DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019

Toward A Metadata Activity Matrix: Conceptualizing and Grounding the Research Lifecycle and Metadata Connections

Sonia Pascua
Drexel University, USA
smp458@drexel.edu
sonia@slis.upd.edu.ph

Kai Li
Drexel University, USA
kl696@drexel.edu

Abstract

The role of metadata to support research cannot be underestimated; and, yet, it is difficult to develop a systematic understanding of metadata activities throughout the research process. In this paper, we preliminary analyzed how metadata activities were embedded in the research and data lifecycles. Specifically, we identified some key metadata activities associated with the components of the generic research process, from hypothesis formulation to disseminating the results and data management. The exploration raised epistemological questions about the presence of metadata activities in conducting research and managing data. This work conceptualized and grounded the connection between metadata and the lifecycles of research and data processes and presented a high-level mapping identifying the cross section of their activities and established the impression of metadata value in the field of scientific research and data management.

Keywords: metadata capital; research lifecycle; data lifecycle; metadata lifecycle

1. Introduction

Metadata plays central roles in the contemporary science, particularly research that involves large quantities of digital data. The increasing amount of data introduces new epistemic relationships between data and scientific activities (Leonelli, 2016). These epistemic relationships lead to a rising position of data in the contemporary scientific enterprise and the task of management data becoming a core scientific activity (Mayernik, Batcheller, & Borgman, 2011). Metadata serves as a platform to instigate the consistency of definitions and the reconciliation of terminologies among researchers. Given the growth of data-drivenness in science, it seems important to evaluate the values of metadata in conducting research, especially using quantitative manners.

The concept of metadata capital introduced by Willis et. al. offered means to measure the importance and quantitative value of metadata. Moreover, it inquired into metadata practices and goals that can contribute to the pool of knowledge for supporting a more interoperable environment—one that is hospitable to interdisciplinary and trans-disciplinary science. For continuum, had the said study discussed the systematic analysis of metadata schemes, our paper focused on the other end of the spectrum and examined the different research and data models with the metadata lenses. We drew the explicit connection between research, data and metadata through the direct tally of metadata activities amongst research and data activities for the perusal of measuring the metadata capital.

The concept of *metadata capital* provides a different perspective in examining the value of metadata by measuring the cost and benifit of metadata along the processes in which metadata is created and re-used (Greenberg, 2014, 2017). Metadata capital research has also been conducted to evaluate how metadata capital was increased during the laboratory process in the Viral Vector Core Laboratory at the National Institute of Environmental Health Sciences (Greenberg et al.,

https://doi.org/10.23106/dcmi.952141672

2014). In this paper, we addressed the need to further explore the relationship between metadata capital and the research processes to advance our understanding of how metadata contributes to the scientific practice.

One approach towards the goal of this study is to analyze how metadata activities and outputs are embedded in the research lifecycle models. While criticisms have been raised about the fact that lifecycle models cannot fully reveal the subtleties of actual procedures (Cox & Tam, 2018), both types of models could be used as valuable proxies of the real-world research setting to draw more contextualized knowledge about metadata (Gil, Hutchison, Frame, & Palanisamy, 2010).

As the first step of our exploration, we assessed different research and data lifecycle models and identified that one model that we subjected to the goals of this study and mapped these models to some higher-level metadata activities towards the formulation of metadata matrix. Although the models are abstractions of generic research lifecycle components, they are helpful in that they represent common processes and steps in which metadata is imbedded.

The research goals of this study are as follows: (1) identify the most proper research lifecycle models in the context of examining metadata activities; (2) obtain the high-level research and data activities of the selected models; (3) understand the scope and expanse of metadata by tracing the metadata activities in the high-level research activities and assess the metadata value for metadata capital.

2. Methodology

This study initially explored some models in the field of research and selected the most popular research and data lifecycle models that took into account some theoretically-driven considerations. The selected models were discussed and used in this study to build connections to how metadata contributes to scientific research and data. It was assessed as to how it served as the vehicle of metadata activities. A matrix of research and data lifecycle activities versus the metadata activities was formulated to map the existence of metadata in conducting research.

The concept of metadata capital was proposed to fill the metadata value gap, by measuring the cost and benefits of metadata along with the processes in which metadata is used (Greenberg, 2014; 2017).

FIG. 1 shows the conceptual framework of Metadata Capital Project that this study is part of. This framework presents the conceptualization of emphasizing metadata connections among research lifecycle model and metadata activities towards quantifying metadata events in their situ contexts. It also proposes a design and formulates a template in conducting metadata capital based on the results and findings of this study. Knowledge of the metadata process varies across different domains, although there are also clear distinctions of phases; Anchoring onto this framework, performing metadata activities and conversely not performing metadata are subjected to quantification of the cost and benefits directing to metadata capital. As part of the project research goals, empirical analysis is conducted, and includes a quality audit of the framework.

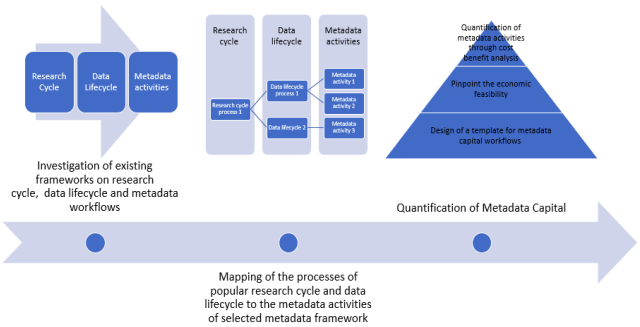

FIG. 1. Conceptual Framework of the Metadata Capital Project

https://doi.org/10.23106/dcmi.952141672

3. Results and Discussion

This section presents our preliminary results based on the selected knowledge models and corresponding discussions about the mapping among these models.

3.1. Research Lifecycle Model

How research is conducted has been a popular topic in the philosophy of science. One particular debate around this topic, i.e., whether scientific reasoning is or should be inductive or deductive, has been going on for a long time (Musgrave, 2011). It has been argued that scientific research is becoming increasingly deductive, after the arrival of data-driven science paradigm (Kitchin, 2014). However, there are also counter-arguments that the bigness of data alone cannot totally influence the traditional scientific epistemology; scientific research in the data-driven mode still needs many traditional inductive elements, such as pre-existing scientific laws and theories (Pietsch, 2017).

Despite the different reasoning styles, there are relatively subjective function unit in the research processes that lead to the final research outputs. These units are an important marker for related services (such as library and/or data services) to be aligned with the in-situ research contexts (Ball, 2012; Vaughan et al., 2013). Moreover, these units are often summarized into research lifecycle models in somewhat linear approaches.

In the National Network of Libraries of Medicine website (nnlm.gov), five models were shown as related resources discussing research lifecycles. They are the following below:

1. JISC Research Lifecycle Model
2. Oregon State University Library – Research Lifecycle
3. Research Lifecycle at the University of Central Florida Libraries
4. University of Virginia Library – Steps in the Research Lifecycle
5. University of Western Australia – Research Data Management Toolkit: Research Lifecycle Subject

Taking into account the first top three that are available for public consumption (without access and credentials needed) including the Pryor research lifecycle, Table 1 shows the comparative matrix of their activities for metadata recognition.

TABLE 1. Evaluation of some popular research lifecycle models

Research Lifecycle Model / Activities	High Level
Prior	Hypothesis
	Research
	Interpret
	Synthesize
	Publish
Oregon State University Library	Organize files
	Assign clear roles for QA/QC
	Document the context of data collection
	Document and manage file versions
	Determine appropriate file format
The UCF Libraries RLC Committee model	Planning
	Project Management
	Publishing and Presenting
	Preserving and Disseminating
	Prestige, Impact, and Discovery

https://doi.org/10.23106/dcmi.952141672

One model that meets both criteria of functionally-driven, rather than logically-driven, and how data objects are involved in the research processes as much as possible, despite taking a deductive assumption, is the one proposed by Pryor (2012, p. 6). It identified the following six parts in a cyclical approach: 1. Hypothesis; 2. Research; 3. Interpret; 4. Synthesize; 5. Publish; and 6. Reuse. Specifically, it took a data-centricity viewpoint, i.e., all research steps are situated around the existence of scientific data objects. This highlighted the importance of data management and activities in each step of scientific research. Even though the author did not aim to further elaborate this model, it was noted that there was a coherence between this research life-cycle model and data that was implicitly considered for the tracing of metadata activities.

3.2. Data Lifecycle Model

A large number of data lifecycle models were included in Ball's impactful review of existing models (Ball, 2012). See below the nine different models evaluated and considered in the manifesto of the Review of Data Management Lifecycle Models issued on February 13, 2012 by Innovative Design & Manufacturing Research Center.

1. DCC Curation Lifecycle Model
2. I2S2 Idealized Scientific Research Activity Lifecycle Model
3. DDI Combined Life Cycle Model
4. DDI Combined Life Cycle Model
5. ANDS Data Sharing Verbs
6. DataONE Data Lifecycle
7. UK Data Archive Data Lifecycle
8. Research360 Institutional Research Lifecycle
9. Capability Maturity Model for Scientific Data Management

Among all models included, the Capability Maturity Model for Scientific Data Management (CMMSDM) was concluded as the most useful model (Crowston & Qin, 2011) because it offered comprehensive details concerning the practice of managing data objects in the context of scientific practice and it takes a large functional approach to the lifecycle of data. It's also the model that this paper used for the formulation of the metadata matrix.

This model identified four stages of data process areas, including (1) Data acquisition, processing and quality assurance, (2) Data description and representation, (3) Data dissemination, and (4) Repository service and preservation. As a model rooted in actual research processes, its steps show a strong coherence with research steps discussed in the previous section: while the different natures of publications and data objects pose distinct requirements for how these objects should be treated, they nevertheless follow a largely similar path where they are deeply involved in the co-production of each other (Harris & Lyon, 2014).

3.3. Mapping metadata activities of research and data lifecycles towards metadata matrix

The above-mentioned selected models of research and data lifecycle and the identification of their metadata activities and the conceptual map among them is illustrated in Table 2. While this map does not mean to describe the temporal relationship between research and data lifecycles, it illustrates the similar procedures of both research and data lifecycle activities involving metadata activities. Both models manifested activities evident of metadata 1) concept and exposition; 2) metadata formulation and construction; 3) metadata management; 4) metadata use and manipulation; and 4) metadata value and signification. It could also be deducted that there were activities in the models evaluated that had not recognized conducting metadata activities. This could be further verified as another stride from this study.

https://doi.org/10.23106/dcmi.952141672

It offers a solid framework for our next steps of research, which is briefly addressed in the conclusion section.

TABLE 2: Towards metadata matrix: mapping of research and data lifecycle activities

Research lifecycle (Prior)	Data lifecycle (CMMSDM)	Metadata Activities
Hypothesis		
Research	Data acquisition, processing and quality assurance	✓
Interpret	Data description and representation	✓
Synthesize		
Publish	Data dissemination and preservation	✓
Reuse		

4. Conclusions

This study offers some preliminary examinations of the relationships among research, data lifecycle, and how metadata activities are involved in these processes. This is the first step of the Quantification of Metadata Capital project, which aims to formulate a framework to evaluate the costs and values of metadata works and outputs in the research process using quantitative methods.

A few conclusions that inform the next step of our project can be drawn from our discussions. First, our analysis of research and data lifecycles demonstrate that these two models together offer a coherent framework of data-driven scientific research, which can be further used to evaluate how metadata adds values to research. Moreover, the abstract nature of these models will make it easier for our framework to be used in different research contexts and draw results for comparison. Second, the limitation of the lifecycle model can also be easily observed from our results and thus should not be ignored. The linear and cyclical metaphors taken by such lifecycle models can only offer brief illustration of the research procedure (Cox & Tam, 2018), unable to cover scenarios where the actual process does not fit into a linear approach (Li, Greenberg, & Dunic, 2019).

Based on these findings, the next step of project will use both qualitative and contextualized methods, such as interview and observation, and quantitative methods to pursue the research problem discussed above. Both observational and content analysis methods would be used to understand what data activities happened in the Metadata Capital Conceptual Framework and their qualitative values. While these works are yet to be conducted, some considerations below about these tasks are discussed in this section based on the works presented in this work-in-progress paper.

First, we fully acknowledge the tensions between a constructivist view of science focusing on *in-situ* scientific procedure (Latour, 1987) that underlies observational studies and an alternative, normative view that is essential for quantifying scientific metadata activities (Greenberg et al., 2014). The framework proposed above aims to address these tensions. One specific aspect of these tensions is the fact that not all studies follow the same procedure. Sociologists of science have demonstrated that scientific procedures are highly contingent: researchers make post-hoc decisions to address situations happen in local research contexts (Cetina, 1995).

Given these considerations, we would totally acknowledge the differences between individual research projects. This, we believe, is the most important approach to bridging the two urgent needs to have a both empirically-driven understanding of scientific activities and the abilities to gain more comprehensive knowledge from projects with many minor differences.

https://doi.org/10.23106/dcmi.952141672

Second, while observational studies at the site of research laboratories would be the most direct research method to the research problem discussed in this work, it is also important to review empirical studies on scientific data management that have been conducted in many different scientific fields focusing on various aspects of data practice (e.g., Borgman, 2015; Edwards, Mayernik, Batchller, Bowker, & Borgman, 2011; Faniel & Jacobson, 2010; Jerokta, Lee, & Olsen, 2013; Zimmerman, 2007). Most of these studies, by adopting observational methods, have created rich knowledge about how data works are facilitated by metadata activities in research processes, and thus could greatly support our own collection and interpretation of first-hand data.

For the future works of this study, metadata capital will be calculated. We are hoping our work will not only contribute to a better know of how metadata works and contribute to the real world in economic terms but also help to develop new methodologies for metadata research to be used by other members in this research community.

References

Ball, A. (2012). Review of data management lifecycle models. University of Bath: Bath, http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf.

Cetina, K. K. (1995). Laboratory studies: The cultural approach to the study of science. *Handbook of Science and Technology Studies*, 140-167.

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), 1–9. https://doi.org/10.1002/meet.2011.14504801036.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: data, metadata, and collaboration. *Soc Stud Sci*, 41(5), 667–690. https://doi.org/10.1177/0306312711413314.

Faniel, I. M., & Jacobson, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work(CSCW), 355-374.

Gil, I. S., Hutchison, V., Frame, M., & Palanisamy, G. (2010). Metadata activities in biology. *Journal of Library Metadata*, 10(2–3), 99–118.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *Acm Sigmod Record*, 34(4), 34–41.

Greenberg, J. (2017). Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. *Journal of Data and Information Science*, 2(3), 19–36.

Greenberg, J. (2014). Metadata capital: {Raising} awareness, exploring a new concept. *Bulletin of the American Society for Information Science and Technology*, 40(4), 30–33. https://doi.org/10.1002/bult.2014.1720400412.

Greenberg, J. (2005, January 31). Metadata Generation: Processes, People and Tools - Greenberg - 2003 - Bulletin of the American Society for Information Science and Technology - Wiley Online Library. Retrieved August 12, 2019, from https://asistdl.pericles-prod.literatumonline.com/doi/full/10.1002/bult.269

Greenberg, J., Murillo, A., Ogletree, A., Boyles, R., Martin, N., & Romeo, C. (2014). Metadata capital: Automating metadata workflows in the niehs viral vector core laboratory. *Research Conference on Metadata and Semantics Research*, (1–13). Springer.

Harman, G. (2015). The metadata lifecycle [blog post]. Retrieved from https://bigr.io/the-metadata-lifecycle-gh/.

Harris, F., & Lyon, F. (2014). Transdisciplinary environmental research: a review of approaches to knowledge co-production. *Nexus Network Think Piece Series,* Paper 2.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 205395171452848. https://doi.org/10.1177/2053951714528481.

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting scientific collaboration: Methods, tools and concepts. Computer Supported Cooperative Work (CSCW), 22(4-6), 667-715.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Harvard University Press.

Leonelli, S. (2016). Data-centric biology: a philosophical study. Chicago: The University of Chicago Press.

Li, K., Greenberg, J., & Dunic, J. (2019). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *ArXiv Preprint ArXiv:1903.06215*.

Library of Congress (2008). Metadata standards & applications: Cataloging for the 21st century. Retrieved from: http://www.loc.gov/catworkshop/courses/metadatastandards/pdf/MSTraineeManual.pdf.

Lytras, M.D., Sicilia, M.-A. (2007). Where is the value in metadata? *International Journal of Metadata, Semantics, and Ontologies*, 2, 235–241.

Overview: Research Lifecycle. (n.d.). Retrieved August 14, 2019, from https://library.ucf.edu/about/departments/scholarly-communication/overview-research-lifecycle/
University of Central Florida Libraries

https://doi.org/10.23106/dcmi.952141672

Matthews, B. M. (2008). Metadata for information management in large-scale science. Retrieved from https://epubs.stfc.ac.uk/work/50499.

Mayernik, M. S. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators. SSRN 2042653. http://dx.doi.org/10.2139/ssrn.2042653.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference.* (417–425). New York, USA: ACM Press. https://doi.org/10.1145/1940761.1940818.

Musgrave, A. (2011). Popper and hypothetico-deductivism. In *Handbook of the History of Logic*. vol. 10, 205–234. Elsevier.

Pietsch, W. (2017). Causation, probability, and all that: Data science as a novel inductive paradigm. *Frontiers in Data Science*, 329.

Pryor, G. (2012). Managing research data. London, UK: Facet Publishing. ISBN: 978-1844047562.

Vaughan, K. T., Hayes, B. E., Lerner, R. C., McElfresh, K. R., Pavlech, L., Romito, D., … Morris, E. N. (2013). Development of the research lifecycle model for library services. Journal of the Medical Library Association : JMLA, 101(4), 310–314. doi:10.3163/1536-5050.101.4.013

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505-1520. doi:10.1002/asi.22683

Willoughby, C., Bird, C. L., Coles, S. J., & Frey, J. G. (2014). Creating context for the experiment record. User-defined metadata: investigations into metadata usage in the LabTrove ELN. *Journal of Chemical Information and Modeling*, *54*(12), 3268–3283.

Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1-2), 5-16.

https://doi.org/10.23106/dcmi.952141672