Extending Curation Profiles to Study Enterprise-level Data Practices

Nicholas M. Weber[1] and Carole L. Palmer[1]

[1] Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract

This poster presents preliminary work in adapting a 'curation profiles' approach to study data practices in a corporate enterprise setting. We outline important similarities and differences between the curation of basic vs. applied research data and present preliminary findings from a pilot study with design engineers at a multi-national corporation that manufactures heavy machinery. We show that reproducibility, quality vs. value, and that discovery-driven quality control are key areas for the development of new curation services in this sector. We conclude with some future directions for extending the curation profiles project to new data-intensive workplace settings.

Keywords: data curation, workplace studies, research and development policy

Citation: Weber, N. M., & Palmer, C. L. (2014). Extending Curation Profiles to Study Enterprise-level Data Practices. In *iConference 2014 Proceedings* (p. 1025–1027). doi:10.9776/14359

Acknowledgements: Thank you to the anonymous reviewers for their helpful suggestions.

Contact: nmweber@illinois.edu, clpalm@illinois.edu

1 Introduction

Data generated through research and development (R&D) activities are increasingly considered valuable assets with high potential for reuse, and important contributors to national economic stability (OSTP, 2013). Library and information science has studied the curation of data for reuse in many diverse scholarly settings (i.e., Treloar et al., 2007; Choudhury, 2008; Cragin et al., 2011), including recent open government data initiatives (Ding et al, 2010). However, there have been few studies that address how corporations are attempting to preserve and curate their data for future analysis (Curry, Freitas & O'Riain, 2010). Moreover, there is little known about the ways academia, government, and the private sector differ or converge on basic data curation issues, such as the organization, representation, tracking, reuse, and sharing of data amongst their various research stakeholders.

Curating and Profiling Enterprise Data (hereafter referred to as CPED) is a project investigating how both systems and services developed for data curation in academic science departments can be tailored to meet the unique demands of corporate R&D settings. This work is aimed most immediately at informing the development of new data curation infrastructures and services for corporate partners in CPED, but these findings will also provide iSchools a better understanding of the skills needed to curate data in diverse R&D contexts and can therefore drive the development of new curation curriculums.

2 Profiling Data Collections

CPED's research methodology draws from the Data Curation Profiles Project (DCPP) which was designed to study how data infrastructures and curation services might be tailored to fit the specific needs of a research field, laboratory, or discipline in an academic setting (Witt, et al., 2009). Specifically, this project attempted to:

> Enrich the understanding of data access and related curation activities through case studies of researchers' data practices.

> Translate and compare needs for archiving and sharing data across campus units and institutions. And,
>
> Convert the results into formalized policies to enhance curation and access to data collections in a repository setting. (Witt, et al., 2009)

Working with corporate partners, including a multi-national corporation that designs, engineers, and manufactures heavy machinery, CPED is attempting to extend the research agenda of DCPP to the private sector. In doing so, CPED will both investigate the research practices of different working groups, corporate laboratories, and individuals in private industries, as well as create rich, generalizable cases studies of data access and preservation issues for each of these groups.

3 CPED Pilot Study

Thus far, we've conducted a single pilot study to investigate how we might extend DCPP curation profiling approach to study enterprise-level data practices. The first phase of this work has included the customization of a curation profile template and an interview guide to be used by the IT departments of CPED partners. We have also completed 12 group interviews with data producers and consumers at one of our partner's headquarters. This pool of respondents includes engineers from manufacturing, testing, and design departments, as well as marketing and IT support staff.

The first curation profile being developed from these interviews is focused on engineers that have responsibilities in the design and drafting of a machine's geometry; including the specifications and limitations for appropriate machine use conditions, and programming lower-level software controls used by these machines. These engineers are unique in that they require access to test data that result from their design prototypes being used in simulation and field-experiments, as well as reliable data about materials (i.e. properties, prices, stock availability, etc.) from external parts vendors. The reliance on multiple sources and types of data therefore makes design engineers an ideal sample for studying data reuse and sharing practices.

4 Enterprise-level data practices

Many data management issues are persistent in collaborative research, such as the need for standardized metadata and consistent file naming conventions, but our initial interviews with design engineers have revealed three aspects of our participant's data practices that are unique to this R&D setting:

> External data are expected to be reproducible.
>
> Research conducted by design engineers is not focused on a basic understanding of a material, but instead on the performance of that material when used in a particular context. This places increased importance on reproducing and verifying material data obtained from parts vendors, as well as creating internal standards of quality that can be shared across teams using similar data. When testing vendor supplied material data it is assumed this information is not accurate and can only be accepted on a "reputation basis" after many rounds of tests that verify the manufacturers reported values.
>
> Quality vs. value is an imprecise, and often flawed distinction
>
> Some design engineers would trace back and obtain legacy data they had produced to verify the quality of a new dataset they received from field-test engineers, others only valued data coming directly from the testing of their latest prototypes- regardless of known quality issues. Value and quality were not well bounded, nor well defined concepts to our participants, and they dismissed any discussion of their own research data along such lines. This is highly unique, as most

participants in an academic research setting report quality or reliability to be the most important criteria for determining the value of a dataset for reuse (Weber et al., 2012).

Discovery is a mechanism for quality control

In many instances, datasets that were continually reused became de-facto standards for design, and were archived in many different systems and shared workspaces. These datasets persisted, not because of their high quality or uniqueness, but simply because they were easy to access, most team members knew of their existence by a generic name, and the errors that the data did contain were likely to have known "work-arounds". Reuse of data with known problems is not unique to enterprise-level data practices (Zimmerman, 2007), but a dataset being archived in multiple locations so that it would become "known" to collaborators diverges from many previous studies of data practices (e.g. Cragin et al. '2010).

5 Next Steps

We will continue to code transcripts from these interviews, and our poster will present further analysis of the themes mentioned above. For the purposes of visualization, we have also developed a set of schematics that map how data move through the design engineer's daily work routines. These maps allow us to trace the data between points of production and consumption, and will be used in the poster presentation to show useful points of intervention for data curation. We believe there is much to be learned from settings where applied R&D research is taking place, and our future work will include a more comprehensive analysis of the differences between basic and applied research data practices.

6 References

Choudhury, G. S. (2008). Case study in data curation at Johns Hopkins University. Library Trends, 57(2), 211-220.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1926), 4023-4038.

Curry, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprises. In *Linking enterprise data* (pp. 25-47). Springer US.

Ding, L., D. DiFranzo, A. Graves, J. R. Michaelis, X. Li, D. L. McGuinness, and J. Hendler. (2010). Data-gov wiki: Towards linking government data. In Proceedings of AAAI Spring Symposium on Linked Data Meets Artificial Intelligence. Menlo Park , CA.

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The data curation continuum: Managing data objects in institutional repositories. D-Lib Magazine, 13(9), 4.

Weber, N. M., Baker, K. S., Thomer, A. K., Chao, T. C., & Palmer, C. L. (2012). Value and context in data use: Domain analysis revisited. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1-10.

Witt, M., Carlson, J., Brandt, S., & Cragin, M. (2009) Constructing Data Curation Profiles. International Journal of Digital Curation, 4(3), 93-103

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.