Ixchel M. Faniel
OCLC Online Computer Library Center, Inc.

Elizabeth Yakel
University of Michigan School of Information

Significant properties as contextual metadata

Note: This is a pre-print version of a paper published in *Journal of Library Metadata*. Please cite the published version; a suggested citation appears below. Correspondence about the article may be sent to ixchel_faniel@oclc.org.

Abstract
Given the desire to maintain both long-term renderability and meaningfulness of data for reuse, increasing attention is being paid to contextual metadata, also known as significant properties. Both the data reuse and digital curation literatures identify significant properties but do not appear to build on one another, in part because of different methodological approaches. In this paper we discuss the significant properties research from both literatures and call for a more integrative approach. We conclude with a discussion of our project, Dissemination Information Packages for Information Reuse (DIPIR), which features the boundary spanning we advocate between these two research communities.

Suggested citation:
Faniel, Ixchel M., and Elizabeth Yakel. 2011. "Significant Properties as Contextual Metadata." *Journal of Library Metadata,* 11,3-4:155-165. Pre-print available online at: http://www.oclc.org/research/publications/library/2011/faniel-jlm.pdf.

Introduction

The *Reference Model for an Open Archival Information System* (OAIS) is an established standard for preservation and curation of digital objects. It covers ingest, archival storage, preservation, and access to digital collections and does this through a series of 'packages': the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Contextual metadata is present throughout the OAIS model. Preservation description information, a key component in the SIP and the AIP, consists of four types of information: reference, context, provenance, and fixity. Within the DIP, representation information ensures not only the proper rendering of data but also the communication of the contextual information, often referred to as significant properties or characteristics, required to make the data meaningful to a designated community. Most of the research on OAIS has primarily focused on SIPs (Ludäscher, Marciano, Moore 2001; Moore and Smith 2006) and AIPs (e.g., Vardigan and Whiteman 2007; Moore and Smith 2007). However, increasing attention is being paid to contextual metadata for both long-term renderability and meaningfulness during reuse.

The digital curation literature primarily focuses on preservation metadata as a means of making reuse technically possible, but does not consider other types of contextual metadata needed to facilitate meaning making by members of a designated community during reuse. Likewise, the data reuse literature focuses on meaning making but not the

technical aspects of digital preservation. This is unfortunate as OAIS states that a repository may ‘‘make a decision between maintaining the minimum Representation Information needed for its Designated Community, or maintaining a larger amount of Representation Information that may allow understanding by a larger Consumer community with a less specialized Knowledge Base’’ (Consultative Committee for Space Data Systems 2002, pp. 2–4). We assert that the current implementations of preservation regimes, such as the OAIS, preference long term renderability of data over intelligibility or meaningfulness of data to the designated communities.

In this article we are particularly interested in contextual metadata that enables reuse. We will refer to contextual metadata as significant properties and utilize the InSPECT definition

> The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record (Grace et al., 2009, p. 4).

We prefer this definition for three reasons: 1) it focuses on use, 2) references a time scale, and 3) alludes to the need for authenticity in digital objects. Both the data reuse and digital curation research identify significant properties. However, this research does not appear to be actively building on the parallel research, and there is a lack of cross-disciplinary dialog between these two research traditions. This article examines the research on significant properties in the data reuse and digital curation communities, discusses the different methodological approaches, including some of the benefits and

weakness. We will close with a discussion of our own project, Dissemination Information Packages for Information Reuse (DIPIR), which features boundary spanning needed between these two research communities and bridges some of the gaps in the research.

Literature Review

Research into significant properties has been going on for over a decade in order to identify those characteristics of digital data that make it renderable and meaningful over time. This literature review examines some of those research studies and the methods used. We define research as those publications that have a research question and a methodology which demonstrates collection of empirical evidence to support arguments. The research on significant properties is hard to summarize because the approaches differ substantially. The methods used, unit of analysis selected, and the categories of significant properties being examined all differ making comparisons difficult.

A number of different qualitative and quantitative methods have been used to identify significant properties. Perhaps in reaction to this, there have also been at least two efforts to standardize collection procedures to identify significant properties – InSPECT (Knight 2009) and the Data Curation Profiles Project being carried out jointly by Purdue University and the University of Illinois Urbana-Champaign (Witt, Carlson, Cragin, & Brandt, 2009).

The InSPECT Assessment Framework modifies the Function-Behavior-Structure (FBS) framework originally created to help engineers redesign systems. There are two parallel

analysis activities, one focusing on the information object where "The evaluator analyses a representative sample of an object type, identifies a set of functions and behaviours that may be achieved, and the properties that are necessary for their performance" and the other concentrating on the stakeholders where "The evaluator identifies one or more stakeholders that have some relationship with the Information Object and analyses the functions that they wish to perform" (Knight, p. 13). Both the object and stakeholder approaches consist of a seven step analysis process. This process was applied and tested in several InSPECT reports (e.g. Coyne et al. 2007; Coyne and Stapleton 2008; Grace et al. 2009).

The Data Curation Profiles are a means of "exploring, learning about and interacting with data producers and collecting information about datasets and collections" (Witt et al. p. 101). The profiles are developed from in-depth interview protocols, which include written reflections on the part of the respondents/data producers. The Purdue/University of Illinois researchers have also developed a standardized template for others who want to use the data curation profiles methodology to collect information about data producers and data sharing and reuse needs.

Whether researchers are using the above methodologies or not, most chose a method from a group that is largely qualitative to identify significant properties: interviews, observations, surveys, and document analysis. Furthermore, two methodological approaches are apparent: one that focuses on the data and one that concentrates on the producers and reusers (stakeholders). Both approaches have roots in the digital preservation community. Stakeholder analysis for digital preservation was pioneered by

Jeff Rothenberg and Tora Bikson (1999) at the National Archives of the Netherlands. Using an ethnographic methodology and collecting interview and observational data, they uncovered work practices and significant characteristics of information critical for digital preservation. In contrast, the digital document or data analysis was used extensively in InterPARES 1 and 2, applying a form of modern diplomatics to digital records (MacNeil 2000). The people-centric and data-centric approaches represent significant philosophical differences to the question of how significant characteristics should be identified and analyzed. Is the essence of significant properties within the digital object itself, or is it in the contexts of creation, archiving, and reuse? These approaches also tend to map on to an emphasis on certain types of significant properties. Using primarily digital document analysis with minimal emphasis on the practices of producers, curators, and reusers, the digital curation literature focuses primarily on properties of the digital objects. The data reuse literature, with its more ethnographic approach and a focus on producers and designated communities tends to address properties that are important for maintaining meaning over time and enabling designated communities to reconstruct meaning.

In both the digital curation literature and the data reuse literature, qualitative and quantitative methods have been used to identify significant properties; however, there is a clear preference for qualitative methods. As a consequence, much of the research is case studies (e.g. Carlson and Anderson 2007; Coyne, Duce, Hopgood, Mallen, and Stapleton 2007), based on interviews and / or observations with a small number of individuals within one discipline (e.g., White 2010; Zimmerman 2008), or based on a few examples of a file format (Ashley, Davis, and Pinsent. 2008; Coyne and Stapleton 2008). One of

the exceptions is Wilson et al.'s (2006) "Moving Digital Images and Sound Archiving Study" which featured a survey (n=92) of digital object creators, curators, and others interested in sound and moving images.

While the disciplinary focus on practice is necessary to identify some of the non-technical significant properties, there is often little to no discussion of the representativeness of the sample (e.g., Coyne et al., 2007; Matthews, McIlwrath, Giaretta, and Conway 2008) and the generalizability of the findings to others in the same discipline. The problems of sampling and completeness are particularly apparent in the case studies where data collection activities are difficult if not impossible to determine (e.g., Manjula 2008; Conway 2009). The number of interviewees, time spent in observations, the types of documentation reviewed are often not discussed. Finally the number of observations in these studies is usually quite low.

Significant Properties

The divergent emphases on data vs. people lead to different findings in terms of significant properties. Those using the digital diplomatics approach tend to address technical properties for rendering digital objects; the stakeholder (e.g. designated community) focus tends to address properties that are important for maintaining meaning over time.

The digital curation literature cites a number of different significant characteristics to preserve along with the digital objects over time, such as the functionality, relationships among data, and appearance (Coyne et al. 2007); look and feel (Hedstrom, Lee, Olson,

and Lampe 2006; Matthews et al. 2008); rights information (proprietary algorithms or copyright) (Matthews et al 2008; Ashley et al. 2008); purpose and use (Ashley et al. 2008); computing environment (Morrissey 2010); and intention of the creator (Coyne et al. 2008), and usage (e.g., Morrissey 2010). The array of significant properties identified is large and diverse and researchers in computer science have noted the difficulties inherent in identifying and representing contextual metadata (Buneman, Khanna, and Tan 2001; Carter and Green 2009). One response has been research into the creation of representation information registries in order to standardize expression of significant characteristics and provide consistency in capturing and delivering them to researchers (Brown 2008).

Data reuse research also identifies important contextual elements cited in this literature, including data cleaning and manipulation processes (Carlson and Anderson 2007), the context of production and explicitness of the data collection methods (Faniel and Jacobsen 2010) , exposing quality checks (Carlson and Anderson 2007), level of documentation (Corti 2007), standards for documentation (Faniel 2009), identity of the data collector (and as a corollary perceived expertise by the data reuser and trust) (Knorr Cetina 1999; Van House 2002a and 1998; Zimmerman 2008), and the rights of subjects and the people associated with the objects (Carlson and Anderson 2007). The literature also points to the difficulties of identifying and maintaining the context for meaning making (e.g. Berg & Goorman 1999; Jirkota et al. 2005; Cragin and Shankar 2006).

Taken together the digital curation and data reuse studies give us an overview of significant properties of digital data that enable renderability and meaningfulness of data

over the long term. We find that there is little overlap between the sets of significant properties identified. Moreover, preservation regimes are primarily working to ensure the former not the latter. For successful reuse over the long term, attention to both is necessary. Moreover, the research on the significant properties of reuse needs to consider whether and how they can be represented and preserved. There is definitely a need for cross-disciplinary work in this area that combines the data- and people-centric approaches and works toward a more integrative model of significant properties (Faniel and Zimmerman 2011).

Dissemination Information Packages for Information Reuse (DIPIR)

The DIPIR project builds on this previous research as it attempts to take the research on contextual metadata to the next level. In this section we will: 1) provide a brief overview of the DIPIR project, 2) discuss the benefits of the comparative nature of our study, and 3) outline our mixed methods approach to the research questions.

The DIPIR project is a three-year Institute for Museum and Library Services funded project to investigate the relationship between significant properties and data reuse. Our specific research questions are:

1. What are the significant properties related to the reuse of digital data in three disciplines?
2. How can these significant properties be expressed as representation information in OAIS to ensure the preservation of meaning to enable the reuse of data?

Our research approach is more people- than data-centric. We are particularly concentrating on the significant properties that enable data reuse but we see these are contributing to a fuller understanding of the essence of objects that needs to be preserved.

Our partners in this project are the Inter-university Consortium for Political and Social Research (ICPSR), Open Context, and the University of Michigan Museum of Zoology (UMMZ). Within these organizations, we are investigating how significant properties affect data reuse across three disciplines: quantitative social scientists, archaeologists, and scientists (e.g., climatologists and biodiversity researchers) respectively. ICPSR was chosen for three reasons: 1) ICPSR draws a wide variety of users from across the social sciences, 2) it is a highly compliant organization utilizes the most up to date digital preservation standards and practices and was recently awarded the Data Seal of Approval, 3) it holds well-structured and homogenous digital data, and 4) it provides us with the opportunity to see how well a strong digital curation regime satisfies researchers' needs when reusing these data.

Open Context is a service of the Alexandria Archive Institute (AAI), a non-profit organization "working to provide innovative digital services for sharing research" (Alexandria Archive Institute 2010). Open Context was selected because it 1) is a new service with a focused disciplinary community, 2) features multiple formats, such as field notes, maps, and images, and 3) it has developed a rich set of contextual information about the digital objects in the repository and through its online services enables users to visualize and remix content to suit their research questions. Data on this site are largely qualitative and unstructured even though ArchaeoML is used to structure the metadata

around the digital objects. Open Context is also midway between ICPSR and the UMMZ in terms of standardized representation of information.

The UMMZ is an internationally recognized center of biodiversity research with outreach efforts from K-12 through PhD. UMMZ is currently working on integrating metadata, images, and the context of the data collection (Field notes) into a persistent digital repository to support research in biodiversity as well as to support K-12 education. The Museum of Zoology was chosen because of the emergent nature of metadata standards in this area, and the nature of the data: unstructured field notes, metadata. It was also chosen because although there has been substantial literature on reuse of scientific data, none of this deals with the vast amount of scientific data in museums. Finally, the problems with integrating contextual information are very salient since much of the metadata has been applied by specialists in a single sub-discipline (e.g., fish, mollusks) with particularly types of reuses or research questions in mind; but now that biodiversity and ecological researchers are interested in these data, different types of contextual information have emerged as being salient.

One of the weaknesses we detected in the significant properties literature was the case study, disciplinary, or format focus of a majority of the research. One of goals of the DIPIR project is to implement a research design across the three sites that allows for comparison and contrast both among and within the sites. Our plan in following this design is to be able to reach some generalizations about significant properties across disciplines, formats, and purposes for reuse.

Our research design employs a variety of qualitative and quantitative methods across the sites (see Table 1 for an overview) and can be roughly divided into four phases: 1) staff interviews, 2) data collection from ICPSR reusers, 3) data collection from Open Context reusers, and 4) data collection from UMMZ reusers. We are currently in phase 2 of the research. We began by conducting interviews with 25 staff members at the three sites. We posed questions to staff concerning their work in data selection and acquisition, ingest, archival storage, and dissemination. These interviews are now being analyzed and we have a good idea about data management practices across the three repositories as well as interactions with the designated communities.

Table 1: Data Collection Methods

	ICPSR	**Open Context**	**UMMZ**
Interviews (Staff)	10	4	10
Interviews (Data reusers)	40 (Summer 2011)	20 (Fall 2011)	20 (Spring 2012)
Survey of Data Reusers	2000 (Fall 2011)		
Web Analytics		On-going	
Observations			10 (Fall 2012)

We are now interviewing novice and expert reusers of data from ICPSR. Our goal is to conduct 40 interviews with academics in a variety of disciplines. The interviews are intended to help us formulate the survey and to assist us in understanding whether

differences exist in the amount and/or kind of contextual information needed depending on expertise and discipline. Our protocol is divided into three sections, personal information about the interviewees' level of experience reusing data, a critical incident discussion about the last time the individual reused data, and more general questions about data reuse practices and information needs. We will use a version of this protocol with researchers at Open Context and UMMZ, although at Open Context we will focus deeply on one discipline and at UMMZ we will interview individuals from the core disciplines as well as the newer biodiversity and ecological researchers who use the historical data in their work.

In order to test several hypotheses, we are preparing a large-scale survey of ICPSR data reusers. Our population for this survey is first authors in the extensive bibliography of articles citing ICPSR data which is maintained by ICPSR. The entire bibliography contains well over 5000 entries. We are sampling the bibliography and using authors who have published in the most recent 3-4 years, after a major revision of the ICPSR delivery system. Our plan is to distribute the survey to approximately 2000 first authors and test several hypotheses about reuse and significant properties. The goal here is to develop a model of data reuse and understand the role significant properties play in that reuse.

We will utilize Web analytics from the Open Context website to examine quantitative patterns of data use. This is an exploratory study to determine whether web analytics can be used to both determine whether online behavior can inform us about significant properties of data reuse. We will be particularly interested in how visitors to the Open

Context site utilize its data analysis tools for visualization and remixing data. Since Open Context gives its users more control over data analysis on the website, we will also examine how significant properties influence users' control and specificity of the data retrieved for reuse.

For our final data collection activity, we will observe data reuse at the Museum through a series of think-alouds with researchers. Our goal is to observe their work processes in order to understand how they juxtapose information and how this aids in the reuse process. Since the museum includes both digital and analog data we are interested in seeing how people navigate between the two types of data and how significant properties are important in these interactions.

Our research design is also constructed to address the lack of specificity in the methods we detected in the significant properties literature. We want to ensure that the claims we make about contextual metadata have sufficient evidence and that we are clear about where that evidence is from and the nature of the data we ourselves are using.

Conclusion:

Digital data has many properties and their significance can differ between users or the purposes of reuse. In this article, we have discussed the research methods used to identify significant properties and how underlying assumptions about how to derive those properties (from the producer or reuser or from the data itself) lead to specific types of properties. A substantial amount of work has gone on to identify significant properties, but we are now ready to take this research to the next level. Many of the articles are not

as transparent as we would like about their methods. Once one moves beyond the formal methodologies established by InSPECT and the Data Curation Profiles, most studies do not provide sufficient information for replication. The in-depth studies of disciplinary practices and formats are invaluable but the next step is to design research that features larger samples across formats and disciplines to enable generalization where possible. The DIPIR project seeks to do just that, develop and test a research design that allows for comparison and works through some of the problems of small sample size by triangulating different types of data (survey, interview, observation, web analytics) to create a better picture of what contextual data is needed for reuse.

References

Alexandria Archive Institute, http://www.alexandriaarchive.org/

Ashley, K., Davis, R., & Pinsent, E. (2008). Significant properties of e-learning objects (SPeLOs). Retrieved April 4, 2011 from http://www.jisc.ac.uk/media/documents/programmes/preservation/spelosfinal.pdf

Berg, M. & Goorman, E. (1999). The contextual nature of medical information. *International Journal of Medical Informatics. 56*, 51-56.

Brown, A. (2008). *White paper: Representation information registries*. Retrieved from http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

Buneman, P., Khanna, S. and Tan, W. (2001). Why and where: A characterization of data provenance. In J. Van den Bussche, and V. Vianu (Eds.). Proceedings from the International Conference on Database Theory (pp. 316-330). London, UK: Springer, LNCS 1973.

Carlson, S. & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication 12*(2).

Carter, P. E. & Green, G. (2009) Networks of contextualized data: a framework for cyberinfrastructure data management. *Communications of the ACM 52*(2).

Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS)*. (CCSDS 650.0-B-1). January 2002. Washington DC: National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf.

Conway, E. (2009). Curating Atmospheric data for long term use: Infrastructure and Preservation Issues for the Atmospheric Sciences community. SCARP Case Study 2, Digital Curation Centre. Retrieved July 17, 2011 from http://www.dcc.ac.uk/scarp

Corti, L. (2007). Re-using archived qualitative data – where, how, why? *Archival Science 7*(1), 37-54.

Coyne, M., Duce, D., Hopgood, B., Mallen, G., & Stapleton, M. (2007). The significant properties of vector images. Retrieved April 4, 2011 from http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf

Coyne, M. & Stapleton, M. (2008). The significant properties of moving images. Retrieved April 4, 2011 from http://www.jisc.ac.uk/media/documents/programmes/preservation/spmovimages_report.pdf

Cragin, M.H. & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work, 15 (2-3).*

Faniel, I.M. & Jacobsen, T. (2010). Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work*, 19(3-4): 355-375

Faniel, I.M. & Zimmerman, A. (2011) Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation* 6/1, 58-69.

Grace, S., Knight, G., Montague, L. (2009). Investigating the Significant Properties of Electronic Content over Time – InSPECT Final Report. Retrieved April 21, 2010 from http://www.significantproperties.org.uk/inspect-finalreport.pdf.

Hedstrom, M. Lee, C. Olson, J. and Lampe, C. (2006). "The Old Version Flickers More": Digital Preservation from the User's Perspective. American Archivist., 69 (1), 159-187.

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., Voss, A. (2005). Collaboration and trust in healthcare innovation: The eDiamoND case study. *Computer Supported Cooperative Work*. 14(4), 369-398.

Knight, G. (2009). InSPECT Framework Report. Retrieved from http://www.significantproperties.org.uk/inspect-framework.pdf

Knorr Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA, Harvard University Press.

Ludäscher, B., Marciano, R., Moore, R. (2001). Towards self-validating knowledge-based archives. *11th Workshop on Research Issues in Data Engineering (RIDE)*. Heidelberg: IEEE Computer Society.

MacNeil, H., Wei, C., Duranti, L., Gilliland-Swetland, A., Guercio, M., Hackett, Y.,…, Xiu, Z. (2000). Authenticity task force report. Retrieved from http://www.interpares.org/book/interpares_book_d_part1.pdf http://www.interpares.org/display_file.cfm?doc=ip1_atf_report.pdf

Manjula, P. & Ball, A. (2008). Challenges and Issues Relating to the Use of Representation Information for the Digital Curation of Crystallography and Engineering Data. *International Journal of Digital Curation* 1/3, 76-88.

Matthews, B. McIlwrath, B. Giaretta, D. Conway, E. (2008). The Significant Properties of Software: A Study. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops

Moore, R.W. & Smith, M. (2006). Assessment of RLG trusted digital repository requirements. Paper presented at the workshop *Digital Curation & Trusted Repositories: Seeking Success*. Chapel Hill: Joint Conference on Digital Libraries. Retrieved from http://sils.unc.edu/events/2006jcdl/digitalcuration/Moore_Smith-JCDLWorkshop2006.pdf

Moore, R.W. & Smith, M. (2007). Automated validation of trusted digital repository assessment criteria. *Journal of Digital Information, 8*(2). Retrieved from http://journals.tdl.org/jodi/rt/printerFriendly/198/181

Morrissey, S. (2010) The Economy of Free and Open Source Software in the Preservation of Digital Artefacts. *Library Hi Tech* 28/2, 211-223.

Rothenberg, J. & Bikson, 1999. T. Carrying Authentic, Understandable and Usable Digital Records Through Time: Report To the Dutch National Archives And Ministry of the Interior. Retrieved June 1, 2011 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

Van House, N. A. (2002a). Digital libraries and practices of trust: Networked biodiversity information. *Social Epistemology 16*(1), 99-114.

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. In S. Poltrock & J. Grudin (Eds.) Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (pp. 335-343). New York: ACM

Vardigan, M. & Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. *Archival Science 7*, 73–87.

White, H.C. (2010). Considering Personal Organization: Metadata Practices of Scientists. *Journal of Library Metadata* 10:2-3, 156-172.

Wilson, A., Wright, R., Polfreman, M., Anderson, S., Tanner, S., Beer, E. (2006). Moving Digital Images and Sound Archiving Study. Retrieved April 20, 2011 from http://www.jisc.ac.uk/media/documents/programmes/preservation/movingpicturesandsoundarchivingfinalversion.pdf

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

Zimmerman, A. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology, & Human Values, 33(5), 631-652.