.2, Version 13.0.

Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y., & Papatheodorou, C. (2009). DCC&U: An extended digital curation lifecycle model. *International Journal of Digital Curation, 4*(1).

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*. Los Angeles: Sage.

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.

Crowston, K., & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology, 48*(1).

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.

Faundeen, J., Burley, T., Carlino, J., Govoni, D., Henkel, H., Holl, S., Hutchison, V., Martín, E., Montgomery, E., Ladino, C., Tessler, S., & Zolly, L. (2014). *The United States Geological Survey Science Data Lifecycle Model* (No. 2013-1265). US Geological Survey.

Faundeen, J., & Hutchison, V. (2017). The evolution, approval, and implementation of the US Geological Survey Science Data Lifecycle Model. *Journal of eScience Librarianship, 6*(2), 4.

Fischer, M., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Prabhune, A., Rigoll, F., Schwarz, K., & Streit, A. (2017). Advancing data management and analysis in different scientific disciplines. *Journal of Physics: Conference Series, 898*(8).

Freitas, A., & Curry, E. (2016). Big data curation. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*, pp. 87-118. Springer Open, Cham. http://dx.doi.org/10.1007/978-3-319-21569-3

Grigorov, I., Carvalho, J., Davidson, J., Donnelly, M., Elbaek, M., Franck, G., Jones, S., Melero, R., Knoth, P., Kuchma, I., Orth, A., Pontika, N. Rodrigues, E., & Schmidt, B. (2014). Research Lifecycle enhanced by an "Open Science by Default" Workflow. Zenodo. http://doi.org/10.5281/ zenodo.49960

Gupta, S., & Müller-Birn, C. (2018). A study of e-Research and its relation with research data life cycle: A literature perspective. *Benchmarking: An International Journal, 25*(6), 1656-1680.

Heßling, H. (2014). The challenge of managing and analyzing big data. *International Journal of Computing*, *12*(3), 204-209.

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

Humphrey, C. (2006). *e-Science and the life cycle of research*. University of Alberta. Available at: http:// datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

Jejkal, T., Schuller, B., & Grunzke, R. (2017). LSDMA-driven advances in data analysis. In C. Jung, J. Meyer, and A. Streit (Eds.), *Helmholtz portfolio theme large-scale data management and analysis (LSDMA)*, pp. 179-197. KIT Scientific Publishing.

JISC (2014). How JISC is helping researchers. Retrieved from https://www.webarchive.org.uk/way back/archive/20140615113149/http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx

Jones, K. (2011, December). *Research360: Managing data across the institutional research lifecycle*. 7th International Digital Curation Conference (IDCC), December 5-8, 2011, Bristol, UK. Retrieved from http://www.dcc.ac.uk/webfm_send/589

Jung, C., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Prabhune, A., Schwarz, K., Stotzka, R., & Streit, A. (2015). Progress in multi-disciplinary data life cycle management. *Journal of Physics: Conference Series, 664*(3).

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).

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

Lyon, L., Heery, R., Duke, M., Coles, S. J., Frey, J. G., Hursthouse, M. B., Carr, L., & Gutteridge, C. J. (2004). eBank UK: Linking research data, scholarly communication and learning. In *eScience All Hands Meeting*. Engineering and Physical Sciences Research Council.

Michener, W., & Jones, M. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution, 27*(2), 85-93.

Patel, M. (2011). I2S2 Idealised Scientific Research Activity Lifecycle Model. Retrieved from http:// opus.bath. ac.uk/35186/1/I2S2_ResearchActivityLifecycleModel_110407.pdf

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library & Archives, 1*, 34-45.

Pepe, A., Mayernik, M., Borgman, C. L., & Van de Sompel, H. (2010). From artifacts to aggregations: Modeling scientific life cycles on the semantic web. *Journal of the American Society for Information Science and Technology, 61*(3), 567-582.

Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation*, *10*(2), 176-192.

Research Information Network (RIN)/National Endowment for Science Technology and the Arts (NESTA) (2010). *Open to All? Case studies of openness in research*. Retrieved from http://www. rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf

Rizvi, S., Heerden, E., Salas, A., Nyikosa, F., Roberts, S. J., Osborne, M.A., & Rodriguez, E., (2017). Identifying sources of discrimination risk in the life cycle of machine intelligence applications under new European Union regulations. In *AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01*.

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

Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., Marin-Tordera, E., Yin, X., & Wang, C. (2016). A data lifecycle model for smart cities. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 400-405.

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)*.

Structural Reform Group (2004). *DDI Version 3.0 Conceptual Model*. Data Documentation Initiative. Retrieved from https://ddialliance.org/sites/default/files/Concept-Model-WD.pdf

UK Data Archive (n.d.). *Research data lifecycle*. Retrieved from https://www.ukdataservice.ac.uk/ manage-data/lifecycle.aspx

University of Central Florida Libraries Research Lifecycle Committee (2012). The research lifecycle at UCF [Online Graphic]. Retrieved from https://library.ucf.edu/about/departments/scholarlycommu nication/overview-research-lifecycle/

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 preprint:1212.5596*.

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