IJDC | *Peer-Reviewed Paper*

Agile Data Curation as a Diversity of Practices Grounded in Shared Values and Principles

Karl Benedict
University of New Mexico

W. Christopher Lenhardt
Renaissance Computing Institute

Joshua Young
University Corporation for Atmospheric Research

Abstract

Current research data management and curation practices can be described as falling along a continuum between highly engineered systems and ad-hoc practices or nothing at all. In recognition of the increasing investment in and importance of research data as an asset for doing research, for evaluating current research results, and as a resource for new science, funding agencies, publishers and some research teams have instituted research data management practices. These practices are often aligned with a data life cycle models, of which there are many, that embody a circular process of activities that include creation, assessment, documentation, use, preservation, discovery and reuse. While these data lifecycle approaches are well aligned with the documentation and preservation of research data - particularly as they have been primarily developed by organizations with a mandate to provide for the preservation of data - this linear (or more appropriately cyclical) model does not necessarily focus on the level of effort required throughout the processes embodied in the lifecycle or the lowering of barriers to subsequent reuse. The agile data curation conceptual model outlined is proposed as a starting point for community consideration of a core set of values, principles and in the long-run recommended practices in the form of research data management and [agile] curation design patterns that may be used to define project-specific activities that are likely to both meet the immediate needs of data producing research projects while also maximizing the net value of data produced by those projects for future research, education, and applications.

Draft from 30th June 2017

Correspondence should be addressed to Karl Benedict, MSC05 3020, 1 University of New Mexico, Albuquerque, NM 87131. Email: kbene@unm.edu

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/

 doi:10.2218/ijdc.v0i0.0

Overview

The challenges that must be addressed by current research data management and curation processes and strategies consist of a combination of established practices that are not compatible with increasing complexity in the data management landscape at the project level; increasing expectations by sponsors, publishers, and institutions relating to data management and curation; and rapid growth in the volume, variety and velocity (three dimensions commonly used to define "big data") of data generated by and used in research. In combination these challenges translate into an increasing need to develop effective and efficient data management and curation strategies that align with a set of shared values and principles that inform management and curation objectives, and implement processes that are well documented and portable across data management projects.

The concept of *agile data curation* outlined in this paper represents an effort by the authors to develop a conceptual model for data management and curation that extends beyond the linear or cyclical model represented by the many data lifecycle models that have been developed (Ball, 2012; Möller, 2013; Park, 2016; Working Group on Information Systems and Services Data, Data Stewardship Interest Group, 2011). These lifecycle models have been created to define processes that are more structured than the commonly used ad-hoc or minimally designed research processes that focus on the immediate needs of the current research activity without consideration of the full arc of activities that are required to meet the needs of both the current research activity *and* those of future users of the data products generated by that activity (Akers & Doty, 2013; Kennan & Markauskaite, 2015; Kervin, Michener, & Cook, 2013; Tenopir et al., 2011; Vines et al., 2014; White, 2010).

In response to this problem of under-design and with the increasing recognition of the value of research data products for assessment, replication, validation, and extension of research, a variety of requirements have been put in place by sponsors (Obama, 2009, 2012, 2013; Office of Management and Budget (OMB), 2009, 2012, 2013) and publishers ("Availability of data & materials," 2016; Public Library of Science (PLOS), 2016) for planning for and executing effective data management, sharing, and curation. While these requirements have resulted in more explicit documentation of *plans* for data curation and management, the impact of these requirements on *practice* has been found to be limited, particularly in reference to data sharing (Mauthner & Parry, 2013).

When considering the practice of research data management and curation, the priorities (implicit or explicit) of the diverse participants in the process must be identified and addressed. The observations made by Greene and Meissner (2005) relating to the backlog challenges faced by archivists - "if we are going to effectively serve our users, we must adopt a much more flexible conception of what it means to 'process' a collection," (Greene & Meissner, 2005: pp. 233) can easily be applied to the challenges faced by researchers and data curators. In an environment of limited funding for research, the technical, semantic, and social challenges identified by Michener (1998) in the context of research data related to forest ecosystem resources remain relevant today, as do the seven 'habits' that he recommended for

effective information management:

1. Allocate a reasonable percentage of research funding for *long-term management* of data and information generated by the research. In most organizations, data management is seriously under-funded, resulting in data losses and delays in translating data to information.
2. Develop and adhere to *data and metadata standards* and best use protocols.
3. Provide funding for *data rejuvenation* (e.g., adding Global Positioning System fixes, i.e., latitude/longitude, to field sites) and rescue (e.g., convert paper records to digital format) to halt further data entropy.
4. *Routinely evaluate data utility, research objectives, and management needs, and reestablish priorities.* Use this information to revise sampling programs (e.g., reduce effort in certain areas, add new parameters) and to streamline data capture.
5. *Coordinate software and systems development* and purchases with other agencies or departments to eliminate duplication of effort and reduce expenditures (i.e., take advantage of economies of scale).
6. *Cooperate with other agencies, scientists, and the private sector* to establish and adopt data and metadata standards, authority files, and thesauruses for data.
7. Establish synthetic research as a top priority. (W. K. Michener, 1998: pp 434. Emphasis added.)

In particular, these 'habits' 1) acknowledge that effective information management requires investment, both in terms of planning and funding; 2) require ongoing evaluation of data value and potential use; and 3) reflect involvement by diverse stakeholders beyond the research teams and data curators with whom they (may or may not) be working.

While the increasing requirements for planning and execution of systematic data management and curation have resulted in additional attention to these topics, there has not been a corresponding increase in funding in support of these activities. The challenge of fitting these required management and curation activities within existing funds is compounded by the continuing (often characterized as exponential) growth (National Aeronautics And Space Administration (NASA), 2016; Turner, 2016) in created, managed and requested data within those limited resources. These increasing demands within a consistently resource constrained environment increase the value of developing data management and curation objectives and strategies that are likely to maximize the current and future value of research data within available resources.

Given this context, the authors have (with contributions from participants in workshops and meeting sessions held over the past two years in multiple venues) been considering the agile software development movement (Beck et al., 2001) as a source of inspiration for the development of a conceptual model for *agile data*

 doi:10.2218/ijdc.v0i0.0

curation that balances the needs for robust documentation and engineered solutions with a development cycle that is designed for incremental delivery of value through an iterative development and investment process. From the discussions held with researchers and data managers participating in meetings of the Federation of Earth Science Information Partners (ESIP), American Geophysical Union (AGU), the Research Data Alliance (RDA), SciDataCon, and the International Digital Curation Conference, the authors have had an opportunity to explore and refine some of the key concepts relating to agile data curation as it is both similar and dissimilar from agile software development.

Figure 1 illustrates a number of the shared and different characteristics that have been identified that may be ascribed to the ends of the continuum between highly designed/engineered processes and ad-hoc processes in both software development and data curation. A common theme that has emerged in the discussions around this topic over the past two years has been that while the agile software development movement partially emerged in response to the observed shortcomings in the commonly employed, specification heavy, and long development cycle "waterfall" development model, the proposed agile data curation model is largely a response to ad-hoc data management practices that are frequently the norm for research projects - particularly small research projects for which there are not dedicated data management and curation resources, dark data in Heidorn's (2008) terminology. While there are exemplars of highly successful software development and data curation practices at all points along the continuum illustrated in Figure 1, the implicit or explicit adoption of agile software development practices in the middle range of the continuum has allowed some projects to achieve success where they may have otherwise been unsuccessful, and likewise data curation activities that have successfully moved from the right end of the continuum towards the center have also provided measurable value to both the current projects that are creating the data and to future users of the data produced by those projects. It is these successful data curation projects that exemplify an emerging set of values, principles and practices of agile data curation that provide the foundation for the design pattern activity of the research team that is described below.

Methods (1000)

** note: Justification of use of agile principles for activities outside of software engineering**

For the development of an agile curation conceptual model and potential design patterns we use a three-part approach. First we look to develop the conceptual model in the abstract. We compare and contrast the current generally accepted models which guide data curation with the key concepts of agile development. At the same time we are engaging the data curation community to understand their views about and approaches to agile data curation concepts and models. The goal is to map the conceptual model of agile software to data curation while explicitly addressing areas of divergence between the two activities. The application of agile principles for software development to other domains is not unprecedented. Applications outside of software engineering includes manufacturing, supply chain

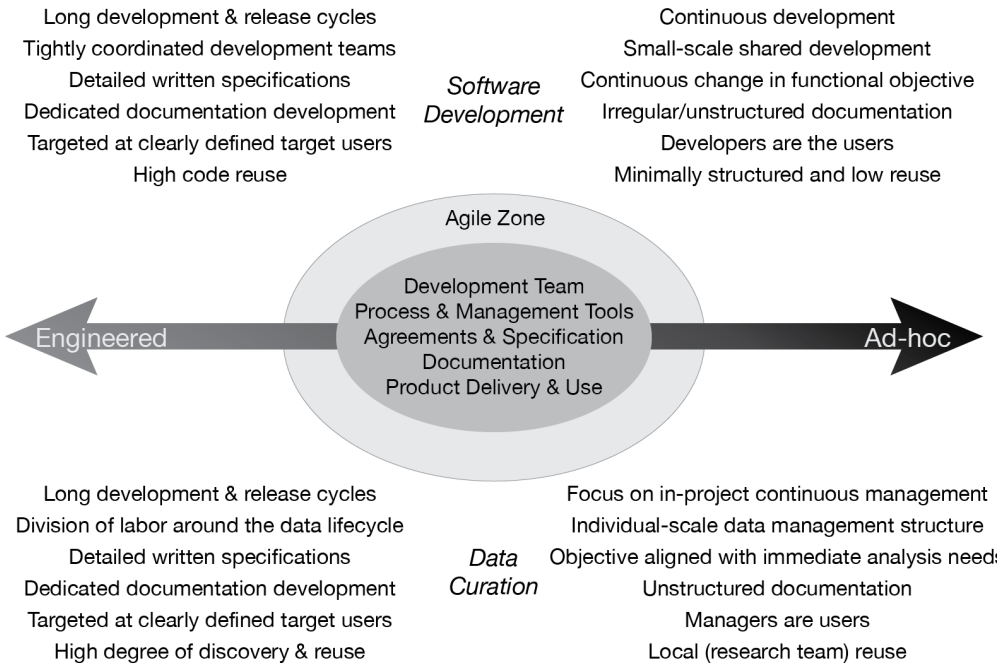

Figure 1. Illustration comparing software development and data curation activities along a continuum between *engineered* and *ad-hoc* highlighting a range of characteristics associated with with each activity, and the mid-point along the continuum where an agile approach can hopefully achieve a balance between the two extremes.

 doi:10.2218/ijdc.v0i0.0

management, and aid relief. [add citations] Second, we will use an applied approach to developing an understanding of agile data curtation through examples. This second activity gathers case study data that illustrate extant practices that reasonably fit the ethos (read 'conceptual framework') of agile curation. Cases are self-selected at this point by the submitter as an example of an agile curation process. Third, we will develop agile curation design patterns based on the agile curation conceptual model and the principles and practices distilled from the case studies. In the context of this work, data curation design patterns are intended to document common *named* data curation *problems*, *solutions*, and *consequences* that provide *descriptions of generalized data components that are customized to solve a general design problem in a particular context* (adapted from (Gamma, Helm, Johnson, & Vlissides, 1995, sec. 1.1)). Following their use in software engineering, data curation design patterns are intended to provide a consistently named and described strategy for solving a clearly defined and generalized data management and curation problem. As a complement to this concept of design patterns, the authors also intend to identify common data curation "antipatterns" in which (as defined by Long) "An"antipattern" is similar to a pattern except that it is an obvious but wrong solution to a problem. Antipatterns have been tried over and over again with much the same result: failure." (Long, 2001, pg. 68)

Discussion

1. Concept mapping (Josh - 800-1000) - have done this, capture here. Talk about vetting of this so far by community. i.e. the papers / posters / sessions we've had to date. What have we learned. What are the areas of push-back?
2. Case studies (Chris - 800-1000)

As part of our research to date we solicited examples from the data curation communities as part of sessions at the AGU Fall 2015 AGU, Summer 2016 ESIP Meeting, SciDataCon 2016, and RDA P6?, P8. (add specifics on these) The examples presented during these panels included agile approaches to managing a physical sample repository, ORNL, and ex3. These preliminary cases facilitated the development of a template to collect data about the cases systematically. Our original goal was to develop more uniform information from a set of agile example cases currently in practice. The uniform information will facilitate the analysis to identify similarities and dissimilarities across the cases. Relevant dimensions for the case study analysis include domain and type of data, data curation requirements, metadata and documentation, processes descriptions, and outcomes. The template has been converted into an online survey. This survey was made available to solicit inputs from the research and curation communities. (Assuming we have data, do we want to make it available?)

However, during our initial phases conceptualizing the research we came to realize that it would be beneficial to develop a richer understanding of the landscape of curation exemplars. That is, we don't assume a case to be agile at the start. Our working hypothesis has evolved to include the possibility that various elements of an agile approach may already be present in existing curation

examples. What may be missing is the systematic application of agile approaches from research conceptualization through to curation and dissemination. Therefore, we expanded our instrument to collect a broader set of information to characterize critical dimensions of agile curation. See figure ## - CurationCube.

[Say something about the validity of a case study approach?]

3. Proposed design pattern process (Karl - 800-1000) -

The application of the concept of agile data curation design patterns is based upon the concept initially developed for object oriented software development (Gamma et al., 1995), and extended into related domains (Ackerman & Gonzalez, 2010; e.g. Daigneau, 2011; Hohpe & Woolf, 2003; Lasater, 2010; Schwinn & Schelp, 2005). The conceptual model that the research team has developed for mapping research data curation functional requirements into design patterns represents a combination of specific research activities that have data-related components (as exemplified in Figure 1) and linkages between those components as envisioned by a model such as the *Open Archival Information System* (OAIS - (Consultative Committee for Space Data Systems (CCSDS), 2012; International Organization for Standardization (ISO), 2012; Lavoie, 2014) - Figure 2). In particular, the research team has developed a model for collecting[1] and synthesizing (through qualitative analysis methods) data curation case studies that can be used as exemplars for identifying existing design patterns or developing new ones that are relevant in data curation.

The process that the research team has developed in support of the identification and, when needed, development of agile data curation design patterns consists of the following steps:

1. Reach out to the research data curation community to seek specific research data problems for which effective (or inefective) solutions have been developed.
2. From this compilation of case studies begin the development of a catalog of *common* problems and solutions from which candidate patterns and antipatterns can be identified
3. Compile a catalog of existing design patterns against which the identified problems and solutions can be compared to identify opportunities for reuse of existing design patterns
4. Develop RFC design pattern proposals for review and comment by the research data management community (e.g. RDA, ESIP Federation, IDCC participants)
5. Integrate feedback on proposed design patterns into final versions for publication.

Conclusions

These are our conclusions: 1. Not all data creation is created equal. Traditional curation models don't take into account, long-tail versus big science continuum.

[1] https://www.surveymonkey.com/r/agile_case

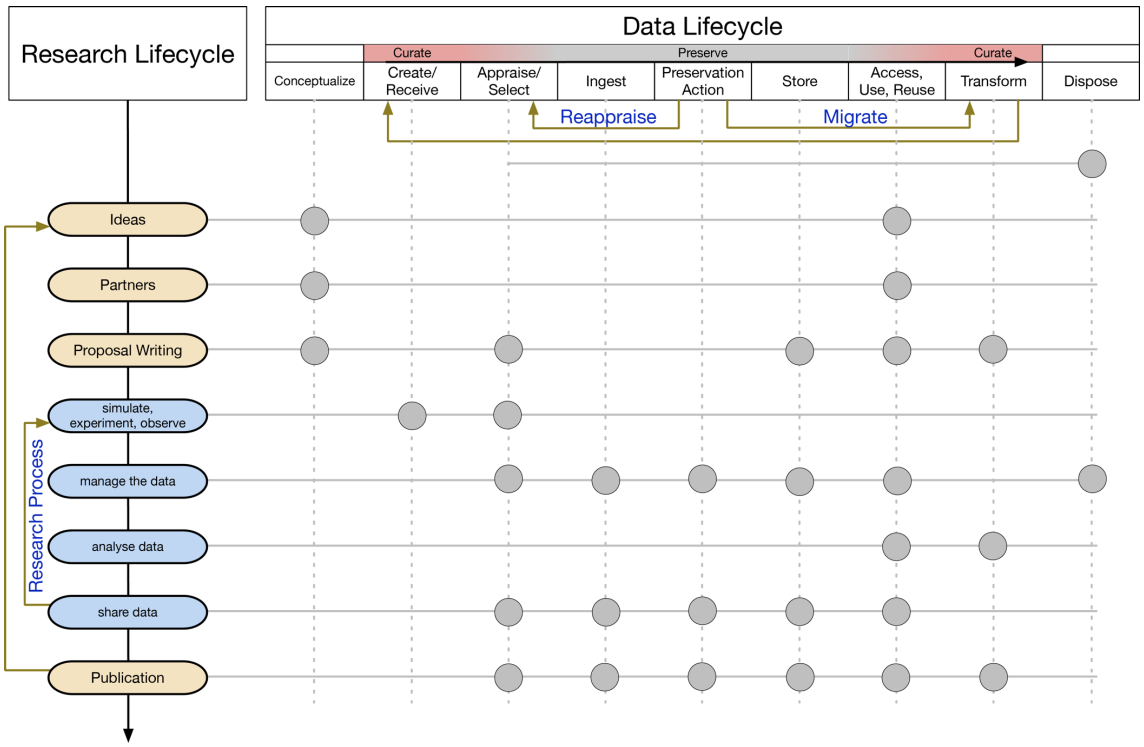

Figure 2. Intersection of Research Lifecycle (JISC, 2014) and Data Curation Lifecycle Actions (Digital Curation Centre (DCC), nd) illustrating high-level research activities that involve data-related functions.

2. There is an analogous curation continuum from none to ad hoc to light touch to highly engineered. 3. There is some correlation between the two dimensions. 4. There is an implicit[?] split/disconnect between the research lifecycle and the curation life cycle. Data often thrown over the fence to repositories. 5. Agile curation provides a means to connect the research lifecycle and the data lifecycle in a more explicit and robust way. 6. The curation / data creation disconnect also hints at the how much curation is needed to overcome data entropy. 7. ?

Next steps: 1. Initiate systematic effort to get more case study/survey data 2. Analyze case data 3. Develop draft design patterns / present to community

Acknowledgements

This would not have been possible without ...

References Cited

====== begin notes ======

The values and principles around the concept of *agile software development* developed by the agile software development community, provides a potential framework from which a set of *agile data curation and management* values and principles can be derived. Once such a set of agile data curation values and

principles have been developed, the community of research data producers and consumers is in a position to develop and use practices informed by those principles.

The objective of this paper is to propose[2] a set of *agile data curation* values and principles that parallel those developed by members of the software development community, but reflect the distinctive characteristics and challenges posed by the research data process and its products.

- Continuum from "Engineered" <==> "Agile" <==> "Ad-hoc" (Josh)
  - Technical debt as another dimension for characterizing
    - Model technical debt as increasing cost/reuse value as time passes
    - Data entropy as a dimension (increased investment in metadata, data structure, preservation can reduce the slope for the entropy curve)
  - Dimensions to think about:
    - Required Formats
    - Required data schemas
    - Required file nameing conventions schemas
    - Required metadata/documentation content
    - Required metadata standards
    - Approvals required
- Recognize cost of capture/creation, management, sharing and preservation and build prioritization into decision making about what products/parameters are maintained within the system.

Ackerman, L., & Gonzalez, C. (2010). *Patterns-Based Engineering: Successfully Delivering Solutions via Patterns*. Addison-Wesley Professional.

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2), 5–26. https://doi.org/10.2218/ijdc.v8i2.263

Availability of data & materials : Authors & referees @ npg. (2016, October). http://www.nature.com/authors/policies/availability.html.

Ball, A. (2012). *Review of data management lifecycle models.*

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... Thomas, D. (2001). Manifesto for Agile Software Development. http://agilemanifesto.org/.

Consultative Committee for Space Data Systems (CCSDS). (2012). *Reference Model for an Open Archival Information System (OAIS)* (No. CCSDS 650.0-M-2). Consultative Committee for Space Data Systems (CCSDS).

Daigneau, R. (2011). *Service Design Patterns: Fundamental Design Solutions for SOAP/WSDL and RESTful Web Services*. Addison-Wesley Professional.

Digital Curation Centre (DCC). (nd). DCC Curation Lifecycle Model | Digital Curation Centre. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software*. Reading, Mass.: Addison-Wesley.

[2] link to a web site where community input can be collected and collated into something like the *Manifesto*

Greene, M., & Meissner, D. (2005). More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist*, *68*(2), 208–263. https://doi.org/10.17723/aarc.68.2.c741823776k65863

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Hohpe, G., & Woolf, B. (2003). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Professional.

International Organization for Standardization (ISO). (2012). ISO 14721:2012 - Space data and information transfer systems – Open archival information system (OAIS) – Reference model. *ISO*. http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284.

JISC. (2014, June). How Jisc is helping researchers : Jisc. http://www.webarchive.org.uk/wayback/archive/20140615113149/http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx.

Kennan, M. A., & Markauskaite, L. (2015). Research Data Management Practices: A Snapshot in Time. *International Journal of Digital Curation*, *10*(2), 69–95. https://doi.org/10.2218/ijdc.v10i2.329

Kervin, K., Michener, W., & Cook, R. (2013). Common Errors in Ecological Data Sharing. *Journal of EScience Librarianship*. https://doi.org/10.7191/jeslib.2013.1024

Lasater, C. G. (2010). *Design Patterns*. Jones & Bartlett Learning.

Lavoie, B. (2014). *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*. Digital Preservation Coalition.

Long, J. (2001). Software Reuse Antipatterns. *SIGSOFT Softw. Eng. Notes*, *26*(4), 68–76. https://doi.org/10.1145/505482.505492

Mauthner, N. S., & Parry, O. (2013). Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology*, *27*(1), 47–67. https://doi.org/10.1080/02691728.2012.760663

Michener, W. K. (1998). Information Management Challenges to Integrated Inventory and Monitoring of Forest Ecosystem Resources. In C. AguirreBravo & C. R. Franco (Eds.), *North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources.* (Vols. RMRS-P-1, pp. 432–434). Guadalajara, Jalisco, Mexico: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. Fort Collins, CO USA.

Möller, K. (2013). Lifecycle models of data-centric systems and domains. *Semantic Web*, *4*(1), 67–88.

National Aeronautics And Space Administration (NASA). (2016, October). HEASARC Data/Usage Statistics. https://heasarc.gsfc.nasa.gov/docs/heasarc/stats/stats.html#arch_data.

Obama, B. (2009, January). *Transparency and Open Government*.

Obama, B. (2012, May). *77 FR 32391: Building a 21st Century Digital Government*. Memorandum.

Obama, B. (2013). Executive Order 13642 - Making Open and Machine Readable the New Default for Government Information. *Federal Register*, *78*(93), 28111–93.

Office of Management and Budget (OMB). (2009, December). *Memorandum for the Heads of Executive Departments and Agencies - Open Government Directive.*

M-10-06. Memorandum.

Office of Management and Budget (OMB). (2012). *Digital government : Building a 21st century platform to better serve the American people.* [Washington, D.C.] : [Washington, D.C.] :

Office of Management and Budget (OMB). (2013, May). *Memorandum for the Heads of Excutive Departments and Agencies - Open Data Policy – Managing Information as an Asset. M-13-13*. Memorandum.

Park, E. G. (2016). Session Two: OAIS Model & Digital Curation Lifecycle Model.

Public Library of Science (PLOS). (2016, October). Data Availability. http://journals.plos.org/plosone/s/data-availability.

Schwinn, A., & Schelp, J. (2005). Design patterns for data integration. *Journal of Enterprise Information Management*, *18*(4), 471–482. https://doi.org/10.1108/1741039051060961

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Turner, V. (2016, October). Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives | The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

White, H. C. (2010). Considering Personal Organization: Metadata Practices of Scientists. *Journal of Library Metadata*, *10*(2-3), 156–172. https://doi.org/10.1080/19386389.2010.50

Working Group on Information Systems and Services Data, Data Stewardship Interest Group. (2011, September). Data Lifecycle Models and Concepts Version 1.0. Committee on Earh Observation Satellites (CEOS).