Documenting provenance in non-computational workflows: research process models based on geobiology fieldwork in Yellowstone National Park

Andrea K. Thomer[1*], Karen M. Wickett[2], Karen S. Baker[3], Bruce W. Fouke[4], Carole L. Palmer[5]

[1]University of Michigan School of Information, 105 S. State St., Ann Arbor, MI 48109 USA
athomer@umich.edu

[2]School of Information, University of Texas at Austin, 1616 Guadalupe Suite #5.202, Austin, Texas, 78701-1213 USA
wickett@ischool.utexas.edu

[3]INTERACT Research Unit, PO Box 8000, FI-90014 University of Oulu, Finland;
School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, Champaign, Illinois, 61820 USA
karensbaker@gmail.com

[4]Department of Geology, University of Illinois Urbana-Champaign, 1301 W. Green Street, Urbana, Illinois 61801 USA
Department of Microbiology, University of Illinois Urbana-Champaign, 601 S. Goodwin Avenue, Urbana, Illinois 61801 USA
Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, 1206 W. Gregory Drive, Urbana, Illinois 61801 USA
fouke@illinois.edu

[5]Information School, University of Washington, Box 352840 - Mary Gates Hall, Ste. 370 Seattle, WA 98195-2840 USA
clplamer@uw.edu

* Corresponding author

Abstract

A comprehensive record of research data provenance is essential for the successful curation, management and reuse of data over time. However, creating such detailed metadata can be onerous, and there are few structured methods for doing so. In this case study of data curation in support of geobiology research conducted at Yellowstone National Park, we describe a method of "Research Process Modeling" for documenting non-computational data provenance in a structured yet flexible way. The method combines uses systems analysis techniques to model research activities, the PROV ontology to illustrate relationships between data products, and simple inventory methods to account for research processes and data products. It also supports collaborative data curation between information professionals and researchers, and is therefore a significant step toward producing more useable and interpretable research data. We demonstrate how this method describes data provenance more robustly than "flat" metadata alone and fills a

critical gap in the documentation of provenance for field-based and non-computational workflows. We discuss potential applications of this approach to other research domains.

Introduction

Methodological and analytical reproducibility are fundamental to scientific practice. Comprehensive documentation of data collection and analysis methods are therefore critically important as well. This documentation is needed to replicate results, assess the validity of a study, and facilitate reuse of data. In other words, future users must fully understand a dataset's *provenance* and *context of production* (Faniel & Jacobsen, 2010; Vertesi & Dourish, 2011; Weber, Baker, Thomer, Chao, & Palmer, 2013): the why, where, when, how, and by whom of a dataset's creation.

Traditionally, data provenance and other important contextual information about laboratory and field conditions have been recorded by hand (literally), in written documentation such as field and laboratory notebooks; concise summaries of these notes are then published in the "materials and methods" sections of journal articles. However, computational modes of analysis have motivated computational methods of capturing context and provenance. An executable record of the algorithm behind a computational analysis – rather than just a description of the algorithm – is needed to make a computational analysis truly reproducible (Mesirov, 2010; Peng, 2011; Stodden, Guo, & Ma, 2013). Consequently, more and more researchers have urged the use of version control platforms such as GitHub to share and collaboratively maintain their codebases; workflow systems such as Taverna or Kepler to capture and publish computational research workflows (also called *in silico* workflows); or workbenches such as myExperiment or the Open Science Framework to aggregate and manage projects. Combined, these tools can be used to create datasets that render *in silico* research methods not just reproducible, but exactly re-executable (see Bechhofer et al., 2013).

However, despite the growing prevalence of *in silico* modes of research, there are many branches and phases of science that are *ex silico* – that is, not conducted on a computer. There is, therefore, still a need to document provenance and research processes "by hand". This is particularly true for sciences that require significant research in the natural world ("fieldwork"), and that use field observations and the analysis of physical specimens as data (hereafter referred to as "field-based" research). Though field-based researchers certainly use computers for data management and analysis, they often must collect and integrate data by hand (and even sometimes on paper). This work often cannot be automated for several reasons. Firstly, field-based research can take place in some of the least computer-friendly data collection settings on earth: volcanic craters, hot springs, cliff-faces, and deep undersea vents that subject equipment and their human operators to all kinds of risks and disturbances. Researchers may need to leave delicate laptops behind during these trips, collect data by hand or with specialized tools, and integrate it later. Secondly, even when computational data processing and analysis *is* possible, workflows are often distributed among multiple computers. For instance, rock samples may need to be sent to other labs for radioisotope analysis, or biological samples may need to be sent to special facilities for genetic sequencing. Consequently, the relationships between these data products and the research processes used to create them can't be tracked automatically because they take place on different, disconnected computers.

Manual, *ex silico,* and distributed research processes are just as important to capture and disseminate as automated, *in silico,* localized processes. Lightweight methods of capturing these

ex silico workflows are needed. However, due to the diversity of data collection methods in field-based sciences, it is difficult to establish broadly applicable best practices for data capture, and harder yet to reconcile *ex silico* documentation with the highly structured provenance graphs produced through localized *in silico* workflow systems. Where *in silico* processes need to be recorded in a way that's re-executable, the activities of field-based science need be recorded in a way that is re-traceable: other researchers need to be able to reconstruct when, where, how, and in what order data points were collected, and under what conditions. Further, the original and interpreted data needs to be represented in ways that support reuse by researchers across broad disciplines.

In this paper, we present a case study of long-term field-based research conducted in Mammoth Hot Springs at Yellowstone National Park by a geobiologist and his research team, developed as part of the Site-Based Data Curation project. This case study enables elaboration of a method for documenting high-level research processes, which we call *Research Process Modeling*, that represents both computational and non-computational processes alike, thus bridging a critical gap in existing workflow capture and documentation methods. We discuss how Research Process Modeling can help researchers and information professionals collaborate by identifying *points of intervention* for improving data curation practices and the division of curatorial responsibilities. We also show how this method can be used to create semantically-rich research objects, in a similar vein as previously described (Bechhofer et al., 2013; Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010; Belhajjame et al., 2014; Corcho et al., 2012; Hettne et al., 2014), and make explicit the often-obscure relationships among components of complex data objects. Additionally, the method may inform growing efforts to develop standard, interoperable and broadly applicable workflows for tasks such as metadata rescue and sample curation (e.g. Hills, 2015). While much of scientific fieldwork will never be done with a push of a button – and consequently, much of a dataset's provenance or context of production will never be automatically recorded – we believe the approach presented here represents a significant step toward in supporting the creation of more robust documentation of field-based research.

Background: Provenance and reproducibility through workflow capture and modeling

Rooted in art history and archival practices, the concept of *provenance* now serves as a guiding principle in documenting the creation of data products, and thereby facilitating scientific reproducibility (Tilmes, Yesha, & Halem, 2010). In art history and archival work, provenance refers to an object's "chain of custody" through time: the chronological documentation of an object's custodian, which can be used to validate claims of authorship, ownership, or authenticity. When applied to scientific data processing, provenance refers to the history of changes made to a dataset in addition to its "chain of custody" from one instrument to another, one process to another, one researcher to another, or one format to another. This is necessary not just for scientific reproducibility, but also for "understanding of data and analyses, auditing, and anomaly resolution" (Tilmes et al., 2010, p. 548).

Where the provenance of art and archival materials tends to be linear, scientific data requires a graph model capable of expressing the complex many-to-many relationships between agents and objects, as well as the one-to-many relationship between a data object and later derived data products. The Open Provenance Model (OPM) exemplifies this approach, in which "[the]

provenance of objects (whether digital or not) is represented by an annotated causality graph, which is a directed acyclic graph, enriched with annotations capturing further information pertaining to execution" (Moreau et al., 2011). OPM and the W3C PROV models (*PROV-DM: The PROV Data Model*, 2013) both give an over-arching view of the particular artifacts, processes, and agents, and the relations among them, that contributed to a specific data product.

A number of ontologies have been developed specifically to describe the provenance and context of scientific data. Particularly relevant to this work are efforts focused on describing data from field sites or derived from physical samples. For instance, ontologies such as the Environment Ontology (ENVO) and the Biocollections Ontology (BCO) can be used to contextualize field observations by supporting detailed descriptions of their originating environment. ENVO describes environmental entities and their qualities (Buttigieg et al., 2016); whereas BCO models the sampling processes used to collect biodiversity data (Walls et al., 2014). Other approaches, such as the ontology for observations and sampling features, aim to provide "domain-neutral" vocabularies describing the observations and their associated properties themselves (Cox, 2015), with alignments to other observation and measurement standards such as ISO 19156 (Cox, 2013). Recent work by Cox and Car is particularly relevant to this study; they most explicitly explores the application of the PROV data model to "Real Things" – specifically, physical samples taken by geologists in the field (Cox & Car, 2015).

Approaches like Cox and Car's provide important formal syntax for the description of scientific data provenance, particularly as data are manipulated in both the physical and digital worlds. However, to be broadly and efficiently applied, they require a mechanism for application, as well as guidelines regarding what to document, and in what detail. Developing these best practices will require drawing on existing approaches to provenance capture and workflow documentation.

The workflow paradigm

As briefly reviewed above, the complex modes of modern data processing and analysis – particularly the dependencies that arise in a computational environment – have necessitated efforts to more fully document the provenance of the specific computational events and entities involved in algorithmic data analysis. It is not enough to simply share abstract, prose descriptions of algorithms: we must instead share the computational processes themselves – preferably in a re-executable format (Mesirov, 2010; Peng, 2011). This "workflow paradigm" (Hettne et al., 2014) is predicated on the ability to automatically record *in silico* workflows (e.g., the precise descriptions of the sequenced execution of a "computational process, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource" (Goble & De Roure, 2009, p. 138) as they are executed. These programs ideally function as a kind of "sheer curation" in which, "curation activities are quietly integrated into the normal work flow of those creating and managing data and other digital assets" (Hedges et al., 2012, p. 1).

The resulting workflows can, in turn, be treated as first order research objects in and of themselves. With careful "process curation" (Goble, Stevens, Hull, Wolstencroft, & Lopez, 2008) and the application of workflow-centric ontologies (e.g. those outlined in the Workflow4Ever Research Object Model; Belhajjame et al., 2013), they can be re-used to make

rote tasks more efficient, and can be shared to make analytical methods more transparent and reproducible. That said, process curation requires considerable care. Computational workflows are often dependent on idiosyncratic third-party resources and operating system specificities, and can consequently be incredibly fragile (Zhao et al., 2012). Thus, data curation best practices are needed that account for the curation of workflows as well as data.

Further, the "workflow paradigm," as currently realized, is only strictly applicable to computational data analysis; the approaches described above are not immediately applicable to non-computational research processes, which must therefore be captured in other ways. Research by those within the data curation community provides a first step toward bridging this gap. For instance, interview protocols developed through the Data Curation Profiles project can be used to guide information professionals in gathering necessary data provenance from data creators (Witt, Carlson, Brandt, & Cragin, 2009). The DCPVocab was developed to provide specific terms for representing relationships among research practices, types of data, and curation roles and activities (Chao, Cragin, & Palmer, 2015). Additionally, the many data lifecycle models developed by curation communities may be thought of as an approach to non-computational process representation (e.g. CCSDS, 2012; Faundeen et al., 2013; Higgins, 2008). These models can be helpful when planning or describing data curation work in broad terms. However, they are ultimately insufficient as detailed models of data provenance for the purpose of data reuse or reproducibility (Ball, 2010, p. 14).

Systems analysis and provenance

Data curation lifecycle models and the computational workflow capture systems share a common ancestor in systems analysis, which we have found instrumental to this work. Systems analysis techniques can be used to capture processes of information creation, flow, and management in a comprehensive manner, thereby describing an entire process as enacted by an organization (Aalst & Hee, 2004). They document and visualize the needs and activities of an organization or individual at the point at which they are performing these actions, as opposed to being rooted in down-stream procedures. In application, systems analysis techniques give an account of agents, objects, and processes spanning computational and non-computational environments, with clear connections to the entire high-level process. Activity diagrams are particularly useful; they model workflows in a relatively simple yet logical manner and, "can be used to describe the current as-is system and the to-be system being developed" (Dennis, Wixom, Tegarden, & Seeman, 2015). Often, processes and workflows are modelled through structured notations such as the Unified Modeling Language (UML), a method of visualizing a system's processes, inputs and outputs through standardized diagrams.

Researchers in data curation have previously utilized systems analysis techniques, despite a lack of best practices for their application to curation environments (for example Ball, 2012; De Roo, De Maeyer, & Bourgeois, 2016; Hills, 2015; Williams & Pryor, 2009). Here, we similarly draw on systems analysis and further define best practices for their application to curation environments. We complement UML-structured Activity Diagrams with provenance graphs and simple inventories of the processes and data products involved in a project (hereafter we refer to these products as "artifacts", in acknowledgement of their role as quasi-archaeological evidence of past research processes – as well as evidence in scientific investigations). Through this

combination of methods, we record detailed information about the collection of observational data and physical samples by scientists in the field. Additionally, we find that systems analysis methods offer a mechanism for the structured application of ontologies – particularly PROV. We develop this approach through a case study of geobiology research at Yellowstone National Park, conducted as part of the Site-Based Data Curation (SBDC) project, through which we developed a method of provenance-enriched Research Process Modeling to support the documentation of field-based research processes.

The Case: Geobiology at Yellowstone National Park

Geobiology is an exemplar of the kind of integrated science increasingly in need of structured, retraceable field process documentation. An interdisciplinary and relatively young domain, geobiology is the study of how microorganisms influence the geology and whole ecosystem of the earth, and how earth environments in turn influence the behavior and evolution of microbes – in other words, of how microbes "eat", "breathe", and "make" rocks to adapt, survive and evolve. These interactions take place in some of the most extreme environments in existence: hot springs, deep-sea vents, and potentially even other planets. Geobiology research is dependent on a combination of manual collection of physical samples, laboratory analysis, and computational analysis. Working with data across sites for systems-driven, integrative work can only be done when data collection methods and each site's conditions and context are explicitly documented (Fouke, 2011; Fouke & Murphy, 2016).

Yellowstone National Park (YNP) is an important and popular site for data collection in geobiology. The park is a well-protected, well-studied, easily accessible and extremely diverse research environment. Researchers at YNP can select study sites from 12,000 thermal features, the largest collection of hot springs anywhere in the world, allowing the careful dissection of undisturbed natural systems that are similar to those that originally formed on the ancient earth (Fouke, 2011). Furthermore, because the hot springs at YNP are all part of the same natural system, studies of individual hot springs can potentially be integrated to support broader investigations of the geothermal system as a whole.

A typical geobiology research project starts with reconnaissance and hypothesis testing in the field. Multiple kinds of field observations (biological, chemical, physical, geological, and genomic) are recorded and water and rock samples are collected. This data is captured in a range of file formats. Some data are handwritten in paper lab notebooks; some are entered into spreadsheets by hand in the field; and some are "born digital" outputs from hand-held instruments. The physical samples may be sent to external laboratories for analysis if facilities for special analysis (e.g. mass spectrometry, radiogenic isotope analysis) aren’t available at the researcher’s home institution). The results are typically emailed back in spreadsheets. Different data components arrive at different times and are ultimately combined and synthesized centrally in a geobiology lab.

Through the SBDC project, we sought to support the aggregation and integration of geobiology data within and across scientifically significant sites. In collaboration with geobiologists and National Park Service (NPS) personnel, we developed a Minimum Information Framework of key information classes that ought to be prioritized for collection and curation (Palmer et al.,

2017). We additionally used the approaches described herein to identify optimal points of curatorial intervention in the research workflow; these are points at which data should be optimally documented and managed, thereby making field-based processes retraceable, and the data collected reliably interpretable and reusable. While our documentation method was developed out of this geobiology case, it can address data quality issues connected to the interpretation of data that have been observed in as disparate fields as condensed matter physics (Stvilia et al., 2015).

Method

Overview

To design effective curatorial interventions, we first analyzed typical research workflows and data products and then sought to determine the appropriate division of labor between researchers and information professionals for curation activities. This strategy was grounded in a prior stakeholder analysis conducted with earth science researchers and park service resource managers at YNP (Thomer et al., 2014).

Our primary collaborator, a Geobiologist, provided us with a hard drive containing over ten years of his and his students' field and laboratory data. We first surveyed the contents of the Geobiologist's research hard drives and paper field notebooks and established the range of data typically collected during a research field trip. We also learned his day-to-day data management practices through review of his file structures, file metadata, paper field notes and file organization system.

After this initial survey, we selected two field seasons of data (e.g. two summers' worth of work, comprising over 400 files representing the range of data types and data collection and management practices) for comprehensive analysis. We created an initial inventory of each season's data through content analysis of individual data products and iterative consultation with their creators. Data products were traced back to the "raw" data from which they originated, and inventoried along with a description the analytical processes or other transformations that created them. Through this work, we realized that a more nuanced approach was needed to fully represent the data collection and transformation processes in geobiology fieldwork. Specifically, information modeling techniques were needed to make the provenance and contents of data objects explicit.

To that end, we developed a method of *Research Process Modeling*. This approach draws on systems analysis and information modeling approaches, and is informed by both our prior work on this project (Palmer et al., 2017), and prior work on computational process curation (Goble et al., 2008) and workflow-centric research objects (e.g. Bechhofer et al., 2010; Belhajjame et al., 2012). The simple inventory described above became one of four components required to document the artifacts, processes, and relationships involved in the collection of physical samples and observational data. The final Research Process Model is composed of two inventories and two diagrams:

1) an activity diagram
2) an artifact inventory
2) a process inventory

4) a provenance graph.

Below, we elaborate on each of these four components through a Research Process Model using data from a 2004 field trip to YNP, in which the Geobiologist's lab collected water, rock and microbial samples and conducted several field experiments at Mammoth Hot Springs in YNP (Fouke, 2011, and references therein).

Activity diagram

The activity diagram (Figure 1) is a flowchart visualizing the different processes (activities) undertaken in a typical research project. In this case, this diagram was created through consultation and collaboration with the Geobiologist and his research team. The geobiologists described each step in their field work; we created a diagram representing those steps and decision points; and they corrected the diagram as necessary.

Figure 1 is constructed using UML activity diagram notation (Dennis et al., 2015). Each rectangle with rounded corners represents an individual process. The control flow arrows connecting the rectangles show the sequencing of the processes. Generally speaking, each process is a discrete set of actions that require certain pre-conditions (other processes that must be complete before the process can begin), inputs, and outputs. Concurrent processes, which run in parallel during the same period in time, are co-located between black bars. In these cases, subsequent processes will not begin until all concurrent processes are complete. Decision points and possible iterations back to earlier processes are represented by diamonds, with branches labeled with the deciding condition. For example, the process, "Identify primary flow path" is followed by a decision point. If the primary flow path (the predominant flow of water in a hot spring) is identified, then the site documentation and data collection processes that follow can begin. If the primary flow path is not identifiable, then the control flow returns to the earlier processes that precede the selection of a data collection site.

For this case, we chose not to visualize data inputs and outputs in the activity diagram, so as to focus the reader's attention on the processes essential to the creation and collection of data and samples. Instead, inputs and outputs are itemized in the artifact inventory and illustrated through the provenance graph.

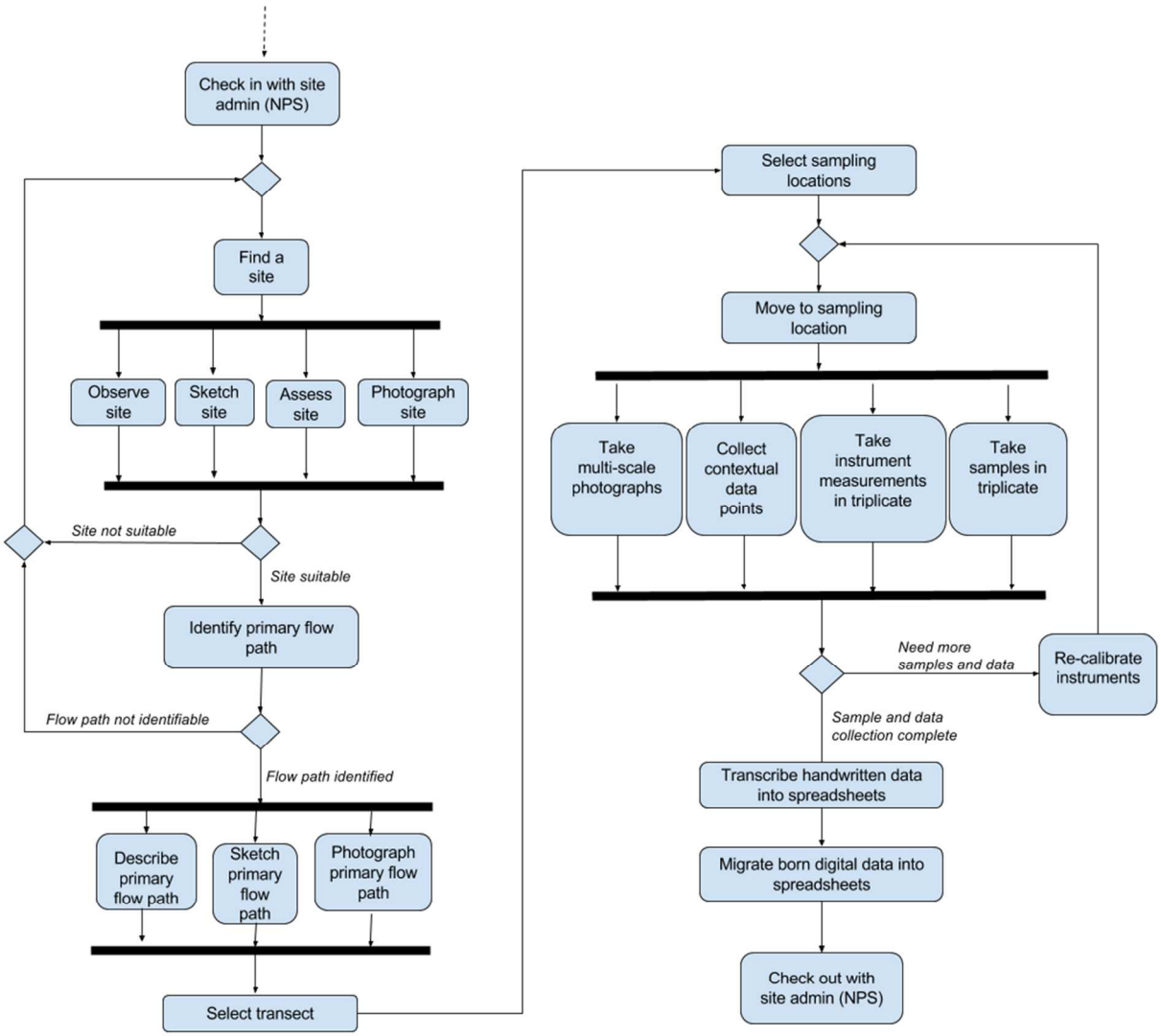

Figure 1: An excerpt of the activity diagram representing the "data collection" phases of research. Rectangles represent a research process; diamonds represent a decision point. A full figure can be found at in our supplemental materials on figshare: http://dx.doi.org/10.6084/m9.figshare.5450809

Artifact inventory

The artifact inventory (Table 1) lists the data artifacts created through the workflow illustrated in Figure 1. Since not all of the data products were preserved on the Geobiologist's hard drive, we list data artifact categories (broad groupings of kinds of data products) as well as the file names of specific instances of those categories. Thus, we are able to represent artifacts that have since been deleted along with those that were preserved.

Each row of the inventory identifies the Artifact UUID, the File Name (if applicable), and the Data Artifact category name; it also describes the Generated By and Used By processes

associated with data collection and transformation, as well as output and input Relationships. Additionally, the inventory identifies the Minimum Information Framework (MIF) superclass and the Formats for each data artifact. The "MIF superclasses" are drawn from our prior work developing a high-level information model for geobiology field data (Palmer et al., 2017). Linking data artifacts to MIF classes is essential for supporting reuse of data for new purposes, and may aid access and retrieval functions as data collections are brought together in repositories over time.

In this example, only a portion of specific artifacts were actually preserved by the Geobiologist and his team; for instance, the original template for "empty data tables" was overwritten in the course of field work and ultimately not preserved. We nevertheless record "empty data tables" in the artifact inventory to support the later creation of provenance graphs. Similarly, we include records for *non-digital* artifacts such as physical samples (noted in the excerpt in Table 1 below) and white board drawings (noted in the full inventory in our supplemental materials) which simply could not be preserved on a hard drive. These, too, must be documented to present a complete view of a dataset's provenance, even if not preserved in a digital data collection.

Artifact UUID	Data artifact category	File Names (if applicable or available)	Generated By [process]	Used by [process]	Related artifacts	MIF class	Format
4ce9a480-0667-479f-882e-d76908818e59	Sample key	Trip Documentation/ Sample_Types 1.doc;	Prepare sample key and labels	Take samples in triplicate	inputFor: physical Samples	Field campaign	.docx
78476b44-fcb9-49cf-beab-d01186683b8e	empty data entry tables	n/a; overwritten by other files	Prepare data entry tables	Take instrument measurements in triplicate Collect contextual data points	usesAsInput: Sampling, data collection, and photography schedules becomes: completed data entry tables	Field campaign - sample plan	n/a
890a52c7-5c51-4f8e-8388-81adb2ee9b46	physical samples	n/a - destroyed through analysis; no permeant field identifiers (e.g. IGSNs) assigned	Take samples in triplicate	[not used further in fieldwork]	usesAsInput: sample key usesAsInput: sample labels usesAsInput: Sampling, data collection, and photography schedules	sample sites and measurements	n/a

Table 1. Excerpt of the artifact inventory. The complete table can be found at http://dx.doi.org/10.6084/m9.figshare.5450809

Process inventory

The process inventory (Table 2) functions as a catalog view of the activities illustrated in the Activity Diagram and provenance graph. Each row of the inventory lists a process's UUID, title, description, responsible agent(s), preconditions, inputs and outputs. Preconditions are described only for processes that result after a decision point has been passed. Inputs and Outputs are artifacts used in or created by a process; these are listed in the Artifact Inventory and the Provenance Graph. Note that we chose to associate responsible agents with processes rather than artifacts; this makes it possible to assign credit and/or responsibility for processes that do not

necessarily create artifacts. While we only describe Agents by a general role in this example, in a repository-ready version of this inventory, they could be listed by name or a unique identifier such as an ORCID (www.orcid.org).

Process UUID	Process Title	Description	Agents	Preconditions	Inputs	Outputs
be4419dc-b9b9-4f1b-b67a-fd4195f6661c	Re-calibrate instruments	reset instruments prior to each new measurement.	Primary researcher; research team	Sample and data collection not complete	equipment operating instructions	n/a
bfd6da52-a732-426f-bfd3-77b8564147d8	Transcribe handwritten data into spreadsheets	type all handwritten data from field notebooks into spreadsheets for back up and later analysis.	Primary researcher; research team	n/a	n/a	Completed and transcribed field notes
addc6e38-7e85-4310-b864-27e02aedaf7f	Migrate digital data into spreadsheets	copy and paste born digital data values from files generated by instrument into spreadsheets	Primary field assistant	Sample and data collection complete	Empty data entry tables	Completed data entry tables

Table 2. Excerpt of the process inventory. The complete table can be found at http://dx.doi.org/10.6084/m9.figshare.5450809

Provenance graph

Where the activity diagram represents coordination and planning among people, the provenance graph (Figure 2) illustrates the movement and transformation of artifacts, as well as the interrelationships among artifacts, agents and processes.

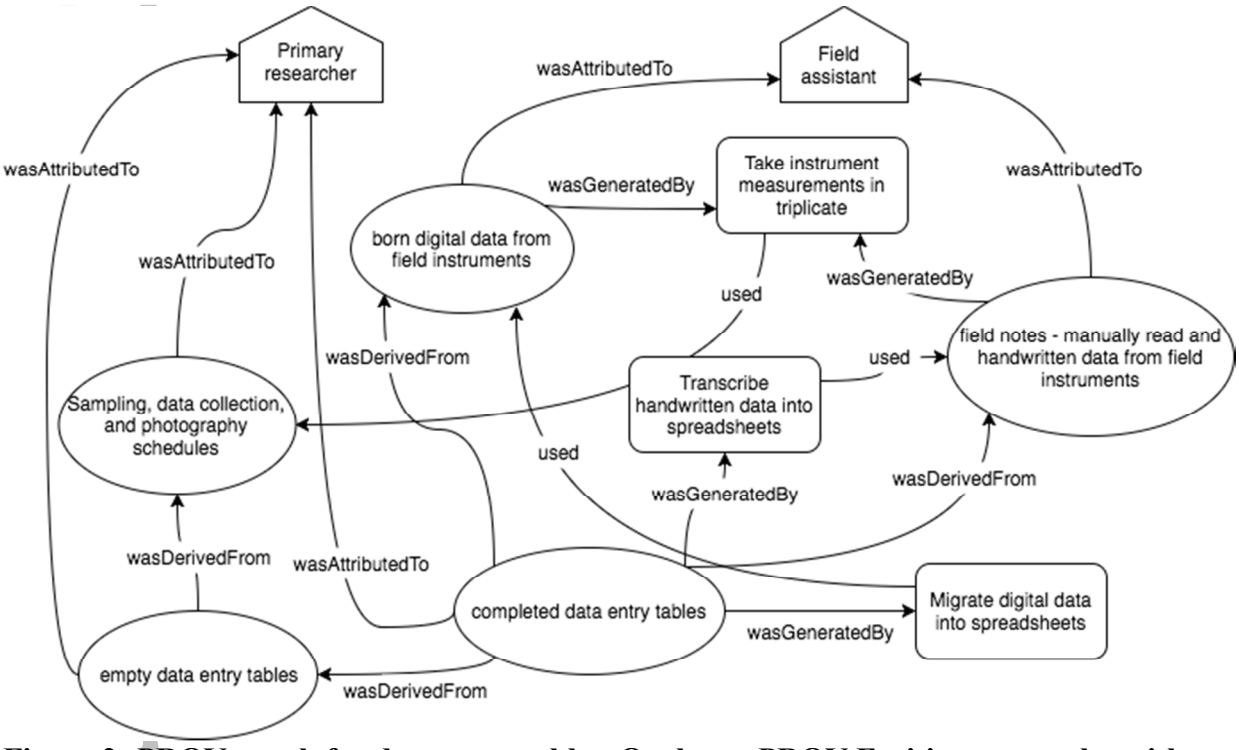

Figure 2: PROV graph for data entry tables. Ovals are PROV Entities, rectangles with rounded edges are PROV Activities, upward pointing pentagons are PROV Agents.

The W3C PROV language accommodates relationships between information objects in wide variety of modes and formats, including physical samples (Cox & Car, 2015). Our graph associates research processes (represented in rectangles) with artifacts modeled as PROV entities (represented in ovals). Agents are represented in the upward pointing pentagons. The representation of Agents in the provenance graph is particularly important for documenting the roles of various contributors to the collection of scientific data and samples.

Arcs in a PROV graph are directed, relating an origin and a target node; the semantic meaning of the arcs vary. Arcs between entities represent the "was derived from" relationship, whereas arcs from an entity to an agent represent an attribution relationship, showing the contribution or control of an artifact. Activities and Entities can be connected by two kinds of arcs. The "used" arc, applied from entities to Activities, indicates which entities were used as inputs for an Activity. The "wasGeneratedBy" arc, applied from an Entity to an activity, indicates the actions that produced the entity as an output.

In Figure 2, we show a PROV graph visualizing the provenance chain for the generation of the "completed data entry tables" entity. It captures the distinct operations related to the two contributors involved in the creation of this artifact: the data spreadsheets as a whole are attributed to the Primary researcher; the intermediate objects containing the digital and manual data are attributed to the Field assistant who generated the measurements in the field. Since PROV is natively expressed in RDF (Cyganiak, Wood, & Lanthaler, 2014), provenance information could be stored as a set of RDF triples, thereby potentially supporting the automated generation of diagrams that focus on particular phases, relationships or on the entire provenance chain for particular objects. However, in this case, the diagram was constructed by hand.

Discussion

Application of the Research Process Modeling approach

At present, field science research processes are documented in diverse, *ad hoc* ways. Research Process Modeling provides a standardized approach for documenting diverse data collection processes in the field, without introducing artificial constraints on the work of researchers. The Geobiologist that provided this case originally documented his field processes via Word documents and spreadsheets inventorying physical samples and field methods; a customized sample labeling system; instrument-specific data and metadata outputs; hand-annotated maps and photographs; and more. Our approach does not supplant these methods, but rather, renders the processes that created each data product (as well as the relationships among data products and contributing Agents) more explicit.

Optimally, Research Process Modeling should be a role for information professionals working in sustained collaboration, or through iterative consultation, with researchers. The models could be created retrospectively after a project is completed, as they were in this case, or drafted prospectively before a project begins (see McPhillips et al., 2015 for further discussion of retrospective and prospective provenance). When constructed retrospectively as done here, Research Process Models can guide consideration on how to curate data earlier and more effectively in future projects. Activity diagrams, in particular, can be used to identify points in the research workflow where curation intervention would be most beneficial. For instance, our case indicated that physical sample curation would be more efficient if samples were registered with the Solid Earth Sample Registry and assigned an International Geo Sample Number (www.geosamples.org). Taking these steps at specific points while data are collected in the field would make subsequent data management more streamlined and efficient.

Data management tasks for a given project are often distributed among team members in unplanned and sometimes counterproductive ways (Wu, Worrall, & Stvilia, 2016). However, through precise articulation of how research processes unfold and what they produce, we can determine the specific curatorial actions needed for each data product, when they should be performed, and by whom. If constructed prospectively, a Research Process Model can help stakeholders to negotiate this division of labor. For example, the Activity Diagram acts as a map of the research process that makes it easier to determine which curatorial tasks can only be accomplished by a researcher (e.g., describing field instrument calibrations, collecting contextualizing data points in the field) and those best delegated to an information professional (e.g., bundling data products for long-term preservation, reformatting data). Moreover, as a project progresses, a prospective Research Process Model can be updated to include unanticipated processes and data products. Given the broad recognition that data curation is most effective when started during the planning phase of a project, prospective Research Process Modelling would likely be optimal in most cases.

Research Process Models can also be used prospectively for training of laboratory staff and students. In this case, we used the Activity Diagram as a teaching tool to prepare undergraduate students in the Geobiologist's "Yellowstone Biocomplexity" class for fieldwork in YNP. We

presented the Activity Diagram to students as part of a tutorial on data management best practices. We asked students to reflect on how this workflow compared to their planned fieldwork, and consulted with each student to identify key data and metadata products that ought to be collected at each step. Students were additionally provided a spreadsheet template outlining important parameters from the Minimum Information Framework; we reviewed this template in light of the Activity Diagram to identify when they might collect key MIF parameters. The students reported finding this prospective discussion of data collection methods and best practices helpful in their work, and many were successful at producing robust metadata in their field books that would be key for curating and sharing their data. The spreadsheet template, as well as an excerpt of one student's completed template are available in our supplemental materials (http://dx.doi.org/10.6084/m9.figshare.5450809); this work is also discussed further in (Palmer et al., 2017).

Whether retrospective or prospective, creating a Research Process Model represents a serious curatorial intervention in and of itself. The diagrams and inventories precisely represent data products, processes, agents and relationships, and therefore are excellent starting points for the creation of standardized metadata for the output of a research study. While the Research Process Model includes many intermediate products not necessary or appropriate for sharing or dissemination (e.g., travel documentation or empty data entry templates), it would nevertheless be straightforward to extract information to produce the metadata necessary for access and reuse by others. Thus, the Research Process Model makes the data bundle more repository-ready by creating a "first draft" of machine-readable metadata. The extension and application of ontologies for workflow-centric, *in silico* research objects (e.g. Belhajjame et al., 2013) could make our approach even more useful in metadata creation; we hope to explore this extension in future work.

Finally, we note that the process of creating the diagrams and inventories outlined above helps reframe workflows as something one does, as opposed to just something one captures. We found that use of systems analysis techniques in general prompts a consideration often overlooked aspects of research workflows, and helped us identify obscure relationships between data artifacts; we believe that explicitly drawing on the wealth of existing systems analysis literature and methods will be a fruitful direction for future work in data curation. Additionally, we found that taking the time to describe one's workflow in detail prompts reflection on and refinement of the workflow itself. The diagrams we produced effectively function as social or boundary objects (see Carlile, 2002; De Roure, Bechhofer, Goble, & Newman, 2011; Dourish, 2001; Star & Griesemer, 1989), and were used to negotiate work arrangements and develop a shared language and understanding of a project. Intentionally creating these boundary objects will help bridge divides that often occur as information professionals and domain scientists collaborate and work to align their contributions to the research process (Palmer, 2006).

Limitations of the case

The case presented here has several limitations. Firstly, though we believe the Research Process Modeling approach could – and possibly should – be done prospectively, this case is retrospective. That said, retrospective construction gives researchers an opportunity to reflect on their work without the pressure of time constraints or deadlines.

Secondly, our case may be limited by its focus on work conducted in a United States National Park. The administrative and organizational structures that govern work in YNP have shaped the Geobiologist's workflow in some distinct ways that would not necessarily translate to other sites; for instance, a considerable amount of his workflow is organized around NPS permitting and reporting requirements. However, many scientifically significant sites are managed by an administrative organization with similar permitting and reporting processes. Research Process Modeling throughout a project's duration could help information professionals leverage administrative structures to their advantage by giving them a means to integrate data management planning with external reporting requirements.

Future work

This work points to a further need for standardized application of process, workflow and provenance terminology and notation. Here we have drawn on existing standards as much as possible, notably systems analysis-based diagramming notations like UML and the PROV specification. However, many process documentation methods are not as explicit in their notation. This could be an impediment to interoperability and interpretability of Research Process Models. We hope to collaborate with others in the workflow capture community to discuss how, and how much, to standardize descriptions of research workflows. As noted above, the Research Object ontology may be a particularly suitable candidate for extension or adaptation, though it presently is designed for application to entirely *in silico* workflows.

This work additionally underscores a need for tools that concurrently create activity diagrams, RDF triples, and artifact or process inventories. The inventories and diagrams presented herein were created manually, using simple diagramming software (Google Drawings, draw.io and PowerPoint) and Microsoft Excel. However, data curation tools that could support the simultaneous creation of inventories, activity diagrams, provenance graphs, and other metadata are needed. A PROV authoring tool that streamlines the process of creating both triples and diagrams simultaneously would be a productive development step. The recently developed PROV-TEMPLATE (Moreau et al., 2017) is promising, but further work is needed to create interactive tools.

Finally, further case studies are needed to truly demonstrate the utility of this approach, and to refine its application. An additional case study with paleontology research has recently been completed (Thomer, 2017) and two others are nearing completion: one focusing on a bioinformatics project, and another of field science conducted with small unmanned aircraft systems (sUAS; colloquially referred to as drones). The sUAS community is particularly interested in an approach to documenting both computational and non-computational workflows in a machine-readable way. In all three of these cases we have created, or plan to create, Research Process Models from the beginning of their projects, rather than retrospectively after their conclusion.

Conclusion

Here we have presented a case study of geobiology data curation, through which we develop an approach to Research Process Modeling: a method for documenting a combination of *in silico, ex silico* and field-based research processes and data provenance. This approach supports collaborative data curation between information professionals and researchers, and is therefore a significant step toward producing more useable and interpretable field-based research data.

Building on existing techniques in systems analysis, and recent work curating research objects and data curation lifecycles, the Research Process Modeling method documents field processes in sufficient detail so as to render them retraceable, much like computational workflow capture makes processes re-executable. It adds value to datasets by acting as a potentially lasting representation of a dataset's context of production in a partially machine-readable way. Additionally, Research Process Modeling can help stakeholders understand one another's work, and negotiate an effective division of labor between different researcher team members as well as information professionals.

Acknowledgements

We thank the two anonymous reviewers, whose feedback helped improve and clarify this manuscript. We gratefully acknowledge the contributions to this work made by other SBDC team members: Timothy DiLauro, Jacob Jett, Abigail Asangba, and G. Sayeed Choudhury. This work was funded by IMLS National Leadership Grant LG-06-12-0706-12 and draws upon the insight of data curation and data practices teams initiated by NSF Office of Cyberinfrastructure DataNet award #0830976 for the *Data Conservancy: A Digital Research and Curation Virtual Organization*. The geobiology case study research was supported on grants to B. W. Fouke by the National Science Foundation Biocomplexity in the Environment Coupled Biogeochemical Cycles Program (EAR 0221743), the National Science Foundation Geosciences Postdoctoral Research Fellowship Program (EAR-0000501), the Petroleum Research Fund of the American Chemical Society Starter Grant Program (34549-G2), the University of Illinois Urbana-Champaign Critical Research Initiative, the NASA Astrobiology Institute Cooperative Agreement No. NNA13AA91A issued through the Science Mission Directorate, and TOTAL S. A., France No. FR5585. We wish to highlight the critically important role played by the U.S. National Park Service in Yellowstone National Park for their ongoing collaboration and permission to collect microbe, water and travertine samples at Mammoth Hot Springs (Permit Number YELL-03060). Conclusions in this study are those of the authors and do not necessarily reflect those of the funding or permitting agencies.

References

Aalst, W. van der, & Hee, K. van. (2004). *Workflow management: models, methods and systems* (1. MIT Press paperback ed). Cambridge, Mass.: MIT Press.

Ball, A. (2010). *Review of the State of the Art of the Digital Curation of Research Data* (Vol. 32). Bath, UK: University of Bath.

Ball, A. (2012). *Review of Data Management Lifecycle Models*. University of Bath. Retrieved from http://opus.bath.ac.uk/28587/

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., … Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599–611. https://doi.org/10.1016/j.future.2011.08.004

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. https://doi.org/10.1038/npre.2010.4626.1

Belhajjame, K., Corcho, O., Garijo, D., Zhao, J., Newman, D., Klyne, G., … Roos, M. (2012). Workflow-Centric Research Objects : A First Class Citizen in the Scholarly Discourse. In *Proceedings of Workshop on the Semantic Publishing, (SePublica 2012), 9th Extended Semantic Web Conference* (pp. 1–12). Crete, Greece.

Belhajjame, K., Klyne, G., Garijo, D., Corcho, O., Garcia-Cuesta, E., & Palma, R. (2013, November 30). Wf4Ever Research Object Model 1.0. (S. Soiland-Reyes & S. Bechhofer, Eds.). Retrieved from http://wf4ever.github.io/ro/

Belhajjame, K., Zhao, J., Garijo, D., Hettne, K., Palma, R., Corcho, Ó., … Goble, C. (2014). The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. *arXiv:1401.4307 [Cs]*. Retrieved from http://arxiv.org/abs/1401.4307

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, *7*, 57. https://doi.org/10.1186/s13326-016-0097-6

Carlile, P. R. (2002). A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development. *Organization Science*, *13*(4), 442–455. https://doi.org/10.1287/orsc.13.4.442.2953

CCSDS. (2012). *Reference Model for an Open Archival Information System (OAIS): Recommended Practice* (No. June) (p. 135). Washington, D.C.: The Management Council of the Consultive Committee for Space Data Systems. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology*, *66*(3), 616–633. https://doi.org/10.1002/asi.23184

Corcho, O., Garijo Verdejo, D., Belhajjame, K., Zhao, J., Missier, P., Newman, D., … Goble, C. (2012). Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceedings of Workshop on the Semantic Publishing* (pp. 1–12). Hersonissos, Creta (Grecia): Facultad de Informática (UPM). Retrieved from http://sepublica.mywikipaper.org/sepublica2012.pdf

Cox, S. J. D. (Ed.). (2013). OGC Abstract Specification: Geographic information — Observations and measurements. Open Geospatial Consortium. Retrieved from http://www.opengis.net/doc/is/om/2.0

Cox, S. J. D. (2015). Ontology for observations and sampling features, with alignments to existing models. *Semantic Web*, *8*(3), 453–470.

Cox, S. J. D., & Car, N. J. (2015). PROV and Real Things. In *21st International Congress on Modelling and Simulation* (pp. 620–626). Gold Coast, Australia. Retrieved from http://www.mssanz.org.au/modsim2015/C4/cox.pdf

Cyganiak, R., Wood, D., & Lanthaler, M. (Eds.). (2014, February 25). RDF 1.1 Concepts and Abstract Syntax. W3C. Retrieved from https://www.w3.org/TR/rdf11-concepts/

De Roo, B., De Maeyer, P., & Bourgeois, J. (2016). Information flows as bases for archeology-specific geodata infrastructures: An exploratory study in flanders. *Journal of the Association for Information Science and Technology*, *67*(8), 1928–1942.

De Roure, D., Bechhofer, S., Goble, C., & Newman, D. (2011). Scientific Social Objects: The Social Objects and Multidimensional Network of the myExperiment Website (pp. 1398–1402). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.245

Dennis, A., Wixom, B. H., Tegarden, D. P., & Seeman, E. (2015). *System analysis & design: an object-oriented approach with UML* (Fifth edition). Hoboken, NJ: Wiley.

Dourish, P. (2001). Process descriptions as organisational accounting devices: the dual use of workflow technologies. *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, 52–60. https://doi.org/10.1145/500286.500297

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Faundeen, J. L., Burley, T. E., Carlino, J. A., Govoni, D. L., Henkel, H. S., Holl, S. L., … Zolly, L. S. (2013). The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4. https://doi.org/10.3133/ofr20131265

Fouke, B. W. (2011). Hot-spring Systems Geobiology : abiotic and biotic influences on travertine formation at Mammoth Hot Springs , Yellowstone National Park, USA. *Sedimentology*, (58), 170–219. https://doi.org/10.1111/j.1365-3091.2010.01209.x

Fouke, B. W., & Murphy, T. (2016). *The Art of Yellowstone Science: Mammoth Hot Springs as a Window on the Universe*. Livingston, Montana: Crystal Creek Press.

Goble, C., & De Roure, D. (2009). The impact of workflow tools on data-centric research. In *Data Intensive Computing: The Fourth Paradigm of Scientific Discovery*. Retrieved from https://eprints.soton.ac.uk/267336/

Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, *9*(6), 506–517. https://doi.org/10.1093/bib/bbn034

Hedges, M., Blanke, T., Fabiane, S., Knight, G., & Liao, E. (2012). Sheer Curation of Experiments: Data, Process, Provenance. *Journal of Digital Information*, *13*(1). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/5883

Hettne, K. M., Dharuri, H., Zhao, J., Wolstencroft, K., Belhajjame, K., Soiland-Reyes, S., … Roos, M. (2014). Structuring research methods and data with the research object model: genomics workflows as a case study. *Journal of Biomedical Semantics*, *5*, 41. https://doi.org/10.1186/2041-1480-5-41

Higgins, S. (2008). The DCC curation lifecycle model. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries - JCDL '08* (Vol. 3, p. 453). https://doi.org/10.1145/1378889.1378998

Hills, D. J. (2015). Let's make it easy: A workflow for physical sample metadata rescue. *GeoResJ*, *6*, 1–8. https://doi.org/10.1016/j.grj.2015.02.007

McPhillips, T., Bowers, S., Belhajjame, K., & Ludäscher, B. (2015). Retrospective provenance without a runtime provenance recorder. In *USENIX Workshop on Theory and Practice of Provenance*. Retrieved from https://www.usenix.org/system/files/tapp15-mcphillips.pdf

Mesirov, J. P. (2010). Accessible Reproducible Research. *Science*, *327*(5964), 415–416. https://doi.org/10.1126/science.1179653

Moreau, L., Batlajery, B., Huynh, T. D., Michaelides, D., & Packer, H. (2017). A Templating System to Generate Provenance. *IEEE Transactions on Software Engineering*, *PP*(99), 1–1. https://doi.org/10.1109/TSE.2017.2659745

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., … den Bussche, J. V. (2011). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, *27*(6), 743–756. https://doi.org/10.1016/j.future.2010.07.005

Palmer, C. L. (2006). Weak Information Work and "Doable" Problems in Interdisciplinary Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–16. https://doi.org/10.1002/meet.14504301108

Palmer, C. L., Thomer, A. K., Baker, K. S., Wickett, K. M., Hendrix, C. L., Rodman, A., … Fouke, B. W. (2017). Site-based data curation based on hot spring geobiology. *PLOS ONE*, *12*(3), e0172090. https://doi.org/10.1371/journal.pone.0172090

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

PROV-DM: The PROV Data Model. (2013). (W3C Recommendation). Retrieved from https://www.w3.org/TR/prov-dm/

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology,

1907-39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Stodden, V., Guo, P., & Ma, Z. (2013). Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE*, *8*(6), e67111. https://doi.org/10.1371/journal.pone.0067111

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., … Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246–263. https://doi.org/10.1002/asi.23177

Thomer, A. K. (2017). *Site-based data curation: bridging data collection protocols and curatorial processes at scientifically significant sites* (Doctoral Dissertation). University of Illinois at Urbana-Champaign, Champaign, IL. Retrieved from http://hdl.handle.net/2142/98372

Thomer, A. K., Palmer, C. L., Wickett, K. M., Baker, K. S., Jett, J. G., Dilauro, T., … Choudhury, G. S. (2014). *Data Curation for Geobiology at Yellowstone National Park: Report from Workshop Held April 16-17, 2013* (p. 41). Center for Informatics Research in Science and Scholarship. Retrieved from http://hdl.handle.net/2142/47070

Tilmes, C., Yesha, Y., & Halem, M. (2010). Tracking provenance of earth science data. *Earth Science Informatics*, *3*(1–2), 59–65. https://doi.org/10.1007/s12145-010-0046-3

Vertesi, J., & Dourish, P. (2011). The Value of Data : Considering the Context of Production in Data Economies. *CSCW 2011*, 533–542.

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., … Wooley, J. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the

Biological Collections Ontology and Related Ontologies. *PLoS ONE*, *9*(3), e89606. https://doi.org/10.1371/journal.pone.0089606

Weber, N. M., Baker, K. S., Thomer, A. K., Chao, T. C., & Palmer, C. L. (2013). Value and Context in data use: domain analysis revisited. *Proceedings of the American Society for Information Science and Technology*. https://doi.org/10.1002/meet.14504901168

Williams, R., & Pryor, G. (2009). *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences.* (RIN report). London: Research Information Network & British Library.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing Data Curation Profiles. *International Journal of Digital Curation*, *4*(3), 93–103. https://doi.org/10.2218/ijdc.v4i3.117

Wu, S., Worrall, A., & Stvilia, B. (2016). Exploring data practices of the earthquake engineering community. *iConference 2016 Proceedings*. https://doi.org/https://doi.org/10.9776/16187

Zhao, J., Gomez-Perez, J. M., Belhajjame, K., Klyne, G., Garcia-Cuesta, E., Garrido, A., … Goble, C. (2012). Why workflows break - Understanding and combating decay in Taverna workflows. In *2012 IEEE 8th International Conference on E-Science* (pp. 1–9). https://doi.org/10.1109/eScience.2012.6404482