Title

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

Authors

Karl Kent Benedict
W. Christopher Lenhardt
Joshua Young

Author Affiliations

University of New Mexico
Renaissance Computing Institute (RENCI)
Unidata

Corresponding Author's Email Address

kbene@unm.edu

Biographical Statement

Abstract

Current research data management and curation practices can be described as falling along a continuum between highly engineered systems and ad-hoc practices. In recognition of the increasing investment in and importance of research data as an asset for doing research, evaluating current research results, and as a resource for new research, funding agencies, publishers and some research teams have instituted research data management practices aligned with a wide variety of data life cycle models that embody a circular process of steps that include such activities as 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 herein provides is intended to propose as a starting point for community consideration a core set of values, principles and in the long-run recommended practices in the form of research data management and 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.

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 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 specific 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; Information Systems and Services Data, 2011; Möller, 2013; Park, 2016). 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 are not explicitly developed to meet the full arc of activities that meet the needs of both the current research activity *and* those of future users of the data products generated by that activity (Akers and Doty, 2013; Kennan and Markauskaite, 2015; Kervin et al., 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, 2013, 2012, 2009; Office of Management and Budget (OMB), 2013, 2012, 2009) 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, it remains unclear what impact they are having on *practice*.

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 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), and SciDataCon 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 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.

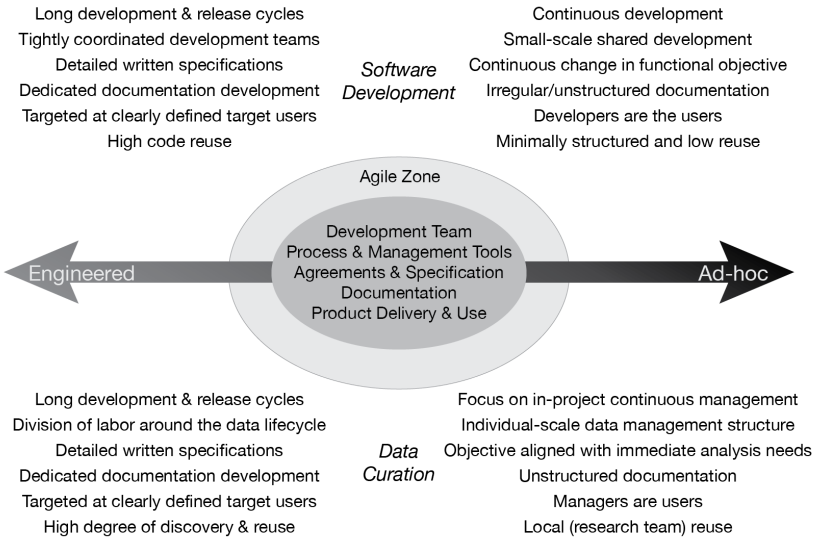

Figure 1: Illustration comparing software development and data curation activities along a continuum between *engineered* and *ad-hoc* highlighting a range of characteristics associated with with each activity, and the mid-point along the continuum where an agile approach can hopefully achieve a balance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

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

Following the model developed by the agile software development communityThe set of values and principles developed by members of the software development community around the concept of *agile software development* provides a potential framework from which a set of agile data curation and management values and principles can be derived. Once a set of values and principles have been developed the community of research data producers and consumers is in a position to develop and use practices that are informed by those principles.

The objective of this paper is to propose[1] 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"

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

    - Technical debt as another dimension for characterizing
    - 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.

Methods

These are our methods

Discussion

This is our discussion

Conclusions

These are our conclusions

Acknowledgements

This would not have been possible without ...

References Cited

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

Availability of data & materials : Authors & referees @ npg [WWW Document], 2016. [WWW Document]. URL http://www.nature.com/authors/policies/

availability.html (accessed 10.30.16).

Ball, A., 2012. Review of data management lifecycle models.

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D., 2001. Manifesto for Agile Software Development [WWW Document]. URL http://agilemanifesto.org/ (accessed 10.31.16).

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

Information Systems, W.G. on, Services Data, D.S.I.G., 2011. Data Lifecycle Models and Concepts Version 1.0.

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

Kervin, K., Michener, W., Cook, R., 2013. Common Errors in Ecological Data Sharing. Journal of eScience Librarianship. doi:10.7191/jeslib.2013.1024

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

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA), 2016. HEASARC Data/Usage Statistics [WWW Document]. URL https://heasarc.gsfc.nasa.gov/docs/heasarc/stats/stats.html#arch_data (accessed 10.31.16).

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

Obama, B., 2012. 77 FR 32391: Building a 21st Century Digital Government.

Obama, B., 2009. Transparency and Open Government.

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

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), 2009. Memorandum for the Heads of Executive Departments and Agencies - Open Government Directive. M-10-06.

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

Public Library of Science (PLOS), 2016. Data Availability [WWW Document]. URL http://journals.plos.org/plosone/s/data-availability (accessed 10.30.16).

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff,

M., Frame, M., 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6, e21101. doi:10.1371/journal.pone.0021101

Turner, V., 2016. 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 [WWW Document]. URL http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm (accessed 10.31.16).

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

White, H.C., 2010. Considering Personal Organization: Metadata Practices of Scientists. Journal of Library Metadata 10, 156–172. doi:10.1080/19386389.2010.506396