SITE-BASED DATA CURATION: BRIDGING DATA COLLECTION PROTOCOLS AND CURATORIAL PROCESSES AT SCIENTIFICALLY SIGNIFICANT SITES

BY

ANDREA K. THOMER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Carole L. Palmer, University of Washington, Chair
Professor Michael B. Twidale
Professor Allen Renear
Professor P. Bryan Heidorn, University of Arizona

ABSTRACT

Research conducted at scientifically significant sites produces an abundance of important and highly valuable data. Yet, though sites are logical points for coordinating the curation of these data, their unique needs have been under supported. Previous studies have shown that two principal stakeholder groups – scientific researchers and local resource managers – both need information that is most effectively collected and curated early in research workflows. However, well-designed site-based data curation interventions are necessary to accomplish this. Additionally, further research is needed to understand and align the data curation needs of researchers and resource managers, and to guide coordination of the data collection protocols used by researchers in the field and the data curation processes applied later by resource managers.

This dissertation develops two case studies of research and curation at scientifically significant sites: geobiology at Yellowstone National Park and paleontology at the La Brea Tar Pits. The case studies investigate: What information do different stakeholders value about the natural sites at which they work? How do these values manifest in data collection protocols, curatorial processes, and infrastructures? And how are sometimes conflicting stakeholder priorities mediated through the use and development of shared information infrastructures?

The case studies are developed through interviews with researchers and resource managers, as well as participatory methods to collaboratively develop "minimum information frameworks" – high level models of the information needed by all stakeholders. Approaches from systems analysis are adapted to model data collection and curation workflows, identifying points of curatorial intervention early in the processes of generating and working with data. Additionally, a general information model for site-based data collections is proposed with three classes of information documenting key aspects of the research project, a site's structure, and individual specimens and measurements.

This research contributes to our understanding of how data from scientifically significant sites can be aggregated, integrated and reused over the long term, and how both researcher and resource manager needs can be reflected and supported during information modeling, workflow

documentation and the development of data infrastructure policy. It contributes prototypes of minimal information frameworks for both sites, as well as a general model that can serve as the basis for later site-based standards and infrastructure development.

To Ms. Donato, Ms. Berge, Mr. Joliffe, Mrs. Kudron (née Baitx), Mr. Rupp, Mr. Kopacki, Professor Hoffman, Professor Grossman, and all my other teachers.

ACKNOWLEDGMENTS

Much of this work was funded through the Site-Based Data Curation Project (IMLS National Leadership Grant LG-06-12-0706-12). SBDC team members include Carole L. Palmer (Principal Investigator), Karen S. Baker, Karen M. Wickett, Jacob G. Jett, Tim DiLauro, Abigail Asangba, Sean Gordon, G. Sayeed Choudhury and Bruce Fouke. Portions of this dissertation are modified from or informed by the following:

Thomer, A. K., Wickett, K. M., Baker, K. S., Fouke, B. W., Palmer, C. L. (in revision). Research process modeling for curatorial intervention. Submitted to JASIST.

Palmer, C.L, Thomer, A. K., Baker, K. S., Wickett, K. M., Hendrix, C. L., Rodman, A., Guenther, S., 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

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. http://hdl.handle.net/2142/47070

This dissertation would not have been possible without the participation of the staff at the Yellowstone National Park Research Permitting Office; the staff at the La Brea Tar Pits Museum; the many researchers that shared their time and insight with me; the students in Bruce Fouke's Fall 2016 Introduction to Biocomplexity course; and Bruce Fouke himself, who provided crucial support and guidance in developing the Yellowstone geobiology case study.

For this particular document: thanks to Karen Wickett for collaboration in developing information models; David Dubin and the Conceptual Foundations Group for conversations about Coombs; Nicholas Weber for feedback and writing playlists; Rebecca Crist for herculean copy edits; and Katrina Fenlon for writing support both grammatical and emotional. Thanks also to the Kelly-Weyerhauser family for feeding me approximately ¼ of my meals over the last six months, and to the GSLIS/iSchool help desk for technical support and occasional grief counseling (RIP CIRSSLAP8).

To my friends at Illinois: In 2010, I crammed my belongings into and on top of a 1992 Honda Accord and traded the tar pits for a little apartment on the prairie. I meant to stay for two years; that somehow turned into seven. I thought I would be desperate to get back to California; I am now so incredibly sad to leave Champaign. To the dear, lovely people that kept me here: this is all your fault. Thank you for your collegiality, creativity, hallway conversations, Symposiums, travel support, good advice, snack-sharingness, Barbarian Sundays, and general willingness to go along with the joke. I couldn't (and wouldn't want to) have done this without you.

To my wonderful committee: I cannot thank you enough for your mentorship, guidance and unflagging support throughout both this dissertation and my graduate studies overall.

And finally, endless gratitude to my incredible family for their encouragement, patience, care packages, and love throughout my graduate work.

TABLE OF CONTENTS

1. INTRODUCTION

Scientifically significant sites are localities that have attracted enough attention from researchers to merit government administration and/or protection, as well as the preservation of associated specimen and/or data collections. Prior work has shown that these sites are logical and potentially efficient points for data curation (e.g. Karasti & Baker, 2008; Millerand & Baker, 2010; Palmer et al., 2013; Thomer, Palmer, et al., 2014). However, they have nevertheless been understudied in research in library and information science, which has largely focused on the curation interests of institutions, disciplines, and data archives, or on the needs of individual researchers working at academic institutions. Work is needed to further understand and support the unique data curation needs and priorities of the resource managers and researchers working at scientifically significant sites.

Resource managers and *researchers* at scientifically significant sites represent the two key stakeholder groups that produce, use, and manage data. These two groups have distinct priorities, which must be carefully aligned. Resource managers (those who manage the natural, human and information resources at a site; e.g. park rangers, research permitting officers, museum curators and collections managers, and library and archives personnel) are concerned with the maintenance of their sites and collections, and need information that helps them plan and facilitate access to sites for a broad range of individuals. Researchers (scientists who are not employed by a site, but visit to collect data), on the other hand, are more focused on collecting data for their individual, idiosyncratic projects (Thomer, Palmer, et al., 2014).

Work conducted through the Site-Based Data Curation Project has shown that despite their differing priorities, both researchers and resource managers have some common needs of the data collected at their study sites: they both need information[1] about key aspects of the natural

[1] This information is often described as "metadata." However, throughout this project I've found it necessary to sometimes describe it more generally as "information," especially when speaking with my study participants. The reasons for this change in vocabulary are discussed further in Chapters 4 and 6.

phenomena and structure of their sites, and about the methods used to collect data (see Palmer et al., 2013, 2017; Thomer, Palmer, et al., 2014; stakeholder needs are discussed further in Chapter 4). However, despite the importance of this contextualizing information, it is often not collected or shared in a consistent manner – if at all. There is consequently a need to reconcile data reporting and curatorial processes with data collection methods. This may require moving away from a conceptualization of metadata (and data, for that matter) as static records or products. Instead, they may be better thought of as a form of scientific communication created through iterative, complex, and often friction-riddled processes (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011). Supporting the creation of effective metadata requires first understanding and supporting the on-going processes used to create it.

In site-based research, the processes that create data and metadata are rooted in scientists' *data collection protocols*. Data collection protocols enumerate what observations or samples ought to be recorded or collected and in what manner. These protocols can be thought of as a kind of proto-information model*:* a scientist's representation of an aspect of the world that will eventually structure the data products that result from field work. For example: a field biologist seeking to understand the biodiversity of a region would likely collect information about where different organisms are found at what time, and in what abundance. Those four factors (what, how much, where and when) would act as a fundamental information model that would guide the structure of her sampling schedule and data collection. This protocol would be further shaped by practical and conceptual factors, such as:

- the real-world limits of what data could be collected in a time- and access-limited field trip,
- the existing theory informing a researcher's understanding of a phenomenon,
- and the specific hypotheses a researcher is seeking to test.

A biodiversity project taking place over three months would have a different sampling schedule – and a different data structure – than one taking place over a weekend. Thus, protocols structure and are structured by more than just spreadsheets; they also structure and are structured by the processes involved in collecting data.

Idiosyncratic data collecting protocols create heterogeneously structured datasets. This heterogeneity is not problematic in and of itself, but it makes it difficult to analyze or curate multiple datasets from a site as an aggregate, despite coming from the same underlying phenomena or study site. The barriers to reuse resulting from heterogeneous data structures are particularly problematic for sensitive field sites, where data and specimens can only be collected in limited quantities; or where phenomena are fleeting and can only be documented once; or in which the site is excavated or otherwise destroyed in the process of data collection. In these instances, already existing data *must* be reused because they document sensitive or ephemeral conditions and cannot be collected again. A site's resource managers thus have a vested interest in encouraging the creation of robust data collections beyond their own administrative needs; well-documented, reusable data are a resource in need of protection as much as any the physical, natural phenomena at a site. Site-based *curatorial processes* must be put into place to ensure that high value data are captured and documented as effectively as possible.

Site-based data collection protocols and curatorial processes, then, must strike a balance between the immediate constraints of fieldwork (e.g. limited time and resources available to collect and curate data) versus the long-term needs of resource managers invested in protecting the site and creating long-lasting data collections. Site-based data collection protocols and curatorial processes must also balance the imposition of standardized data collecting methods on a researcher against the value of the structured data and metadata products they guide into creation. Consequently, the often-idiosyncratic data practices of researchers must be aligned with the more standardized normative modeling constructs demanded by curatorial workflows and infrastructures.

In the natural history collecting and curation traditions, curators and researchers have long depended on a kind of site-based curatorial tradition to organize and contextualize data in a relatively standardized, yet broadly usable way. For instance, researchers augment their structured data with less structured, qualitative field notes describing their collecting methods and site conditions. Natural history museum collections are also often organized according to their specimens' locality of origin. These practices have been effective for long-term curation of specimen collections useful for broad ranges of researchers and applications, and could inform a site-based data curation framework. However, some practices are unique to the physical and

organizational structure of a physical museum, and additional work is needed to understand how physical sample collecting and curatorial practices might inform digital curation.

1.1 STUDY OVERVIEW

In this dissertation, I examine how independent researchers' and site resource managers' needs of site-based data collections differ. I further show how researchers and resource managers bring their data collection protocols and curatorial processes into alignment through the use of contextualizing data points describing unique natural features of their study sites, and how representation of a site can be used to mitigate the structural differences of idiosyncratic data collection methods. I also discuss how we as information science researchers can work with site stakeholders to develop and encourage the use of site-based information frameworks that balance the needs of researchers conducting idiosyncratic research projects at a site, and of resource managers caring for or managing a site over time.

1.2 RESEARCH QUESTIONS

I answer the following research questions. The first relates to different stakeholder perspectives on "sites."

1) What aspects of a site are most important from the perspective of researchers? What aspects of a site are most important from the perspective of curators? And how do these differ amongst and between sites?

The second question relates to "sites" in the process of data collection.

2) How do the data collection protocols followed by researchers represent study sites? How do curation protocols followed by natural history museum curators represent sites?

The third question relates to "sites" in the products of data collection.

3) How do these protocols and practices appear in the information models associated with datasets created by researchers and curators?

The last relates to how to accommodate the practices and interests of researchers and curators

4) How can researchers and curators develop data collection protocols that balance their respective needs of data from a scientifically significant site?
    a. What aspects of traditional natural history work practices might inform site-based standards development?

To address these questions, I develop two case studies of data collection and curation practices at scientifically significant sites:

Case 1: Geobiology research at Yellowstone National Park: The hot springs at Yellowstone National Park (YNP) are home to a diverse range of thermophilic microbes that shape the structure and behavior of the springs over time, and are consequently an important study site for the growing field of geobiology (the study the interaction between microbes and the sedimentary environments they live in and create). Hundreds of geobiology researchers conduct studies at YNP every year, and while they sometimes collaborate through research coordination networks and other organizations, they do not share or publish much of their data, and subsequently face many barriers to data integration. This case was initially developed through my work as a research assistant on the IMLS-funded Site-Based Data Curation Project (Palmer et al., 2017).

Case 2: Paleontology research and curation at The La Brea Tar Pits: The La Brea Tar Pits are an incredibly rich ice age fossil locality in the heart of Los Angeles, CA; the asphaltic deposits have produced an estimated 3-4 million fossils, ranging from microscopic ostracods (a kind of crustacean) to 6-foot-long Columbian mammoth tusks, and are an important site for paleontology research. Since 1969, museum staff have been using roughly the same data collection protocol in their excavations: an incredibly detailed, grid-based measurement system, through which multiple coordinates are recorded for every specimen over ¼" in size. While this data theoretically makes it possible to reconstruct the position of every single fossil, and could serve as a powerful backbone for data integration, there are not (and have not been) any computer programs built for this task. Thus, the location data is largely unused. Current collections managers are eager to explore alternative excavation protocols.

These case studies are developed through interviews and participatory engagement with stakeholders at each of my sites. I additionally draw on methods from systems analysis to model and analyze data collection and curation workflows.

1.3 STUDY CONTRIBUTIONS

This dissertation contributes to our understanding of how data from scientifically significant sites can be aggregated, integrated and reused over the long term, and how potentially conflicting stakeholder needs can be aligned, reflected and supported during the work of information modeling, workflow documentation and data reporting policy development. It additionally contributes prototypes of data standards called, "minimal information frameworks" for both study sites, as well as a general information model that may be applicable to site-based data curation at other sites.

2. BACKGROUND

2.1 TOWARDS SITE-BASED DATA CURATION

As briefly described in Chapter 1, scientifically significant sites such as Yellowstone National Park and the La Brea Tar Pits are important loci for data curation, management and preservation. There has been some prior work towards the development of data curation best practices at scientifically significant sites by researchers at sites such as the Long Term Ecological Research (LTER) network (e.g. Karasti, Baker, & Halkola, 2006) and at biological field stations (e.g. Brunt & Michener, 2009). However, these sites and best practices are primarily meant to support long-term monitoring projects, rather than the "small science" work conducted (after Cragin, Palmer, Carlson, & Witt's use of the term (2010)) at sites such as national parks. Additionally, monitoring networks and field stations do not function with the same constraints and resource management responsibilities as national parks and other sites that need to coordinate work done by external researchers or enforce permitting and other means of oversight. Thus, data curation best practices are needed that accommodate the special organizational considerations of sites while improving access and reuse of the highly valuable data produced at the sites over time.

The "small science" research supported by scientifically significant sites is not in the least small in its scope and impact. By some measures "small science" research projects make up 80% of all science (Heidorn, 2008) and is expected to produce more data over time than "big science" (Carlson, 2006). But because of the complexity of its practice and culture, small science is very poorly served by curation and repository services (Cragin et al., 2010). Researchers tend to work independently or in small groups, on hypothesis-driven questions, keeping their data private for local analysis. Their communities are heterogeneous in their methods and the types of data they produce and use; standards are not always applied; and cultures of data sharing are still developing (see for example Borgman, Wallis, & Enyedy, 2007 on data sharing in habitat ecology; Delson, Harcourt-Smith, Frost, & Norris, 2007 on paleoanthropology; Douglass, Allard, Tenopir, Wu, & Frame, 2014 on government employees; Hampton et al., 2013 on ecology; Mounce, 2014 on paleontology; Wallis, Rolando, & Borgman, 2013 on sensing networks; and Wieczorek et al., 2012 on biodiversity data). That said, small science has much to gain from systematic curation and a tremendous amount to contribute to integrative, cross-disciplinary

research driving the move toward national and global networked data (Hey, 2009; National Science and Technology Council, 2009; National Science Board, 2005).

Numerous researchers have studied the data practices (Cragin, Chao, & Palmer, 2011; Cragin et al., 2010; Palmer & Cragin, 2008) of small science domains that conduct work at scientifically significant sites, though much of this research has focused on identifying barriers to data sharing (e.g. Borgman, 2012; Kowalczyk & Shankar, 2011; Wallis et al., 2013 as well as references above), or reuse (e.g. (Darch, 2014; Faniel & Jacobsen, 2010; Frank, Kriesberg, & Yakel, 2015; Palmer, Weber, & Cragin, 2011; Weber, Baker, Thomer, Chao, & Palmer, 2013; Zimmerman, 2008)[2] rather than on data collecting practices and how these might inform the work of resource managers. That said, several studies have additionally focused on the relationship between data collection and curation processes and data products, and are therefore particularly relevant to this work. For instance, Mayernik et al. discuss how researchers at the Center for Embedded Networked Sensing (CENS) "un-earthed" their data collection procedures to incorporate new sensing technologies (Mayernik, Wallis, & Borgman, 2013), noting that making data collecting processes explicit was critical to their ability to re-engineer them. Edwards et al., in their discussion of data and metadata "friction", argue that metadata can be understood as an "ephemeral process of scientific communication, rather than as an enduring outcome or product" (pg. 667, 2011).

2.2 DATA AND METADATA AS PRODUCT: STANDARDS EFFORTS RELEVANT TO SITE-BASED DATA CURATION

Before discussing work relevant to Edwards et al.'s conception of metadata-as-process, it is important to first review the wealth of prior work on metadata-as-product. The creation and maintenance of metadata has been a central concern in data curation research because structured descriptive information is essential to facilitating data discovery in information systems and is

[2] There several other articles that discuss these topics that were not included in this section because they do not take a "data practices" perspective, or do not focus on site-based sciences. For instance, Vertesi and Dourish (2011) discuss differences in data sharing practices between different lab groups at the Jet Propulsion Laboratory, but astronomy data stretches our definition of scientifically significant sites in an unhelpful way; thus this paper was not included.

needed by users to understand the potential and fitness for data to be reused. However, LIS and domain researchers alike have struggled to develop ways of creating metadata efficiently and without undue burden on data creators. In this subsection, I briefly review relevant standards initiatives, focusing particularly on communities that work at scientifically significant sites.

At the most basic level, metadata should capture the "who, what, where, when, why and how" of a dataset's collection. Additionally, detailed information about data collection methods, units of measurement, and variable definitions is necessary for data to be usable by anyone beyond the original collector (Fegraus, Andelman, Jones, & Schildhauer, 2005).[3] However, there are numerous obstacles to creating this documentation. Many researchers are not aware of or do not implement existing standards for describing data collection methods (Tenopir et al., 2015), in some cases because of the standards' complexity (Mayernik, Batcheller, & Borgman, 2011), and in others, because of their insufficiency (Chao, 2014). Additionally, many researchers regard the creation of metadata – methodological or otherwise – as something to be done long after research is concluded. Consequently, many important details about early phases of research, especially data collection, are in danger of being lost simply through the passage of time and the fading of memory. Embedding data curation interventions "upstream" or earlier in research workflows has consequently been of on-going concern in data curation research.

Many research communities have developed their own data and metadata standards and data publication or aggregation infrastructures to support easier publication and reuse their data. However, these too tend to lack critical contextualizing information. The Protein Data Bank and GenBank are acknowledged as canonical models of success in the curation of large aggregations of shared scientific data (Berman, Henrick, Nakamura, & Markley, 2007; Howe et al., 2008), yet nevertheless lack critically important contextualizing site-based data. Work on the SBDC project showed that the lack of contextualizing information about scientifically significant sites in databases such as GenBank impeded reuse of the sequence data by geobiologists (Palmer et al., 2017). Curation of these resources has focused on annotation and assembly of metagenomes and

[3] Text from this and the following paragraph has been modified from Palmer et al., 2017.

genomes. Consequently, only minimal, if any, environmental contextualization of environment is required and published.

The lack of metadata describing scientifically significant sites isn't for lack of standards; there are numerous data standards and data publishing initiatives that capture data and metadata about research localities and data collecting methods. For instance, the EarthChem portal[4] aggregates, stores, and provides data formatting guidelines for geochemistry data; and the EarthChem data templates express much of the critically important site metadata needed by geobiologists and resource managers alike (e.g. Gordon et al., 2014). The various biodiversity data standards developed by the Taxonomic Database Working Group[5] (TDWG), such as Darwin Core (DwC) and the Access to Biological Collections Data (ABCD) standards, include classes for collecting event, locality and even some collecting methods; the Global Biodiversity Information Facility publishes what collecting event data is available (though with some limitations due to historical incompleteness of individual records (Thomer, Baker, Sacchi, & Dubin, 2012). The Ecological Metadata Language (Michener, Brunt, Helly, Kirchner, & Stafford, 1997) developed out of work by the Ecological Society of America and utilized by the LTER network has several data classes for methods and site-based data. The National Environmental Methods Index (NEMI)[6] is a database of analytical and field methods for environmental monitoring methods and protocols – essentially an extensive controlled vocabulary of environmental field research methods. And finally, Biosharing.org, a resource focused on life sciences data standards, databases, and policies, includes over 600 standards,[7] including over 80 "Reporting Guidelines," largely drawn from an initiative working to harmonize minimum information standards, called Minimum Information about a Biomedical or Biological Investigation (Taylor et al., 2008).

In many cases, this site and collecting event metadata *is* collected by researchers; it's just not published or published in sufficient richness and detail (Chao, 2015) . This is a known issue with natural history collections: details about data collecting trips are often described in prose

[4] http://www.earthchem.org/
[5] www.tdwg.org
[6] https://www.nemi.gov/home/
[7] https://www.biosharing.org/standards/

narrative in field notebooks, but not included as part of a structured catalog, and therefore must be migrated by hand (Thomer, Vaidya, Guralnick, Bloom, & Russell, 2012; Tulig, Tarnowsky, Bevans, Kirchgessner, & Thiers, 2012). This isn't to say that field methods are without standards or best practices. Rather, there's an immense diversity of field methods, which can be difficult to formalize into a standard format or publication method. Many, such as field notebooks, are necessarily qualitative and unstructured; others are uniquely tailored to the particular needs of different sampling efforts and studies (the USGS, for instance, lists 142 different field sampling methods at the time of this writing;[8] NEMI lists over 200).

Field notebooks are a common and historically important method of recording site metadata and describing collecting methods in many sciences, and thus deserve some special consideration in a study of site-based data collecting practices and protocols. Field notebooks typically contain handwritten narratives describing each day's tasks, sampling locations and methods, sketches and maps, itemized specimen/sample lists, and so on. Even as modern field-based sciences become less descriptive and more quantitative in their approaches – and even as digital photography has made it relatively simple to rapidly and cheaply document a site through images – keeping a field notebook is still considered a fundamental best practice of fieldwork (Greene, 2011; Mogk, n.d.), and field notebooks are a source of much of the site-based data and metadata.

2.3 DATA AND METADATA AS PROCESS: PROTOCOLS, PLANS AND INFORMATION STRUCTURES

Where work on metadata and data standards focuses on data and metadata-as-product, work in the fields of Computer-Supported Cooperative Work, systems analysis and on the conceptual foundations of information modeling have focused more on the processual nature of metadata. This research is important to consider, given this dissertation's focus on the alignment of data and metadata standards as used in data collection protocols and curatorial processes. Data collection protocols are as much enacted as they are written: they are both plans and data structures. They are plans in that they lay out a future doing of work, and therefore shape how

[8] http://www.usgs.gov/science/science.php?thcode=2&term=379

work arrangements will happen (see Steinhardt & Jackson's work on anticipation work and plans and scientific work (2014, 2015)). Yet they are also data structures in that they literally act as data schemas and dictate how data and information will be structured and stored. Applying a "standard" to one's data collecting means not just changing data structure and content, but also one's plans for work. Thus, data collecting protocols can be visualized and understood through two methods of systems analysis: information modeling and process modeling. In this subsection, I briefly describe these modeling approaches, both of which inform my choice of methods (outlined in Chapter 3).

Information and data modeling are fundamental to computer and information science. They are the mechanism by which we represent information in a sufficiently structured manner that a computer can understand, and map the relationships between, the roles and structures of objects and data (Kent, 1978). Taking the implications of data modeling seriously will be foundational to future cyberinfrastructure and repository development; further, information modeling as an analytic technique can (and has) illuminate critical problems or quirks in data systems that must be addressed in cyberinfrastructure development (Sacchi & Wickett, 2012; see Renear & Dubin, 2008; Renear, Sacchi, & Wickett, 2010; Wickett, Renear, & Furner, 2011; Wickett, Sacchi, Dubin, & Renear, 2012 for examples).

Where information modeling maps objects and relationships, process modeling creates a representation of how a system operates. In business systems analysis, the "model" consists of multiple components, such as an activity diagram (a flowchart illustrating different steps of work) supplemented by use case descriptions and diagrams (Dennis, Wixom, Tegarden, & Seeman, 2015). Systems analytic techniques such as process modeling have been used by software engineers (e.g. Curtis, Kellner, & Over, 1992), and LIS practitioners, as a way of linking human processes with information flows. In LIS, they are predominantly found in the form of “lifecycle” models: broad, overarching, idealized workflows meant to map and visualize a data point’s path from collection to use to curation (e.g. the OAIS model (CCSDS, 2012); the DCC model (Higgins, 2008)). Though not explicitly described as process models, they nevertheless are generalized depictions of data curation processes. These frameworks tend to focus on curatorial activities that take place in repositories and research centers long after the initial point of data collection. Consequently, they have been critiqued for,

> [hiding] a great deal of the complexity that exists in real-world research; while the OAIS and DCC Curation Lifecycle models flesh out the detail of the archival stage in the data lifecycle, neither these nor the other models above provide a similar amount of detail when modeling the early stages of the lifecycle. (Ball, 2010).

Other examples of process modeling in LIS include use of "information flow maps" created through the Research Information Network case studies project; they use systems analysis to document research workflows in with the intent of re-engineering workflows to push "data curation upstream" – that is, to give researchers the tools and methods they need to document and curate their data before it is handed over to a repository (Williams & Pryor, 2009). However, this method has not been broadly adopted.

2.4 CONCEPTUAL FOUNDATIONS

This dissertation seeks to build on the literature reviewed above but is particularly rooted in two prior works: Clyde H. Coombs' *A Theory of Data* (1964), as well as Bruno J. Strasser's definition and explication of natural history collecting practices. In this subsection, I describe both of these bodies of work and their applicability to this study.

2.4.1 COOMBS' THEORY OF DATA

Coombs, a mathematical psychologist, developed a theory of data through "an analysis of the foundations of psychological measurement." This theory is meant to account for ways in which a researcher's epistemological perspective, study design, and interpretations impact and shape a dataset's content and structure at different points within the research process. Though developed via analysis of the work of psychology, it has application to many other fields of science, particularly those dependent on the collection of field observations or specimens.

Coombs defines three phases during which a dataset is shaped by a scientist's interpretation and method of recording (Figure 2.1):

- Phase 1: data collection, in which the researcher decides which of the universe of potential observations he will collect
- Phase 2: the first phase of interpretation, in which the researcher transforms observations into data through interpretation (note that observations are not data in Coombs' view until they have been interpreted). Phase 2 "involves a classification of observations in that individuals and stimuli are identified and labeled" (pg. 5).

- Phase 3: The second phase of interpretation, in which the researcher detects relations, order and structure, following the data and the model used for analysis.

Coombs theory emphasizes that researchers select – rather than construct or indiscriminately collect – their data from an external "universe of potential observations." He argues that scientists enter each phase "in a creative way" (pg. 5) and that their decisions at each phase have impact for subsequent phases.

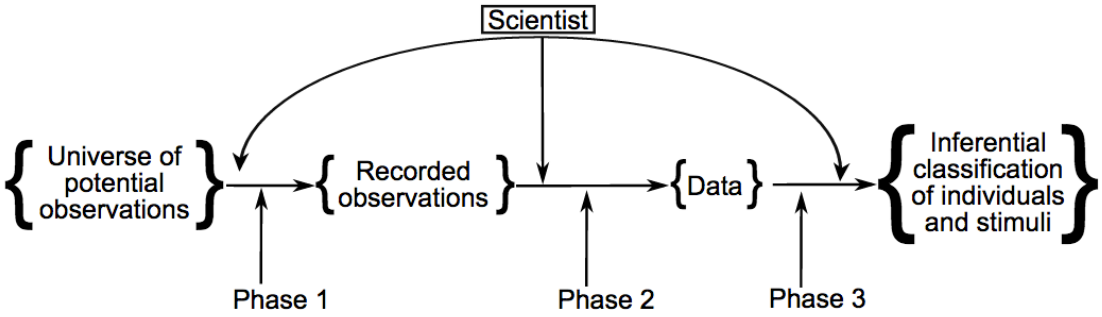


Figure 2.1. From Coombs 1964. Figure adapted by David Dubin.

Coombs' model of data collection and analysis processes emphasizes the role that scientists' decisions and actions play in their data collection, interpretations and conclusions, while also acknowledging that there is an external material reality from which to derive observations. Thus, Coombs' theory provides a helpful bridge between the post-positivism of the scientists I work with and theories of the social construction of knowledge. I use Coombs' theory of data as a lens through which to analyze the data collecting practices in each of my case studies.

2.4.2 Strasser's natural history and the "Experimenter's Museum"

"Natural history" is a broad term with a long history; in this dissertation, I draw on science historian Bruno Strasser's definition and framing of it:

> By "natural history," I do not refer to the study of whole organisms—a recent meaning of the term—but to the different practices of collecting, describing, naming, comparing, and organizing natural objects, practices usually associated not with the laboratory but with the wonder cabinet, the botanical garden, or the natural history museum. Indeed, if there is any distinctive to the natural history approach, it is its reliance on collections, which

> have played a crucial role in natural history from the early modern period to the late nineteenth century, when Victorian sensibilities brought such collections to widespread popularity

This passage is from the introduction to Strasser's paper "The Experimenter's Museum" (2011). In it, Strasser highlights the role that natural historical collecting practices and "ways of knowing" (pg. 93) played in the creation of GenBank, a critically important public database of molecular sequence data.

Strasser argues that while sequence databases are the product of laboratory work, they are also nevertheless a product of natural history collecting tradition, in that they "[rest] on the collection and comparison of natural facts, often across many species" (pg. 93). In his view, allegedly recent modes of scientific inquiry such as “Big Data” are not new to natural historians, who have been mitigating an information explosion since renaissance explorers started returning from the New World with literal boatloads of new discoveries to catalog and organize (Strasser, 2012). The natural histories (biology, geology and their subdisciplines) were developed to organize, manage and make sense of that data; the system of biological nomenclature is fundamentally a method of information storage and retrieval; the science of geology is rooted (metaphorically, anyway) in efforts to map and classify different strata of rock. Even paleontology, a discipline stereotyped as being fundamentally descriptive and specimen-focused, might be considered data-driven in its reliance on fossil catalogs, statistical analyses and data visualizations long before computers (Sepkoski, 2012). Thus, the natural histories are as much about data wrangling as they are about collecting. The tools for this task may have changed, but many of the fundamental practices and goals have not.

Strasser also argues that while digital databases and datasets are often regarded as the *products* of scientific knowledge, they, "like earlier natural history collections, are not mere repositories; they are tools for producing knowledge" (2011, pg. 63). Collections are not just objects; they are infrastructures (National Science and Technology Council, Interagency Working Group on Scientific Collections, 2009). By placing modern biomedical sample databases on the same continuum as physical specimen collections, he draws important parallels between historical and modern collecting practices, which are necessary to understand modern data practices in a historically situated manner. As Strasser writes, “Today, *in silico* biology complements *in vivo* and *in vitro* approaches, and it is vital to the success of the experimental enterprise" (pg. 63). It

follows that approaches to *in silico* natural history data collection and curation must complement *in vivo* approaches as well.

A natural historical framing of site-based sciences – even those that do not aim to create museum collections– relates them to one another by merit of a) their data collecting practices in the field, and b) their reliance on collections as tools for knowledge production – rather than as mere repositories of already produced knowledge. Furthermore, resituating the datasets that result from field work as part of the natural history tradition of work practices may inform efforts to build infrastructure to support them, and may help bridge an artificial gap between studies of physical specimen curation and digital data curation.

3. STUDY DESIGN

3.1 OVERVIEW

This dissertation develops two cases of research and curation occurring at scientifically significant sites: Mammoth Hot Springs at Yellowstone National Park (hereafter referred to as the YNP case) and La Brea Tar Pits and Museum in Los Angeles (hereafter referred to as the La Brea case). Within each case I also develop two embedded subcases: once which describes the work of scientists conducting research at a site, and one which describes the work of curators and resource managers at the sites. Prior work has shown that these two groups have distinct and sometimes opposing needs and practices (Swan & Brown, 2008; Thomer, Palmer, et al., 2014) , and developing them as separate subcases is important to bringing those differences, dependencies, conflicts and overlaps to light.

Additionally, I use participatory methods in each case to develop or revise data collection and curation standards for each site; I refer to these as *minimum information frameworks*. At YNP, participatory engagement was conducted as part of the Site-Based Data Curation (SBDC) project, initially through a workshop hosted at the park, and continued through several years of on-going collaboration with key participants. At La Brea, I draw on a participatory action research approach and work with site stakeholders to prototype a new excavation protocol for potential future use.

I use semi-structured interviews, memos, and data artifacts as evidence in this case study. I also analyze the minimum information frameworks developed with my participants via techniques adapted from systems analysis. The information framework development process, as well as the systems analysis techniques are approaches developed through the SBDC project; I refine these methods through my dissertation work. Specifically, I model my stakeholders' data collection workflows, and compare them to the minimum information frameworks to identify points of curatorial intervention.

In this chapter I describe my study design; the development and applicability of participatory action research to case study research; potential differences in the scope and analysis of each

case, data sources, completed and planned data collection; my plan for analysis; and the limitation of this study.

3.2 METHOD: MULTI-SITE CASE STUDY WITH EMBEDDED SUBCASES

Case study methods are well-suited to studies seeking to describe or explain a phenomenon (e.g. answering "what is happening" or "how or why is something happening?"); studies in which a phenomenon is studied within its real-world context, or in which a phenomenon cannot be easily disambiguated from its context; and evaluation studies, in which some sort of intervention or change is being assessed (Yin, 2012). Thus, a case study method is appropriate given this dissertation's focus on describing and evaluating how data collection protocols and information models are intertwined, and how that interdependency is managed over time.

Case study approaches have been previously used in relevant studies of scientific information practices (Fry, 2006; Vaughan, 1999; Zimmerman, 2007), scientific data curation, sharing and reuse (Cragin et al., 2010; Faniel, Barrera-Gomez, Kriesberg, & Yakel, 2013; Faniel & Jacobsen, 2010; Frank et al., 2015; Hou, Thompson, & Palmer, 2014; Wallis et al., 2013); and research data flows and recordkeeping processes (Bates, Goodale, & Lin, 2014; Hills, 2015; Khoo & Rosenberg, 2015; Star & Griesemer, 1989; Thomer, Baker, et al., 2012; Thomer & Twidale, 2014). This wealth of prior work demonstrates the appropriateness of this method for this study.

Case studies can use a single-case or multiple-case design. A single case is analogous to a single scientific experiment; multiple-case designs, thus, follow a replication logic. Each case in a multiple-case design should be selected so that it "a) predicts similar results (a *literal replication*) or b) predicts contrasting results but for anticipatable reasons (a *theoretical replication)* (Yin 2009 pg 54). Yin further notes that this logic is distinct from a "sampling" design, in which points are selected according to their representativeness of a whole. Multiple cases therefore do not represent multiple data points within the same experiment; rather, they represent multiple "runs" of the same experiment.

Results from multiple case study designs are often considered more robust than single case studies. They offer the possibility of direct replication of results, or of valuable contrasting results which can inform future work and redevelopment of theory (Yin 2009). This dissertation takes a multiple case study approach to compare and better understand the needs of sites

depending on their curatorial and administrative structures, the kinds of research being done at them, or their underlying natural structures.

3.2.1 STUDY DESIGN COMPONENTS

Yin outlines five key components to case study research design: research questions; propositions; units of analysis; how propositions and data will be linked; and the criteria for analysis.

In this section, I outline my research questions, propositions and units of analysis. I then discuss my data collection plan in section 4.3. I discuss how I will link my propositions to my data, as well as my criteria and plan for analysis in section 4.4.

Research questions and propositions

As described in Chapter 1, my research questions are as follows:

1) What aspects of a site are most important from the perspective of researchers? From the perspective of curators?
2) How do the data collection protocols followed by researchers represent study sites? How do curation protocols followed by natural history museum curators represent sites?
3) How do these protocols and practices appear in the information models associated with datasets created by researchers and curators?
4) How can researchers and curators develop data collection protocols that balance their diverse needs for data from a scientifically significant site?
    a. What traditional natural history work practices can contribute to current site-based curation and standards development?

These research questions were developed through work on the YNP Case via the SBDC project, and my prior experience working at La Brea.

Propositions are statements that direct a researcher's attention to specific aspects of a case that should be studied or tested. They are developed through reflection on one's research questions, through prior knowledge of a case, and prior research in related cases or studies (Perry, Sim, & Easterbrook, 2004; Yin, 2009). My prior work, as well as the literature reviewed in Chapter 2, informed my development of the following initial propositions, which are numbered according to the research questions to which they respond:

P1-1: Different stakeholders value the aspects of the site that help them do their jobs. These will differ depending on the natural features of a site and the goals of the stakeholders' work.

P2-1: Scientifically significant sites have unique natural features and idiosyncratic administrative arrangements which need to be accounted for in any data collection and curation protocols, practices and processes at the site.

P2-2: Researchers develop data collection protocols and practices based on their project goals, hypotheses, access to resources, access to a site, and on prior work in their field. These factors lead them to select certain data points for collection over others.

P2-3: Resource managers develop data curation processes and practices based on their curatorial mandates, access to resources, access to a site and to data collected at a site, and best practices at other similar sites. These factors lead them to prioritize curation of different aspects of a dataset over others.

P3-1: Data collection protocols and curatorial processes are often not explicitly described in a dataset's metadata, and this makes datasets more difficult to curate or reuse.

P3-2: Data collection protocols and curatorial processes structure a dataset's underlying data model. These heterogeneous data models impact later attempts to aggregate, curate or reuse data.

P4-1: Researchers and resource managers need different data and/or metadata about a site for their work.

P4-2: Site-based data information frameworks make it possible to manage or mitigate heterogeneous data structures. Specifically, collections organized around aspects of sites' structure may be more usable over time.

P4-3: Researchers and resource managers can balance their needs from a site through collaboration on research projects.

P4a-1: There are existing site-based data collection and curation practices, such as the curatorial processes used by museums, community-developed data collection standards used by natural historians, or the databasing practices that led to the development of GenBank and the Protein DataBank that are applicable to site-based data curation processes outside of museums.

These propositions were revised or expanded throughout my data collection and analysis processes; the final versions of them are presented in Chapter 6.

Units of analysis

In case study research, the case itself is the central unit of analysis; each case may additionally include embedded subcases. Part of the challenge of case study research design is clearly delineating the boundaries of a case and subcases: though the case itself is, "generally a bounded entity (a person, organization, behavioral condition, event, or other social phenomenon)," the boundary between the case and its context can be difficult to identify (Yin 2012, pg. 7). Though boundaries and case definitions can shift throughout a study, major reconfiguration of case structure entail major re-configurations of research questions, propositions and data collection methods, and thus, should be avoided if possible.

For this study, the boundaries of each case are rooted in the concept of a "scientifically significant site," an entity central to the SBDC project. The term is meant to distinguish and draw attention to sites that merit special administrative oversight because of their scientific importance, and whose data merit special, long-term curation and preservation for both scholarly research and resource management.[9] While it is true that any natural site has potential to be scientifically significant depending on one's scientific interests, those that are specifically preserved or administered as long-term research sites for the public good require special consideration and curation work. Thus, a well-studied research locality (e.g. a hot spring at YNP) with associated data, specimen and archival collections, managed as a scientific resource by a formal administrative body, would be considered a scientifically significant site; a formally-managed but not well- or frequently- studied locality (e.g. unincorporated sections of land managed by the Bureau of Land Management) might not be. Similarly, localities that are managed primarily for recreation (vs. research) purposes (e.g. Kickapoo State Park) would not be considered scientifically significant sites because they lack a critical mass of associated scientific research products, and administration aimed to support scientific research. A site could of course

[9] This definition of a scientifically significant site is revisited in Chapter 6

become more or less scientifically significant over time; however, for this work, I am interested in well-established (e.g. decades-long administrative structures and data collections) sites.

As noted above, cases can be hard to distinguish from their context, particularly for "conceptual" entities that are not strictly defined through physical or organizational boundaries. Indeed, much of the "blurriness" between each case of a scientifically significant site and its context is rooted in the inevitable disparity between the administrative bounds of a site and its broader geographic, ecological or geological extent. The hot springs in Yellowstone National Park are not constrained to the Mammoth area, let alone the park as a whole – there are springs outside of the parks' boundaries, and the geological structures that underlie and feed the springs extend for hundreds of miles in all directions. Further, the fossil deposits at La Brea are not constrained to the section of Hancock Park that the Page Museum manages[10]; in fact, the current active excavation project at the Page is of 23 deposits that were moved *en bloc* from the lot of land just west of the Page's jurisdiction (described further in Chapter 4). In these cases, a site's administrators must manage and navigate the blurriness between the land and data they control and manage, versus the land and data that is relevant to what they control and manage. Researchers also need to be aware of ambiguous boundaries between a natural site and its administrative boundaries, in order to understand how their data collection might be limited or constrained. Thus, this blurriness between case and context is a feature of my study design, rather than a bug: a case study method is particularly well suited to exploring the ways that researchers and curators navigate because of its tolerance for ambiguous boundaries.

Several of my research questions focus on the interaction between independent researchers working at a site intermittently and the resource managers and curatorial staff working at a site over the long-term. Curators and researchers were identified as two distinct classes of stakeholders through the SBDC project (Thomer, Palmer, et al., 2014). Thus, my study will consider these groups as two distinct embedded subcases within each case.

---

[10] As one of the staff members at La Brea put it, "No one told the animals to stop dying west of Ogden."

3.3 DATA COLLECTION

As outlined in the introduction to this chapter, this study uses semi-structured interviews, memos, field notes and data artifacts as evidence. Additionally, my analysis is informed by the process of developing minimum information frameworks at each of my site: models of the core data and metadata elements needed to foster reuse and effective resource management at each site. I additionally model the data collection workflows at each of my sites using methods from systems analysis.

This data collection and creation strategy is rooted in prior experience on the SBDC project, in which we collaborated with key site stakeholders and developed a framework for site-based data curation at YNP. At YNP, we used several forms of participatory engagement to collect data and develop the minimum information framework; however at La Brea, I draw on methods from participatory action research. Thus, because of the differences in scale, scope, and duration of these two cases, they each have different data collection schedules and structures. Below I describe my data collection in each case, starting with YNP. I then briefly discuss how my work at YNP influenced the data collection plan for the La Brea case, and the reasons for selecting a study design informed by participatory action research. I then outline data collection for the La Brea case.

In both cases, data collection activities are grouped into four themes to facilitate comparison across cases, and to show how these different study designs are nevertheless comparable. The four themes include: interviews and other stakeholder engagement; artifact inventory and workflow modeling; information framework development and comparative workflow analysis; and framework refinement and external comparative analysis. Table 3.1 summarizes data collection activities according to these four themes.

Data collection category	YNP	La Brea
Interviews & other stakeholder engagement	2013 focus group with 9 participants; approx 20 telecons with YNP resource managers; 4 follow up interviews w/researchers; 1 follow up interview with a resource manager.	Interviews with 10 La Brea staff and 12 researchers who use the La Brea collections and data; on-going collaboration with 2 core La Brea staff; 3 site visits
Artifact inventory and workflow modeling	Inventory and systems analysis of Fouke data; workflow modeling of data collection methods	Inventory of La Brea collections data sources; data collection workflow modeling
Information framework development and comparative workflow analysis	Development of minimum information framework for geobiology @ YNP; Deployment of MIF with undergraduate field workers for Fouke's CHP 395 class; and analysis through comparison to workflow	Development of new excavation protocol; refinement into minimum information model for La Brea; analysis through comparison to workflow
Framework refinement and external comparative analysis	Compare MIFs from each case to other relevant standards to heuristically assess external applicability	

Table 3.1. Data collection

3.3.1 DATA COLLECTION AND INITIAL DEVELOPMENT OF RESEARCH APPROACH THROUGH YNP CASE

Data collection for the YNP case was completed through the following activities:

Stakeholder engagement & interviews: As part of the SBDC project, a two-day workshop with nine geobiology researchers and seven YNP personnel held in April 2013 at Yellowstone National Park. This workshop included structured discussions, a data sharing activity in which researchers were asked to describe what they would need to make use of each other's data, and two focus groups on data curation needs and data sharing attitudes. One focus group was with researchers (geologists, geochemists, microbiologists), and the other was with park resource managers (permit managers, museum curators, archivists and librarians). This initial workshop laid the foundation for on-going relationships with our participants for future work and follow up interviews. Workshop discussions were recorded and partially transcribed; focus groups were recorded and fully transcribed. All transcripts were reviewed and coded for common themes by SBDC team members, and findings and key themes were summarized in a workshop report (Thomer, Palmer, et al., 2014); I revisited and recoded these transcripts for this dissertation.

From 2013-2016, I continued to work closely with two members of the YNP resource management staff through approximately 20 teleconferences, a two-day face-to-face meeting in

2014; meetings at the 2014 YNP Biennial Science Conference (Fouke et al., 2014; Thomer, Gordon, et al., 2014) ; and numerous email exchanges. Memos and notes from meetings and emails were used as data sources in this dissertation.

I additionally conducted several follow up interviews with researchers. I interviewed two participants from the researcher group at the workshop in 2014; and I interviewed an additional two geobiologists who did not participate in the workshop in the summer of 2016; and one resource manager who did not participate in the workshop in early 2017[11]. The interviews in 2014 helped refine the minimum information framework for geobiology at YNP, whereas the interviews in 2016 contributed to the verification of its usefulness and validity. These interview transcripts were also used as data sources in this dissertation.

Artifact inventory and workflow modeling: Working closely with SBDC co-PI Bruce Fouke, I led the inventory of over 10 years of his field data. Through this work, I additionally led the development of a method of systems analysis called *research process modeling*, which was used to structure to the inventory. A paper describing this method is under review in the Journal of the Association of Information Science and Technology. Files were inventoried according to their type, series, relationships to other files, field notebooks, and published papers, and their "fit" with community-developed data and metadata standards. We additionally inventoried the research processes that created his files, focusing primarily on those that take place in the field (rather than at the laboratory bench). We "linked" these two inventories through an activity diagram (a kind of flowchart) illustrating his workflow, and a provenance graph showing how the documents were related. This provided us with a more structured way of documenting the complex sociotechnical contexts that created these research products, and will additionally help guide later analysis. I drew on this work and adapted portions of this approach for use in this dissertation; this is described further in Chapter 5.

---

[11]I originally hoped to interview up to 5 additional geobiologists, but had very poor response to my requests for interviews. I also found that the follow up interviews were not as informative as I'd hoped in verifying the efficacy and relevance of the MIF. To ultimately decided to add a second phase of external comparative analysis to my research design to verify the MIF: I compared the MIFs to other related to data standards to assess the presence or absence of different key elements. This approach was, in the end, far more informative than the interviews.

Information framework development and comparative workflow analysis: One of the primary goals of the SBDC project was to develop a framework for the curation and collection of geobiology data at YNP. Immediately following the 2013 stakeholder workshop, co-PI Fouke and a workshop participant began developing what eventually became the Minimum Information Framework (MIF) for Geobiology at YNP. The MIF outlines key data elements that must be collected and preserved for every geobiology field trip at YNP; it is, in other words, the backbone for a recommended community data collection protocol (Palmer et al., 2017). We refined and re-engineered the MIF over from 2014-2016 through the stakeholder engagement and interviews described above. I additionally explored implementation of the MIF in the field, as a field assistant to Fouke's Fall 2016 undergraduate Introduction to Biocomplexity course. In this class, Fouke brings students to YNP to conduct experiments at Mammoth Hot Springs. I tutored students in their use of the framework before and during their field work.

While work with Fouke's class was informative, their experiences in the field didn't necessarily reflect those of a professional scientist. I consequently developed a method *of comparative workflow analysis* for this dissertation, in which I assessed the feasibility and completeness of the MIF by comparing it to the workflow models of Fouke's field processes developed through the workflow analysis described above. This work is described further in Chapter 5.

External comparative analysis: As part of the SBDC project, we compared the MIF to a range of existing data standards, eventually focusing our comparison to the EarthChem Vent Fluids data template. I revisit this work here to supplement my follow-up interviews in providing sources for external comparative analysis. I also compare the fundamental structure of this MIF to two more general site-based standards, the BioCollections Ontology and the Environment Ontology.

3.3.2 Data collection and use of participatory action research in La Brea case

The SBDC project took place over several years, and with the benefit of grant funding and multiple project partners; this of course was not replicable with my second case for the purposes of this dissertation. Instead, my work with La Brea is rooted in participatory action research methodology. This approach allows me to roughly mirror some of my work with the YNP case in a way that is more suitably scaled to a dissertation (and a single investigator), and further, to build on methods developed through SBDC. Participatory action research explicitly prioritizes

"learning through doing"; I will gain insight into existing data collection and curation practices by re-engineering them through methods of research process modeling and information systems analysis prototyped through the SBDC project to more rigorously document the research processes and protocols that create data objects.

Participatory action research: "The best way to understand something is to try to change it."

Where in a straightforward case study, one would focus on describing and analyzing each case as an observer, the participatory action research (PAR) approach, " 'aims to contribute both to the practical concerns of people' in problematic situations and to the academic goals of science 'by joint collaboration with a mutually acceptable ethical framework' " (Rapaport 1970, p. 499, as cited in Hayes, 2011). Thus, I work with my stakeholders to assess and revise their existing data collection and curation protocols.

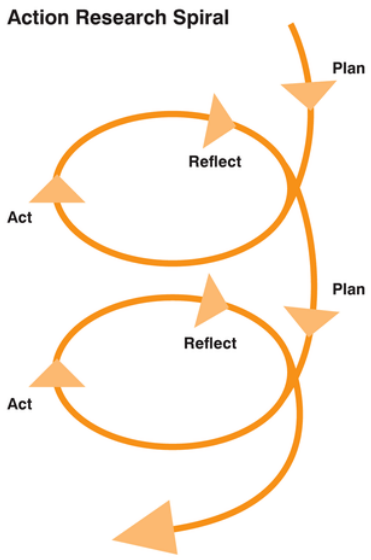


Figure 3.1. The Participatory action research spiral (Hayes 2011).

The PAR approach is rooted in a belief that "the best way to understand something is to try to change it" (Lewin, 1943). As reviewed by Hayes (2011), Lewin's initial formulation of PAR in the 1940s aimed to gain scholarly knowledge (the "research") through the process of effecting change in study populations through training and sustained practice in their home environments (the "action"). Later workers would reframe the process of research-through-action as

participating in an "ongoing dialogue" with the study community (rather than perhaps imposing the researcher's will on a community), and emphasize the importance of sustainable change towards an agreed upon goal, or fulfilling a pre-existing need (Hayes, 2011; Peters & Robinson, 1984). Hayes continues,

> Rather than being distant observers, [researchers] could now be engaged in problem solving alongside their 'research subjects.' Furthermore, judgments of [research] quality would now include the requirement that a workable solution to some real life problem be developed (2011, pg 15:5).

PAR is similar to participatory design (PD) and user-centered design (UCD), in that all three depend on collaborative development of sociotechnical interventions; however, PAR is considered somewhat more scientifically rigorous that UCD and PD, because PAR explicitly aims to generate scholarly knowledge through reflection and analysis rather than solely generating design artifacts or implications for design (Hayes 2011). Thus PAR also has many similarities with straightforward case study research, which is also rooted in reflective analysis.

Though specific implementations of PAR vary according to the discipline using them, the community being studied, and the focus and goals of a specific project (see Cassell & Johnson, 2006 for one review), they're united by a common practice of iterative problem formulation, community engagement, action or intervention, evaluation, and reflection. Where UCD is often depicted as a circle, PAR is depicted as a spiral, wherein work is constantly built on reflection on prior work (Figure 3.1). Further, PAR approaches are united by a post-positivist embrace of the researcher's role as a co-creator of scientific knowledge:

> AR requires scientists to observe their own roles in the process, recognizing and embracing their own influence in the research not as 'contamination' or 'bias' but as an inevitable part of the social construction of scientific knowledge (Hayes 2011, 15:7).

In a traditional PAR approach, part of this iterative process would include reformulating research questions in collaboration with one's study participants. Due to the limited timeframe of this project, I will not be reformulating my research questions. However, I reviewed my research questions with La Brea collections staff members prior to beginning this study, and they agreed that they were relevant to their work at the Page.

Applicability to this study

The PAR approach was particularly well suited to data collection for the La Brea case, given my extensive prior experience at the site. I volunteered at the Page Museum for four years (2003-2007), worked there full time as an excavator for three (2007-2010), and have continued collaborating with them over the last six years of my graduate work (e.g. Thomer & Farrell, 2011). My experience gives me valuable insight into key social factors and tacit knowledge at each of the museum and excavation sites. With the PAR approach, I was able to use my domain knowledge and prior experience to my project's (and my communities') advantage.

An PAR approach is also suitable given this dissertation's focus on data standards and protocols work. Participatory approaches are crucial for the successful development and deployment of information standards (Yarmey & Baker, 2013). PAR has previously been used for information system development (e.g. Bishop, Van House, & Buttenfield, 2003; Lau, 1999), standards development (Millerand & Baker, 2010) and as a way to understand recordkeeping practices (Khoo & Rosenberg, 2015). Data standards development is a uniquely sociotechnical enterprise: it requires ontological engineering and systems analysis as well as a robust understanding of existing community data practices, and sustained engagement with said communities for long-term adoption. PAR provides an excellent framework for all these tasks.

Data collection for La Brea case

I roughly mirror the data collection in the YNP case, including the artifact collection informed by research process modeling. However, in the La Brea case I explicitly use a PAR approach to guide the collaborative re-engineering of data collection protocols, and gather data from this process. This re-engineering process takes place over a shorter time period, but is somewhat more iterative than my work on the YNP case, in that I build several rounds of protocol review and re-design into my study design.

La Brea is quite well positioned for this work; because of a recent turnover in curatorial and laboratory staff, museum administrators are eager to revisit their existing data collection and curation guidelines. Museum staff have been excavating specimens according to the same guidelines since 1969, but have lately been questioning whether this protocol is still sufficient for their work. My dissertation work will hopefully inform their excavation processes going forward (A. Farrell, pers. comm., 29 Dec 2015).

Specific data collection activities were as follows:

Interviews & other stakeholder engagement: I interviewed 10 staff members at the La Brea Tar Pits Museum. This includes one curator, two collections managers, four excavators, two curatorial assistants, and one interim laboratory supervisor. I was also embedded at the site for three nonconcurrent weeks while conducting interviews and re-engineering the data collection protocol; during this time I collaborated with the collections staff (primarily the collections managers) and solicited feedback on the design of the minimum information framework. Notes and memos from these informal discussions have been used as data.

I also interviewed 12 researchers who work with the collections. Eight of these researchers were identified through consultation with the collections managers, the remaining four were targeted through snowball sampling.

Artifact inventory and workflow analysis. As I did for the YNP case, I inventoried the La Brea Museum's data holdings, and modeled the La Brea staff's data collection and curation workflows.

Information framework development and comparative workflow analysis: Where at YNP I created a minimum information framework from scratch, at La Brea, I revised their existing excavation protocol based on data collected through interviews and conversations with the collections managers. This re-engineering took place through a PAR-informed, iterative process of:

1) requirements gathering and brainstorming through an initial on-site meeting
2) requirements gathering through formal interviews
3) on-site discussion of the protocol with the collections staff
4) further revision based on feedback

Steps 3 and 4 were repeated three times through three site visits. I then verified this new protocol's "fit" with La Brea through comparison it to the workflow models collections processes described above. This work is described further in Chapter 5.

External comparative analysis. As in the YNP case, I compared the MIF to a relevant of existing data platforms, focusing on the Paleobiology Database and Neotoma. As in the YNP case, I also

compare the fundamental structure of this MIF to two more general site-based standards, the BioCollections Ontology and the Environment Ontology.

3.4 ANALYSIS

As Yin describes there are five key components to a case study analysis. In the sections above I outlined the first three – my research questions, propositions, and units of analysis. Here, I describe my criteria for analysis.

My overall analytical strategy is guided by the revision of the theoretical propositions described above. These propositions are "linked" (in Yin's phrasing) to my data through explanation building and several rounds of internal and external comparison and analysis. My research questions all seek to understand how different aspects of site-based data work occur; thus, explanation building is an appropriate framework for my data analysis. Explanation building is typically done through iterative construction of narrative case reports, in which the researcher:

- makes an initial proposition
- compares the initial findings of a case against that proposition
- revises the proposition
- compares other details of the case against the revision
- and then compares this revision to the facts of a second or more cases (from Yin 2009).

I roughly follow Yin's steps in my analysis, but with some alterations due to my inclusion of embedded subcases in my study design, and my reliance on the comparison to external data standards in each case (Figure 3.2).

Initial coding and analysis: I use the YNP Case as my grounding case. Using Atlas.ti, I qualitatively code interview transcripts, memos, and artifacts according to a coding schema rooted in my research questions. I then analyze my coded data by writing narrative memos comparing my initial findings to my propositions and revise my propositions accordingly.

Embedded subcase analysis: Because my case study structure includes embedded subcases, and several of my research questions require comparing these subcases, I additionally compose separate narratives exploring propositions regarding the differences between research and curatorial work processes.

Intercase and intersubcase analysis: After tying the propositions to the facts of the core YNP Case, I compose a narrative describing the La Brea Case, and compare the revised propositions to the La Brea Case data. I again pay special attention to my embedded subcases.

External comparative analysis: Because my case study structure is built around the creation of minimum information frameworks, I then compare the frameworks to the external data standards listed above. I revise my propositions accordingly.

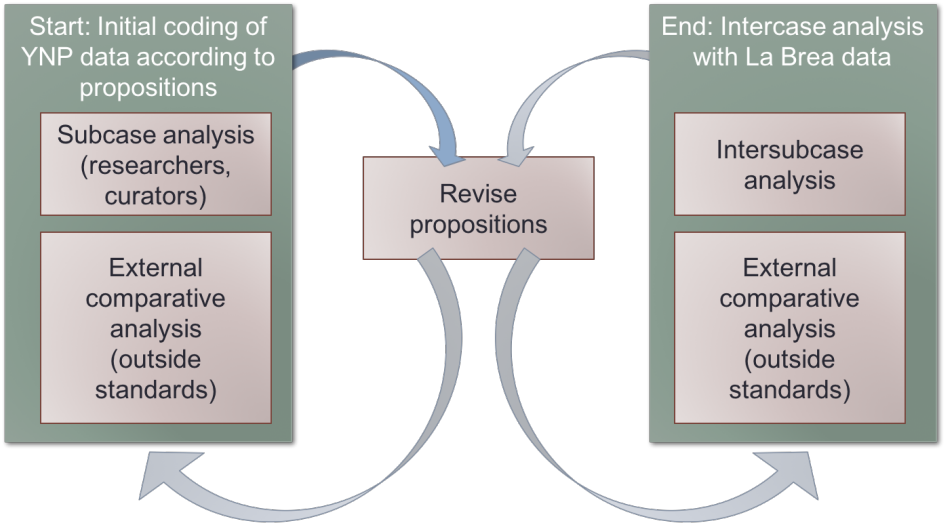


Figure 3.2. Analytical design. Data analysis is organized around iterative revision of propositions

3.5 HUMAN SUBJECTS

My application for data collection was submitted to and approved by the University of Illinois at Urbana-Champaign Institutional Review Board (IRB) in May 2016; the application materials are included in Appendix A. Participation in the study was voluntary, and participants were free to withdraw from the study at any time. Because my study sites are quite unique, and the particular details of that uniqueness will be critical to explaining the cases, I will not be de-identifying the sites. Thus, though I will be referring to my participants with coded pseudonyms, there is still potential for re-identifiability. Participants were told of this potential through the consent forms, and in-person prior to each interview. I have additionally shared drafts this dissertation with participants to ensure that they are comfortable with how they are being presented, and to give them an opportunity to ask that sections be redacted or more thoroughly anonymized (for instance, they are given the option of appearing completely anonymously, rather than being identified by their job title).

3.6 Limitations of Study Design

This study, like all case studies, is not intended to be generalizable. Case studies are frequently critiqued as being too singular to be informative for other situations; however, my use of a multiple case study design with external validation, as well as an iterative, explanation building analytical technique, will mitigate this concern as much as possible.

This study may also be limited in its choice of sites and domains. YNP and La Brea are both incredibly unique sites, both for their administrative structures and their natural features. The hot springs at YNP are some of the largest in the world; and the fossil deposits at La Brea are some of the richest. However, despite these unique features, preliminary findings indicate that there are indeed features of site-based data work that are common across these sites, and that may indeed be common across many others.

4. CASE STUDY NARRATIVES

In this chapter, I present my case study narratives, as developed through the iterative processes described in Chapter 3. Each narrative begins with a description of the natural site, its administration, and the research conducted at the site. I then describe the perspectives of each of my subcase/stakeholder groups (Resource Managers and Researchers). The Researcher and Resource Manager perspectives are summarized around four key themes that emerged from my coding: Roles; Values or Priorities; Data Needs; and Interaction with Information Structures (e.g., databases, data collecting protocols, and other information organization systems). I conclude with a brief summary of these comparisons. This chapter provides background for the the minimum information frameworks developed for each site, which are presented in Chapter 5.

4.1 THE YNP CASE: GEOBIOLOGY AT YELLOWSTONE NATIONAL PARK

From 2013 to 2015, I studied the data curation and integration needs of geobiology resource managers and researchers at Yellowstone National Park as the primary research assistant (and 2014-2015 project coordinator) for the IMLS-funded Site-Based Data Curation project (SBDC). One of the primary goals of the SBDC project was to develop "a framework of principles that helps to articulate and support upstream and downstream processes as a general model for site-based data curation" (Palmer et al., 2017). Though the NPS sets certain curatorial guidelines, they did not have robust policies or infrastructure regarding data curation; our work was an attempt to develop policies, protocols, and information models that could inform future policy and infrastructure development. In creating this case description, I revisit and integrate data collected during that time as well as follow-up interviews conducted over the summer/fall of 2016 (all described in Chapter 3).

4.1.1 THE SITE: GEOTHERMAL FEATURES AND THEIR MANAGEMENT AT YNP

Established in 1872 through an act of Congress, Yellowstone National Park is recognized as not just the first national park the United States, but in the world. The park is particularly famous for many of its rare natural features: It is the home to a number of "charismatic megafauna" such as bison, wolves, and moose. It has also been the subject of numerous first-of-its-kind ecology research, including longitudinal studies of the reintroduction of wolves to the park and studies documenting the impact of and unexpected rebound from massive forest fires. It is also the site

of the many thousands of dramatic geysers, hot springs, mud pots, and fumaroles that are the focus of research and resource management described in this case study (Figure 4.1).

Figure 4.1. **Photographs of two different thermal features at YNP. Left: An eruption at Castle Geyser. The red sediments surrounding the geyser are microbial mats. Photo from (Flicka, 2007). Right: a thermal feature at Mammoth Hot Springs; note the many travertine "terraces" formed by interactions between the water chemistry and microbes Photos taken by author.**

YNP's boundaries can be defined in several different ways. Geographically, the park is located primarily in the northwest corner of Wyoming, though it also extends into Montana to the north and Idaho to the west. Ecologically some aspects of the "site" extend even further: YNP's 2.2 million acres are the core of the 19-million acre "Greater Yellowstone Ecosystem" – one of the last largely intact (i.e., uninterrupted by human development) ecosystems of its kind ("Greater Yellowstone Ecosystem - Yellowstone National Park (U.S. National Park Service)," n.d.). Geologically, the site extends even beyond the ecosystem, below the earth's surface and under the entire Snake River Plain in Idaho. YNP sits atop the Yellowstone Supervolcano, which is fueled by an underlying "hot spot" in the earth's mantle. Though the Yellowstone Supervolcano has not erupted in 640,000 years, it is still considered active (and, somewhat worryingly, potentially overdue for an eruption). This volcanic activity heats and mobilizes underground waters and gases, creating the diverse geothermal features in and around YNP (Figure 4.2).

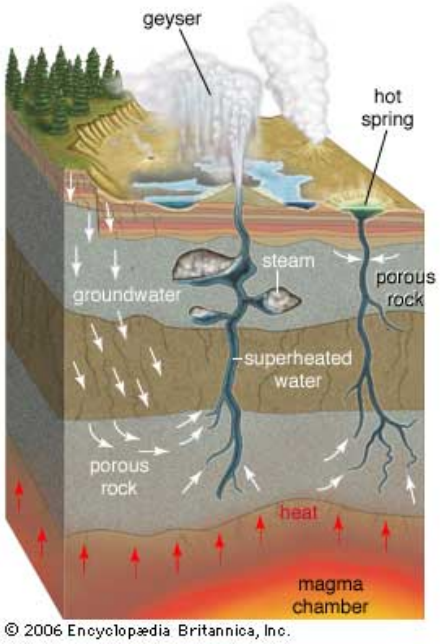


Figure 4.2. Anatomy of a hot spring ("hot spring geology," 2016).

These geothermal features are of interest to researchers such as geobiologists, microbiologists, and geochemists because they are the home to numerous charismatic microfauna: the rare and unique thermophilic (heat-loving) microbes that live in superheated waters of the springs. These microbes exhibit complex interactions with the spring water and the sediments in which they live; not only do the geological and geochemical characteristics of their environment impact their ecology, but the microbes themselves play an active part in creating geological and geochemical processes (Dilek, Furnes, & Muehlenbachs, 2008). Travertine (a kind of limestone) is precipitated by the hot springs at a rate of more than 1.5 m per year – an unusually fast rate of growth. Studies have shown a high correlation between microbial diversity and the specific type of travertine facies (the shape, structure, and chemical composition of the rock (Fouke et al., 2000)), but it is unclear to what extent the microbes impact travertine growth. Better understanding these water–rock–microbe interactions will help refine models of the precise drivers of mineral growth and weathering, the transformation of organic carbon into fossil fuels, and more. Additionally, understanding how biological and geological processes interact can help us answer fundamental questions about how life first appeared on Earth (Fouke, 2011).

Microbial thermophiles in YNP's hot springs have industrial applications as well. *Taq polymerase*, an enzyme critical to polymerase chain reaction (PCR; a method of DNA amplification that must be done at very high temperatures) was first isolated from the bacterium

Thermus aquaticus, which is found in the hot springs at YNP. PCR and *taq* patents have netted their holders over $2 billion a year (Fore, Wiechers, & Cook-Deegan, 2006), making hot springs, deep sea vents, coral reefs and their occupants prime targets for bioprospecting.

Although interactions between sediment and microbes can be observed in a wide range of locales, YNP offers some of the most accessible. The geothermal features at YNP are not only numerous, but also reliable and accessible; many of the springs are near paved roads, boardwalks, and trails, and though individual springs may "turn off" without warning, there are often other accessible springs nearby for study. Additionally, the rate at which microbes precipitate limestone in the YNP hot springs makes it possible for researchers to conduct observational research and experiments over relatively short periods of time.

Research access & resource management policy at YNP

As noted above, YNP was established via an Act of Congress setting it aside as "a public park or pleasuring-ground for the benefit and enjoyment of the people" ("History (U.S. National Park Service)," n.d.). This phrasing is informative: YNP was not necessarily established for research or ecological conservation *per se*, but initially for recreation. Further legislation has been needed to protect federal lands specifically for the purposes of ecological conservation, cultural heritage, and scientific research. Researchers must now apply for a permit before conducting any work in the park, particularly if it involves the collection of physical samples, specimens, or artifacts.

These permitting processes and scientific protections were established slowly. For instance, the Antiquities Act of 1906 provided broad protections of any kind of "cultural or natural resources," and laid the groundwork for the modern permitting process that regulates research access to national parks. Section 3 restricts specimen and artifact collection only to institutions

> ...properly qualified to conduct such examinations, excavations, or gatherings... [and for investigations found to be] ...for the benefit of reputable museums, universities, colleges, or other recognized scientific or educational institutions, with a view to increasing the knowledge of such objects. ("Antiquities Act of 1906," 2016)

These general guidelines have needed enforcement through more specific protections and laws, which have been set more slowly. For instance, it was not until the 2009 passage of the Paleontological Resource Protection Act that specific regulations were put into place requiring that the collection and management of paleontological resources on federal land rely on

"scientific principles and expertise" ("Paleontological Laws and Policy," 2015). Additionally, since the late 1980s there have been often-heated debates about the role and regulation of "bioprospecting" within the NPS, leading to the establishment of Cooperative Research and Development Agreements (CRADAs; also called "benefits-sharing" agreements), in which a park enters into an agreement with a non-federal scientist or company to share "reasonable benefits" (whether monetary or not) from "the commercial use of a discovery or invention resulting from research originating under an NPS Scientific Research and Collecting Permit" ("NPS Benefits Sharing," 2017).

The piecemeal development of permitting conditions and policies remains relevant today because permitting conditions and processes remain somewhat idiosyncratic to each park, and even to certain domains of scientific study. Although there are NPS-wide research permitting conditions ("NPS General Conditions," n.d.), each park has a fair amount of leeway in setting park-specific permitting requirements. YNP has several park-specific policies: for instance, researchers must agree to "carry out all of [their] activities out of public view" unless specifically authorized otherwise (per Condition 5). Additionally, there are several conditions regarding research reporting:

> 6. If you collect specimens that are to be permanently retained, regardless of where they are kept, they must be accessioned and cataloged into the National Park Service's catalog system, and must bear National Park Service accession and catalog numbers….
>
> 10. The Permittee agrees to notify the Chief of Resources of Yellowstone National Park (YNP) of every subject discovery or invention that relates in any respect to research results derived from YNP research studies or use of any research specimens or other materials collected from YNP, or that may be patentable or otherwise protected under the intellectual property (IP) laws of the United States or other jurisdiction.….
>
> 12. Each year, investigators are required to submit copies of journal articles, theses, and dissertations that resulted from park research activities to the Research Permit Office…. (From "YNP Permit Conditions," 2016; see Appendix B for a full list of research policies and conditions)

Per condition 6, all "permanently retained" specimens collected within a national park remain the property of the federal government and are accessioned into the Department of the Interior's collection[12], but may then be loaned back to their collector's home institution on a semi-permanent basis. However, samples and specimens collected for destructive analyses (e.g. water samples, small rock or biological samples) are not necessarily retained. This has resulted in what some consider a loophole in policy regarding the management of geobiology specimens: researchers are able to collect samples that result in the production of high-value data, but don't need to be accessioned. These data are subsequently difficult to track because the data are not "attached" to a physical specimen. Condition 10 does not require that this data be reported unless it is patentable. Moreover, although Condition 12 requests all copies of final research products, it does not specifically request raw or any other processed variations of the data. Further, my YNP collaborators have repeatedly indicated that Condition 12 is difficult to enforce.

Reporting and permitting conditions are primarily enforced by YNP park rangers – particularly those working in the Research Permitting Office (these rangers are hereafter referred to as permitting staff). Permitting staff interact with researchers at three primary points in time: when a researcher submits a permit application; when a researcher is at YNP conducting his or her fieldwork; and when researchers submit their Investigator Annual Reports (IARs) describing the year's activities (Appendix C). At the permit application stage, resource managers must ensure that the proposed project does not violate any laws, and will not negatively affect any of the natural features or other park visitors' experiences. At the fieldwork stage, resource managers must ensure that all safety protocols are being observed, and that the researchers are not deviating from their proposed plan of work. And at the year-end reporting stage, resource managers must do their best to ensure that they get as much of the currently mandated end-of-project reporting (see Condition 12 in the YNP Permit Conditions) as they can.

[12] Condition 6 actually reflects an NPS-wide policy regarding physical samples, specimens, and artifacts; it is unclear why this is listed in the YNP Permit Conditions and not in the NPS General Conditions.

4.1.2 Resource management perspectives at YNP

In this subsection, I present a subcase narrative describing the perspectives of the resource managers I spoke and worked with at YNP.

Over the course of this project I spoke with a range of personnel at YNP, including park rangers working in the Research Permitting Office (permitting staff); archivists and librarians working at the Heritage Resource Center (HRC); curators and collections managers working at the museum; and staff scientists who oversaw internal studies of key natural features. I refer to this group collectively as "Resource Managers": personnel tasked with managing information resources or natural resources, and/or enforcing the permitting conditions outlined above, as well as in related federal mandates.

The archivists, librarians, and museum staff typically had traditional full-time job descriptions and work allocations (i.e., they worked with their respective collections for 100% of their work hours), whereas the permitting staff often split their time between multiple roles (e.g., 25% in the permitting office, 25% working on an environmental monitoring project, 50% on other park management responsibilities) and moved around within NPS departments relatively often. Over the course of this project, I witnessed a notable amount of turnover in the Research Permitting Office (RPO) particularly. The two permitting officers I worked with most extensively left the RPO by 2015 (one still works at YNP but in another department; the other left the NPS entirely), and the two park rangers who took their places moved on to other positions in 2016.[13]

When speaking with the resource managers I found it helpful to rephrase one of my research questions. Because of their roles as gatekeepers/managers, they did not necessarily value natural aspects of YNP as a site; rather, they valued their ability to manage these aspects and/or make them accessible to current and future researchers. Thus, my questions were sometimes rephrased as "what information and information infrastructures at YNP have the highest value?" or "what do you value most about the site?"

---

[13] I am in contact with the current permit officer at YNP, who has expressed interest in this work and has been exploring ways of implementing some of the guidelines developed through this and the SBDC project; this is discussed further in Chapter 6

Resource manager roles: research support and permit condition enforcement

In general, the resource managers saw their roles as divided between supporting researchers in their work, managing park resources (either natural sites or specimen collections and archives, depending on the individual's specific job), and enforcing various NPS permitting conditions and regulations. At the April 2013 workshop, some further emphasized their roles as managers rather than strategic planners. As one member of the permitting staff put it, "The types of positions that we have are managing the information that is given to us but not necessarily determining what is important to get..." (YNP-RM-2). In general, they felt limited in what they had the authority to demand of visiting researchers – both in terms of what power they had been granted by the NPS, and in terms of what they had the right to demand of independent researchers working on their own idiosyncratic projects.

Many of our conversations centered on issues of governance, compliance, and enforcement: Who would be responsible for ensuring that researchers followed any data standards, if developed? How would these standards be enforced? As it stands, at the end of a project, resource managers already ask researchers for three categories of documentation: the NPS-mandated IARs, a YNP hot spring–specific field survey form (Appendix C), and any papers, theses, or other publications that result from the research (as dictated by YNP Permit Condition 12). Information collected through the permitting applications, IARs, and the field survey form are needed primarily for internal purposes related to environmental impact and safety; the resulting publications are seen as important parts of the park's history and heritage, as well as necessary for resource management and strategic planning.

Resource managers noted that they already had problems enforcing existing permitting conditions, largely because they had very little leverage once a permit was granted:

> This is the frustrating part of the permitting process. It really focuses on what happens on the front end, making sure that the person is providing all the deliverables before and during the [permit application] process and that they're maintained, and not impacting resources during their [fieldwork]. (YNP-RM-2)

That is, the only real "enforcement" mechanisms that resource managers (particularly, the permitting staff) had over visiting researchers was their ability to a) deny researchers a permit, or b) kick researchers out of the park while they are conducting fieldwork. This latter mode of

enforcement is very serious, would require support from law enforcement, and would only be enacted for egregious safety violations or damage to natural resources.

Thus, by the time researchers finished their projects and created the kinds of data products that needed to be shared with the park, resource managers had no real disciplinary mechanisms to use to enforce compliance. The rare exception to this was for researchers who were submitting a new permit application, but had not yet finished the reporting for their last permitted project. In these cases, resource managers could withhold a new permit until prior reporting was completed:

> It's difficult, but we are able to go back and say, "I'm sorry, we really can't have discussions with you right now because you still haven't finished your reporting requirements from your last permit." And we did have [one researcher who hadn't filed IAR paperwork] last year who approached us [for a renewal] and we took that tack and it worked. (YNP-RM-2)

Resource managers consequently had to come up with creative workarounds to obtain the information they needed for their work. For instance, because they received so few theses and papers sent back to them by researchers, they turned to setting search alerts for Yellowstone:

> I don't think we'd get 60% of the publications we get if [YNP-RM-3] and I and the gals didn't have Google Scholar alerts for Yellowstone. We went from having maybe 40-50 a year submitted to over 120 because we are actually looking for them. (YNP-RM-2)

Experience with information structures: infrastructural constraints

In addition to feeling constrained by their enforcement capabilities/responsibilities, resource managers felt constrained by their information infrastructure. In short, their information systems were designed to facilitate access to documents, not data. Information technology policy and protocols were generally guided by the NPS. YNP information infrastructures were largely dependent on existing NPS databases, computers, and web services. Two databases in particular were discussed – the Integrated Resource Management Applications (IRMA) system and the Interior Collections Management System (ICMS). IRMA is an NPS-wide, multipurpose database intended to consolidate a number of older, siloed databases containing scholarly literature, species range maps, species lists, and other "official" sources of parks data. ICMS is a collection management database used by all NPS museums to manage physical collections and accessions. (See Appendix C for more detailed descriptions of both.)

Both systems are capable of storing digital datasets to some degree, but they are not specifically designed for that purpose. For example, ICMS has a "related documents" field, which allows datasets to be attached to specimens or artifacts, but these account for only a small percentage of digital data collected in the park. IRMA, on the other hand, should be able to store datasets or metadata about them, but at the time of our workshop the database had been primarily used for data generated internally by the NPS. Resource managers were unsure if IRMA was expected to be used for, or could scale to include, non-NPS data, noting that prior attempts to store researcher data in IRMA had not been successful:

> We tried to put [remote sensing data] in IRMA but [the file size] was too big at the time, and now we might be able to put it in IRMA, but right now it's on a drive in [an NPS staff member's] office. (YNP-RM-2)

This experience suggests that my participants faced the same data management issues as at many other organizations, whether scientific or not: data tend to be stored in an *ad hoc* manner on local computers. There was a reliance on individual staff members' memories for locating data with no overarching structure and uneven formal documentation. The lack of documentation or systems for retrieval created a risk that information would be lost due to staff turnover.

This lack of formal data infrastructure made RMs understandably hesitant to collect data. They balked at the idea of asking researchers for additional data without being able to assure researchers that they had infrastructure in place to store and curate said data:

> There is no point in us asking for data if we don't have anywhere to put it, or any system to put it in, or any way to retrieve it. We couldn't even justify to people why we even need their data. (YNP-RM-1)

This RM went on to say, "I just don't feel honest about [asking for data] unless I have somewhere reasonable to put the data after they've documented it." The RMs were extremely aware that any further request for data would mean additional work on the part of the researchers, and the RMs were consequently hesitant to make those additional demands on their time.

The RMs also worried that additional data reporting guidelines would mean that they would receive more data than they had the capacity to review. The permitting staff particularly worried that they would eventually wind up with more files than they had the capacity to manage. They

disliked the idea of blindly collecting data without checking that it was appropriately curated, and further, disliked the idea of collecting a "huge massive pile of data" when there was only one dataset in it that might be potentially useful (YNP-RM-1). In short, there was a fear of overburdening their data curation workload through naïve over-collection of data. (This sentiment was echoed by RMs at Rancho La Brea, and will be discussed further below.)

Resource manager needs: What data would they want, if they could get it?

Resource managers generally wanted more (or more access to) information *about* research activities rather than access to the raw data itself. This included information that could be described as "metadata": information about the data collection activities going on in the park, and about how data was being collected (the "methods metadata" (Chao, 2014)) – in short, documentation of the "larger context" of data collection and project activities (YNP-RM-4). Several RMs noted that they simply needed to know where researchers' data were being stored. In this case, this may not have even required collecting extra information from the researchers; the archivists, librarians, and museum staff indicated that simply obtaining access to the research permits once they were filed would have been an improvement. Several archivists also indicated that they would have liked to store copies of data management plans, especially if they were already being written for various funding agencies:

> What would be helpful to get in the short term… is the data management plan. What are they going to do with the data when they are no longer active and people and can't find them anymore to ask directly? And I think now that they're going to have to often decide what repository they are going to put the data in sooner rather than later. That could be really useful. Because if they do retire and they can't be contacted anymore by somebody following up, we know at least where they thought they were going to deposit that information and that gives us a lead to provide to the researchers. (YNP-RM-4)

Data management plans (DMPs) were thus seen as a way of getting at least an idea of what researchers would be doing after they left the park. Resource managers recognized that this was an imperfect solution given that researchers may move, or may not follow through on the DMPs as intended, but considered it a start toward more complete data collection. One RM suggested that this sort of information could be requested on the Permit Application form (some information about where data are stored is currently requested in the IARs, but those are not always submitted or thoroughly filled out; requesting this information on the permit application

would at least provide the RMs the opportunity to critique the data management plan, even if tentative).

RMs were also interested in data that could help with strategic planning. Despite their not necessarily being responsible for this sort of work, they recognized it as something that they needed to support, if not personally direct:

> I would think then it would be worthwhile looking at that as a group and deciding are there things that are so critical or so fundamental for the park or in the future that we would want to make sure that we had access to that going into the future... it would be good if anyone knew of any and we actually have, you know, examples where that made or break can make or you know where has that been a huge hit to us just because we lost track. (YNP-RM-1)

Some of this awareness stemmed from an understanding that much of YNP's institutional record is stored in the current resource management staff's memories, rather than in the archives. At the 2013 workshop, two permitting staff discussed whether they felt like they had a sense of a) what was going on in the park at any given moment, and b) whether that represented a permanent archive:

> YNP-RM-1: I don't know that anybody has a handle on any year what data gets collected in Yellowstone National Park
>
> YNP-RM-2: [defensively] I think we do!
>
> YNP-RM-1: You guys, you do! [Laughter]
>
> YNP-RM-2: But yeah but you're right. You have to come talk to us. And then, what if there is a lot of turnover? What if [YNP-RM-3] and I are gone next year? You know what I mean? There is not that long-term memory where it's like 'oh yeah you wanted climate data? You go to this person over here.'

Thus, there was recognition of a need to record broader research trends in a more permanent manner. This exchange is all the more notable given that (as of this writing) all three of the permitting staff speaking or mentioned (YNP-RM-1, -2 and -3) have left the Research Permitting Office.

Finally, RMs expressed a need for better connections between siloed data sources (rather than necessarily needing more or new data). The resource managers were acutely aware that data relevant to their needs were often collected by the NPS or published by researchers – but the data

were stored in so many different places that it was difficult to aggregate or pull together. The museum collections staff particularly struggled with this:

> YNP-RM-5: We might have the physical specimens downstairs, there might be documentation somewhere in the archives but then the [digital] data will be somewhere else. Or sometimes the specimens and the data will be somewhere else but then they'll return the specimens to us but the data will still [be kept elsewhere].
>
> YNP-RM-2: It's like a spider web. No way to really know where the central body is.

Though they were sometimes able to store links to external databases in that "related documents" field in ICMS, there was no easy or programmatic way of keeping this information up-to-date or accessible. This is another example of *ad hoc* solutions to digital data management.

One archivist argued that the move to electronic data storage media themselves led to these silos in the first place:

> What we have a lot of in the archives now are the old-style field notebooks and those pieces, you know, they have all of that all in one piece. They've got a little diary narrative about who was part of the group and… then it also has the data as part of that whole notebook. And now its gets all parsed out into this form and that form. And sometimes it's hard to figure out how this one path equates to something else because there is not an overall connecting description of what was going on. (YNP-RM-4)

In the natural history tradition of work, data and contextualizing metadata (the "diary narrative" describing a day's field activities) were all captured through unified field notebooks. Through the piecemeal adoption of various computer-based reporting and data collection mechanisms, these unified, well-contextualized datasets have been split up and severed from their contexts.

4.1.3 Researcher perspectives at YNP

Over the course of this project I spoke with researchers who conduct work at YNP. Nine researchers attended the April 2013 workshop; this group included geologists, geochemists, and microbiologists. Follow-up interviews were conducted with two of these researchers as well as two additional researchers who had not been involved in the initial workshop. Additionally, I worked closely with Bruce Fouke (a prominent geobiologist and co-PI of the SBDC project) for three years, refining our understanding of YNP data needs, developing a Minimum Information Framework for geobiology (presented in chapter 5), and conducting several information and

process modeling exercises (also presented in chapter 5). I refer to this group of participants as Researchers: scholars who visit YNP to collect data, but who are not employed by the park.

Researcher roles and values

As described above, YNP Resource Managers were tasked with facilitating research within YNP while also protecting the park from those wishing to make use of it: it is the researchers, then, who needed both support and management. These researchers viewed the park as a "living laboratory"[14] – a place to conduct experiments and collect data – and they tended to prioritize their ability to complete that work over other factors (for instance, the preservation of natural sites). Occasionally tensions arose when scientific pursuits required slightly more-destructive methods than the NPS deemed appropriate. This is not to say that the researchers were not invested in preserving the park and its natural features; rather, they had a different idea of what preservation entails – and what its purpose is.[15]

Of the researchers who attended the workshop, five held faculty positions at U.S. universities; two were PhD students at U.S. universities; and two worked for government agencies (i.e., the Jet Propulsion Laboratory, USGS). Though several of the researchers already knew each other (or at least knew of each other's work), they were not necessarily collaborators. As one microbiologist described it,

---

[14] There is some interesting literature around whether this is a "living laboratory" or a "living museum" — the former phrase has popped up in legal proceedings around the copyrightability of YNP data, and the latter implies a different sort of preservation and data sharing mandate than "laboratory." (see Wood, 2000 for a brief discussion). Further investigation is out of scope for now but possibly something worth looking into in the future.

[15] A perhaps illustrative anecdote: Per YNP Permit Condition 5, all researchers at YNP must agree to "carry out all of your activities out of public view" unless specifically authorized otherwise. One researcher described having to literally camouflage himself and his students, and having to duck out of view whenever a park visitor approached. The researcher felt that this was a bit excessive: Why not let park visitors see that scientific research was active and ongoing at the park? The resource managers, on the other hand, wanted to a) ensure that the park visitors had an unobstructed view of natural features, and b) make sure no visitors got the idea that it was acceptable to venture beyond the boardwalk, onto sensitive and possibly dangerous ground. For the scientist, "preservation" of the site ought to have been for scientific data collection, but for the resource manager, "preservation" meant facilitating science while also discouraging activities that could damage the natural features and providing non-scientist park visitors a pleasant and uncomplicated visit.

> Although I think we have a very collegial community, it's really an individual effort. Whereas, like, coral reefs and [the] Deep Carbon [Observatory] project I suspect it's a real big team project. So, I don't know exactly how that will translate into difference in the way that the data results [are shared] and the interactions [among researchers] but I think it's probably important. (YNP-GeoBio-3)

That said, there have been prior efforts at coordination and community-building among YNP hot spring researchers, most notably through an NSF-funded Research Coordination Network (RCN)[16]. One of the primary RCN products was an interactive database of YNP geobiology datasets, with a map-based interface; the researchers involved in the RCN as well as the SBDC project expressed interest in building on these earlier successes through the SBDC project.

My participants described their research as involving multiple kinds of field observations (biological, chemical, physical, geological, and genomic) and sample collection. These heterogeneous data were captured as a suite of files and in a range of formats. Some data were handwritten in paper lab notebooks; some were entered into spreadsheets by hand in the field; and some were "born digital" outputs from handheld instruments. The physical samples sometimes needed to be sent to external laboratories for analysis such as mass spectrometry or radio isotope analysis if these services were not available at their home institutions.

Key parameters for reuse

In contrast to the Resource Managers, the researchers had a somewhat easier time identifying specific aspects of YNP that they found important. This is likely at least in part due to the nature of their work as scientists: they already spent a lot of time considering what aspects (or "parameters" or "data points," in their vernacular) of the site were more important than others. They also had a bit of an easier time speculating as to potential uses of a data store, albeit with some caveats regarding specific reuse, discussed further below.

The researchers broadly agreed that the following categories of data were highly reusable:

Genetic data. DNA and RNA sequences were unanimously seen as highly reusable. This perhaps makes sense considering the reliability of sequencing methods, the success and uptake of

[16] http://www.rcn.montana.edu/

sequence repositories such as GenBank, and the long history of impactful genetic discoveries at YNP. Researchers speculated that these data stores would only become more valuable in coming years:

> Right now we understand community by gene sequence. But we're probably within about five years of understanding community by thousands of genomes at once. And so the data sets are going to climb and grow and be huge, and it's going to become a very genomic thing, with whole genomes. (YNP-GeoBio-1)

Metagenomic approaches will likely bring a need for new methods of storing, sharing, and documenting that data; this could be an important design consideration for future information infrastructures and data reporting standards (though it is out of scope for this present work). It is also worth noting that genetic data are a product of "lab" work rather than fieldwork, and as such were considered somewhat out of scope for the SBDC project. However, given their centrality to geobiology, it is nevertheless important to consider the centrality of these downstream data products and their relationship to field data

"Basic geochemistry," data about the hot springs – particularly the water temperature and pH. Beyond temperature and pH, however, definitions of what "basic" geochemistry included varied. Some said "dissolved solids," other said "dissolved gases," and others cited chemical isotope data. One of the factors indicating whether a data point was "basic" or not seemed to be the data collection method: Data points collected through well-established methods would be easier to reuse (or perhaps more trustworthy) than data collected through less commonly used methods.

Photographs, at a range of scales. The researchers were almost unanimous in their enthusiasm for the usefulness of "meso" scale field photographs of the hot springs, noting that, "You can tell a lot from the picture and can tell a lot from the seasonal changes and the pH" (YNP-GeoBio-6). Another researcher stated,

> I don't get to go [to YNP] with my students so the first thing I always ask is, what did the spring look like? I've gotten into the habit that they have to show a picture and be able to point where they took the sample from and whether it was water or soil or wherever. You can tell a lot with that. (YNP-GeoBio-1)

Researchers broadly agreed that simple, point-and-shoot photographs of the spring and sampling sites made it possible for them to quickly assess site conditions at the time of sampling. Photographs also provided a basic understanding of how active (or inactive) a spring was, and of

other seasonal conditions that may impact interpretation (e.g, snow cover, microbial mat growth, etc.). Researchers agreed that developing systems to support sharing these photographs could be important for the community.

Researchers also noted that micrographs (photographs taken with a microscope) were highly useful. These sorts of images are taken once samples are returned to the lab. Micrographs include fluorescence *in situ* hybridization (referred to as FISH) images that show the relative density of different microbial communities (Kubota, 2013), and images from thin sections. Like genetic data and more complex chemistry data, micrographs are the product of lab rather than fieldwork.

Finally, researchers also valued some more abstract qualities about their data, and about YNP as a site. The concept of "context" came up repeatedly; well-contextualized data sets were more valuable and useable than non-contextualized datasets. "Context" could include information about data collection methods; information about environmental conditions; or information about project goals and hypotheses. Researchers did not necessarily distinguish between kinds of context; a researcher's method of data collection was bound up with the "intended use" of the data, and environmental conditions and hot spring descriptions were similarly tied to sampling strategies. Additionally, researchers also valued mechanisms that encouraged rigor in data entry. As discussions moved to issues of data standards implementation and reporting requirements, they were all aware of their own tendencies to simply check boxes to comply with NPS mandates. Researchers worried that this would result in lower quality data, and expressed a desire for UI/UX mechanisms to help ensure that they were not blindly entering data – but rather, were being mindful about data entry.

Kinds of reuse

Though researchers had little difficulty describing what kinds of data they may have wanted from a data aggregate or system, they hesitated at the idea of directly reusing someone's data, particularly without consulting with them personally. One argued that he would not feel right using someone else's data because "that's their work" (YNP-GeoBio-6); data reuse in this case felt akin to plagiarism. Many, though, attributed a reluctance to reuse data to a need to "know" their dataset in a way that is only possible if you yourself collect it:

> I always tell grad students and postdocs, if you want to do and work on a question, you need to get your own DNA, you need to get your own sample so then you're working

> with it in the context of what it is. That's kind of an important thing, getting your own sample. Because you know it better. (YNP-GeoBio-1)

In other words, researchers wanted an experiential knowledge of their data to feel comfortable using it, which is incredibly difficult to replicate through metadata.

Two kinds of reuse seemed more amenable to researchers: reference reuse and retrospective correlative reuse. By reference reuse, I mean using a data store as a baseline against which to compare new data (e.g., comparing genetic sequences to other sequences in GenBank). Reference reuse could also include assessing a data store to identify current gaps in knowledge and develop new project ideas. As one researcher described,

> Data sharing… it's going to be not to use that data – it's going to help me formulate the next set of questions. And so I want to cater to generating the next generation of questions. (YNP-GeoBio-2)

Retrospective correlative reuse, on the other hand, is more akin to a meta-analysis. The phrasing of "retrospective correlative study" is a direct quote from one of my participants who expressed optimism that this would be a valuable approach for researchers at YNP in the future:

> I do really think there's a real future in correlating the geochemistry with the biota and the genetic content. And that's going to be a retrospective correlative study – it's going to try to mine what's there. And we talked about: does the environment or the history determine what's there in the microbial world? I've got to say that the data looks suspiciously like there really is limited genetic flow. How that holds up is going to depend upon more data points and being able to retrospectively ask, are these other observations consistent with it? And do we find things in the genetics that correlate with different chemistries? Because we've still got the majority of these genomes that we don't actually know what they do. (YNP-GeoBio-4)

Through additional conversations with YNP researchers, we got a better sense of what "further data points" exactly it would take to make this kind of "retrospective correlative study" possible. Firstly, researchers would need to correlate data points with a lot of geological and geochemical context: for instance, information describing from which part of a spring a sample was collected (e.g., near the "vent" of the spring – the source of the water – or near the drainage system); or even simple descriptions of whether the spring was located within the Caldera at YNP or outside of it. Secondly, researchers would need to come to some sort of agreement about what variables within a spring are "key" to understanding the geobiological relationships:

> At the same time, just an *ad hoc* collection of geochemical data correlated with the microbial samples – it fails to recognize sometimes what the key variables are if the people collecting it aren't thinking really about those key variables or what have you. (YNP-GeoBio-2)

Retrospective analyses are much more feasible if researchers – even when working independently – strive to develop and adhere to community guidelines that identify key data points for collection.

Experience with information structures: concerns about data integration

The YNP researchers were well aware of the impact their data collection protocols had on the reusability of their datasets, and they were also well aware that information modeling and information organization choices could impact their data's fitness for use. Some of this discussion became very abstract; one researcher described a data array (including its organizational structure) as the characterization of a spring – not just the parameters:

> This concept of an experiment or an observation – the super row in the spreadsheet, the whole array of data, and the question…. This is [a] characterization of a spring, as close temporally as we can assemble the data for [it]. (YNP-GeoBio-4)

Another said that the data protocol itself constituted context for a sample: "…The context in the sampling, like you guys said – that's everything. That is what the data actually relate to, is how the person took the sample" (YNP-GeoBio-2). Many of these comments seemed grounded in the researchers' training in and use of statistical methods of analysis. Other topics of discussion included whether it would be feasible to integrate datasets collected with different protocols into the same data frame within statistical software, and whether averaged measurements were more or less reusable than the raw data. Some of my participants' concerns about data reuse were thus likely rooted in concerns about the feasibility of integrating two different models in a study.

The researchers suggested certain kinds of monitoring at sites as a way to get around the challenges of integrating heterogeneous data structures, for instance, "temperature monitoring in certain sentinel springs." However, given the reluctance of YNP RMs to allow permanent installation of equipment in the springs, this seems unlikely at this time. Thus, some of their concerns regarding reuse were also reflective of the fundamental challenge of *ad hoc* environmental monitoring – that is, monitoring different site variables (temperature, pH) without true monitoring infrastructure. YNP is an environment in which permanent monitoring

equipment is not possible, the natural system is dynamic, and researchers work in small, not-necessarily coordinated groups – all factors that will make true environmental monitoring a challenge.

4.1.4 YNP CASE SUMMARY

Summary of resource manager perspectives

In summary, the resource managers I spoke with did not necessarily prioritize data collection around specific aspects of Yellowstone's natural features; rather*, they valued data that improved their ability to broadly serve the mission of the park and the needs of visiting researchers.* This included metadata describing research activities overall; documents such as research permits; and broad "contextualizing" descriptions of researchers' goals and methods in conducting samples and specimens. *They wanted to understand research activities in aggregate and to preserve the research heritage and institutional memory of the park.*

Resource managers additionally valued *efficient and well-justified reporting requirements as well as immediately useful and well-justified collection of data.* Any additions to existing reporting requirements must not put an unreasonable burden or workload on researchers, and they must not saddle the RMs with data they have no concrete need for or resources (infrastructure, time, people) to manage.

Finally, resource managers valued the ability to *link/query contextual data across data silos*. They needed to understand individual data points in the context of the field projects from which they were collected, and therefore in the context of a broader data collection. They were, however, constrained by existing NPS infrastructure and curatorial practices.

Summary of researcher perspectives

YNP geobiology researchers by and large worked independently, but described their community as collegial. They consulted with YNP Resource Managers intermittently, and primarily during early stages of their work (permitting and data collection). They valued several categories of data points from the hot springs, notably, genetic data, geochemical data (particularly that which was assembled through reliable and well-established methods), photographs, and complex contextualizing information describing data collection methods, environmental conditions, and a project's hypothesis. At the same time, they were hesitant to reuse other researchers' data without consulting with them first. They were more interested in slightly more oblique forms of

data reuse: reference reuse and retrospective correlative reuse. Some researchers were quite aware of the impact that information modeling and information organization practice had on their research, going so far as to argue that their information organization/modeling strategies represented a form of context, and had ramifications for data reuse. This was possibly a result of their familiarity with and reliance on statistical methods, which would likely prime them to think about the impact of sampling and representation strategies.

Researchers preferred curatorial or resource management policies that facilitated research access to the sites. They prioritized their scientific work over other curatorial conditions, which is not to say that they did not value resource management or data curation – they just attended to it after they attended to their own scientific goals. Consequently, they valued reporting mechanisms that keep them "honest" by somehow forcing them to report carefully, rather than through rote input of information.

4.2 The La Brea Case: Paleontology at The La Brea Tar Pits

My second case focuses on research and curation at the La Brea Tar Pits and Museum. I have a long history with this site and in paleontology: from 2004 to 2007, I volunteered at the La Brea Museum as a lab technician and excavator; in 2007, I was hired as a full-time excavator with their new excavation project (Project 23) and promoted to the position of Senior Excavator in 2009. I continued in this position until I began my master's degree in 2010. During my time at La Brea, I helped create new data collection and documentation standards for Project 23, designed a relational database and information model for our field data (Thomer, 2009; Thomer & Farrell, 2011), and prototyped a small field data visualization project (Thomer, Thara, & Wilson, 2009). After beginning graduate school, I continued working with paleontology collections via a summer job as a curatorial assistant for the natural history collections at Petrified Forest National Park (PEFO) in 2011. At PEFO I assisted in data cleaning, curation, and inventory, and occasionally worked in the field with the research team. I draw on my prior experience at La Brea and my knowledge of paleontological curation and research practices in writing this case description. Taking a participatory action research approach, I worked with stakeholders at La Brea – the curators, collections managers, and excavators that worked at the museum, as well as the researchers that used the collections – to develop a new excavation

protocol for the site. The case study below is drawn from interviews and experiences conducting this work.

4.2.1 The site: Entrapment and excavation at La Brea

The La Brea Tar Pits are a cluster of incredibly rich fossil deposits that just so happen to be located in the heart of Los Angeles. Though colloquially called "tar pits," these geologic features are actually a series of naturally occurring asphalt (also called bitumen, a form of crude oil) seeps fed by the underlying Salt Lake Oil Field (one of the many hundreds of oil fields beneath the L.A. basin). During the Pleistocene era (~10,000-50,000 years ago), this asphalt began leaking to the ground surface, creating shallow puddles of sticky oil. These asphalt puddles would often become covered with leaf litter from surrounding oak groves, and water from nearby streams, creating and camouflaging a flypaper-like environment that quickly entrapped unsuspecting megafauna such as mammoths, grounds sloths, and camels. An initial entrapment likely caused a chain reaction and led to many more; the initial stuck animal would appear to be easy prey to passing predators – who would attack, only to become trapped along with their intended prey. It would have been quite a grim scene: Pack animals such as saber-toothed cats and dire wolves would dive in, and get stuck, along with all their kin (Figure 4.3).

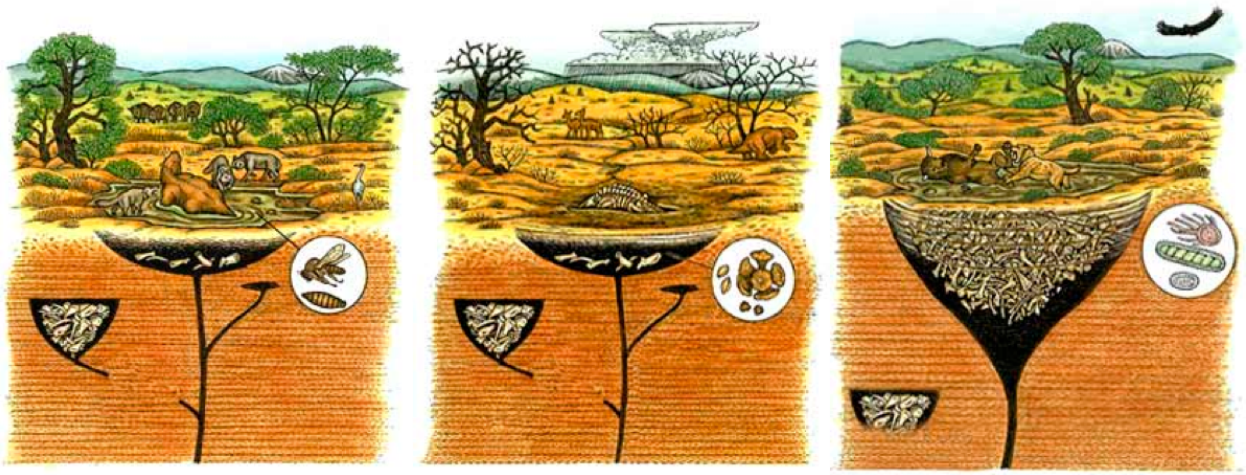


Figure 4.3. Diagram representing how animals became entrapped in asphalt. From (Harris, 2007)

The oil is an excellent preservative, though it does not technically fossilize bones.[17] Rather than replacing bone with minerals, the oil seeps into the bones' pores and prevents decay. This process preserves bones in incredible detail, and at all scales – everything from microscopic ostracods (a kind of crustacean) to 10-foot-long Columbian mammoth tusks have been recovered. An estimated three to four million specimens have been recovered in total, from 231 species of vertebrates, 234 invertebrates, and 159 species of plants ("La Brea Tar Pits FAQs," 2015).

While there is evidence that indigenous Americans used the asphalt from La Brea as a construction material, the first written descriptions of fossils at La Brea do not appear until 1848. The bones were thought to be of cattle until the first excavations in 1901, when field workers employed by the University of California at Berkeley dug 101 exploratory pits in the park. Approximately 20 of these pits were fossiliferous (fossil bearing). After removing the biggest fossils from the pits, they usually refilled. Field notes taken at the time indicated that one of these excavations, Pit 91, was filled in without being completed. In 1969, Pit 91 was "reopened" (the fill was removed), and excavation began again under the jurisdiction of the L.A. County Museum of Natural History (NHMLA). In 1970, a museum was built specifically for La Brea, originally named the George C. Page Museum of La Brea Discoveries, now simply the La Brea Tar Pits Museum. The La Brea Museum is operated as a branch of NHMLA.

Pit 91 was the sole active excavation at La Brea from 1970 to 2008, when the neighboring museum, the L.A. County Museum of Art (LACMA), uncovered 16 fossil deposits while attempting to build an underground parking garage for the newly constructed Broad Contemporary Art wing. Rather than excavating individual fossils *in situ*, the deposits were

---

[17] Fossils form in a variety of ways, but most typically through some sort of rapid burial in damp sediment (e.g. in a riverbank after a flood, in a mudslide; the study of how organisms become fossilized is called *taphonomy*). As a result, the fossil record may show considerable bias towards animals that happened to live near riverbanks, seashores, and at the base of unstable hills. After an organism is buried, its soft tissues (skin, organs) rot away, but its bones remain. Over time, the water in the sediment precipitates minerals into the pores of the bones (or in the case of plants, in between cell walls), which eventually replaces the bone itself, essentially turning skeletons (or stems) into stone. This process is called *permineralization*. Fossils can also form without mineralization in extremely dry and undisturbed environments such as caves (e.g., Prideaux et al., 2007), or in anoxic environments such as peat bogs or oil and asphalt seeps.

excavated *en bloc*, crated into 23 large tree boxes, and physically moved to the northernmost edge of LACMA's property, out of the way of construction (Figure 4.4). These 23 boxes sat at the north edge of LACMA's grounds until 2008, when LACMA provided the La Brea Museum with a small amount of money for curatorial work and craned the boxes over to the La Brea Museum's side of the park, where paleontologists could excavate the specimens. This excavation project is now referred to as "Project 23" (for the 23 boxes), and it continues to this day. Excavation in Pit 91 is currently on hiatus until the completion of Project 23.

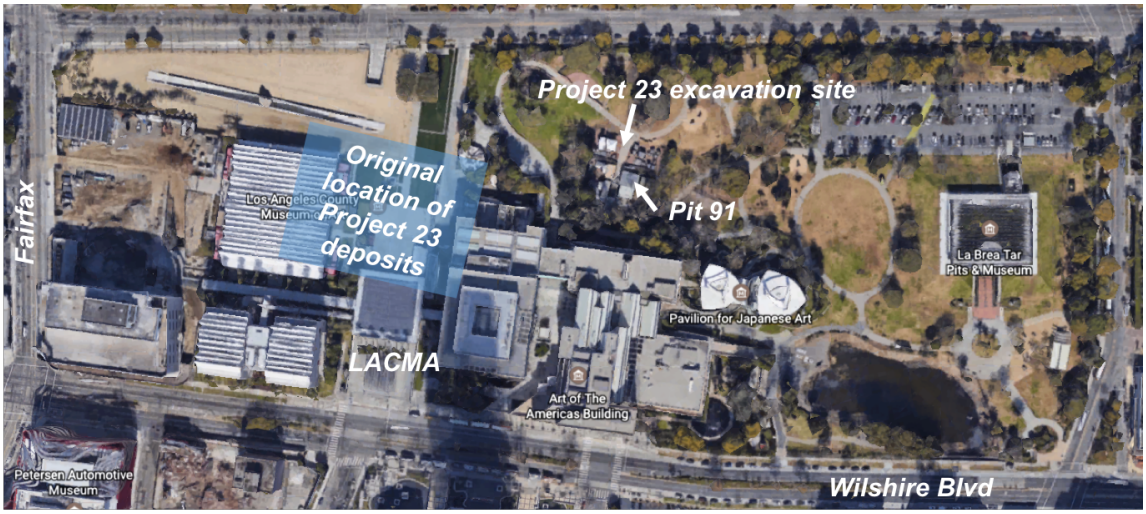


Figure 4.4. Map showing original location of Project 23 deposits, their new location (after being moved to the Tar Pits' side of the park), and the locations of the LA County Museum of Art and the La Brea Tar Pits museum.

Data collection and collections management at La Brea

The La Brea Tar Pits Museum is located on the same city block as the tar pits. As one excavator described it, "it's a site museum. It's all here. We're sitting on the environment, we're sitting on the fossils.… we have this opportunity to look at our field site all the time" (RLB-Exc-1). The main fossil deposits are all a few hundred yards west of the museum (Figure 4.4). The Museum houses almost all the fossils recovered from the site,[18] as well as the fossil preparatory lab (where bones are scrubbed clean of oil via an array of chemical solvents and donated dental tools); a

[18] Notable exceptions include a fairly large collection of fossils from early excavations that are somewhat notoriously stored in the bell tower of the UC Berkeley Museum of Paleontology (El Shafie, 2016).

small library and archive; curatorial and administrative office; public exhibits describing the deposits and the history of Ice Age Los Angeles; and, of course, a gift shop.

La Brea is somewhat unique among fossil sites in that the deposits are excavated by museum staff and dedicated volunteers, rather than visiting researchers. At many other sites – even those on federally managed lands – fossil prospecting (searching for fossiliferous localities) and excavation are carried out by external research teams. For instance, though the museum at the Petrified Forest National Park accessions and manages all fossils excavated within the park, they are usually excavated by visiting researchers, and then loaned to them for study via the curatorial mechanisms described in section 4.1.

The reliance on "in-house" excavators and the uniquely dense qualities of the La Brea *lägerstatte* make it possible to use a highly standardized data collection method. Since 1969, La Brea staff have used an incredibly detailed, grid-based measurement system, in which multiple coordinates are recorded for every specimen over ¼ inch in size (Shaw, 1982; Figure 4.5). This excavation and data collection method is rooted in archeological practice because the researchers who began the excavation believed that they would find human remains. While this positional data theoretically makes it possible to reconstruct the position of every single fossil within the deposit, there are not yet any computer programs built for or capable of this task.

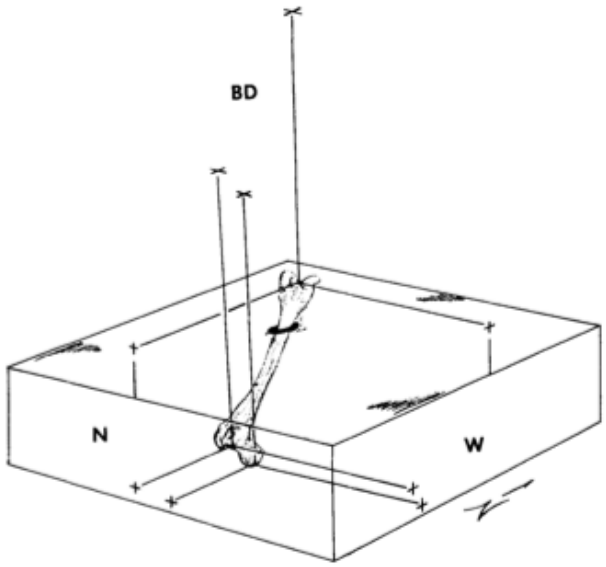
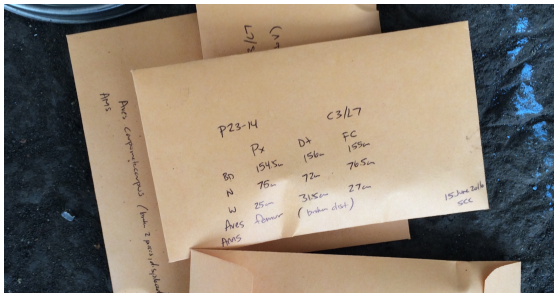


Figure 4.5. Illustration of the "3-point measuring system" currently used to record fossil locations at the La Brea (Shaw, 1982).

These positional data are referred to as "measurements" by the Rancho La Brea (RLB) staff; the act of taking measurements is referred to as "measuring out a bone." Recording measurement data takes up the bulk of excavators' time (the volunteers can help with the actual process of digging, but cannot record this data). Measurements are recorded in each excavator's field notebook and an accompanying 3x5-inch index card with a carbon copy of the field description (Figure 4.6). The field notebooks also contain short narrative accounts of the day's notable events (e.g., puzzling geologic discoveries, the names of volunteers on site for the day, descriptions of strange interactions with odd park visitors). The field notebooks are archived in the museum, and are in various states of digitization; books from the 1960s and '70s have been transcribed on typewriters, others have been typed into word processor documents, and the most recent have been input into Excel spreadsheets to provide collections managers with a rough inventory of yet-to-be-cataloged fossils.



	P23-14	**C-3/L7**	
	Px	Dt	FC
BD=	154.5cm	156 cm	155 cm
N=	75 cm	72 cm	76.5 cm
W=	25mm	31.5 cm	27 cm
	Aves femur (broken dist)		

Figure 4.6. Left: a manila envelope showing a "3 point measurement" of a bird femur from Box 14. This measurement was originally handwritten into a field notebook; the envelope features a carbon copy of the notebook entry. Right: a transcription of the 3-point measurement. The code in the top row is a Project 23 grid number, and the abbreviations in the second row are anatomical abbreviations meaning *proximal*, *distal* and *fovea capitis*.

This rough inventory is important because many of the fossils in the La Brea collections have not yet been cataloged. It is estimated that they hold more than one million fossils, of which approximately 560,000 been cataloged (and therefore, counted).[19] Cataloged fossils are identified to taxon (species if possible) and element by a curator or collections manager and assigned a catalog number. This information, plus a description of the fossil's locality (the pit, grid, and

[19] Some estimates by past collections staff put this number at far over 1 million – closer to 3-4 million. This includes the great number of microfossils found at the site, as well as many small fragments of fossils.

measurement data, see Figure 4.6 above) is recorded in a large paper catalog ledger. These paper catalogs remain the primary method of organizing, accessing, and documenting the data collections; as in many museums, La Brea collections staff have found that paper ledgers are more sustainable and reliable than digital methods of data management.

Like the field notes, these catalog ledgers exist in various stages of digitization. The original pre-1970 fossil catalog (the Hancock Collection) was almost completely digitized into a database called Paradox at some point in the 1970s or 1980s (current staff are unclear on the exact date; however, it was created long enough ago that it was programmed using punch cards – some of which are now part of an exhibit in the museum (Figure 4.7)). This database was used sporadically, primarily by collections management staff, and primarily to create inventories, look up basic information about fossils, or print labels. Though Paradox is a relational database system, the Hancock database is a single table flat-file. This database was maintained through the ’80s, ’90s and early 2000s, and was migrated periodically between computers and versions of Paradox. Around 2008, NHMLA began migrating many of its collections databases to an off-the-shelf relational database system called KE EMu, designed specifically for use with museum collections. The collections staff now use this database and several spreadsheets containing auxiliary information for collections management.

Figure 4.7. Photo of exhibit featuring the old punch cards used to database the collections. Photo taken by author.

Researcher access to La Brea data and collections

The collections at La Brea are not typically loaned out; researchers must use the La Brea collections *while physically at La Brea*.[20] There is no standardized permitting process to access collections; rather, researchers must directly contact the collections managers to request access to the collections ("Current Research," 2015). Typically, they are asked to submit a project proposal, outlining their project's goals, the number and type of specimens they wish to use, and describing in detail any potentially destructive analysis that they may perform (e.g., radiocarbon dating, which requires a small bone sample). Destructive analyses are particularly scrutinized, and steps are taken to mitigate potential harm to the fossils. For instance, a researcher might be asked to create casts of fossils before taking samples for isotope analysis in case the sampling processes affects the integrity of the fossils.

Figure 4.8. A view down the "primary range" holding the fossil collections at La Brea. Photo taken by the author.

Some researchers browse the collections drawers themselves (Figure 4.8) to select specimens; others work with collections managers to find appropriate fossils. Even the researchers who roam

[20] Exceptions have lately been made for cans of "bulk matrix" (cleaned sediment containing microfossils), but this is a relatively recent development, and the loan process was carefully planned as part of an NSF project in which the La Brea curator is a Primary Investigator.

the collections freely check in and out of the museum at the beginning and end of each day, and work closely with the curatorial staff to guide their specimen selection. At times this level of consultation becomes collaboration; collections staff (meaning the collections managers and curators – less often the excavators) often collaborate with researchers on grants and papers.

4.2.2. Resource management perspectives at La Brea

As noted before, La Brea is somewhat unique for having a full-time excavation staff, as well as several other full-time collections staff. This, however, has not always been the case. Budget cuts (as well as a dramatic embezzlement scandal involving the then-museum director and his secretary) led to massive layoffs of collections staff at both NHMLA and La Brea in the 1990s (see Biederman, 1993a, 1993b; Feldman, 1995; Glionna, 1995 for media reports on this). Consequently, the La Brea Museum spent more than two decades (from approximately 1993 to 2016) without a dedicated full-time curator – instead, a single curator was appointed to manage both the RLB collections and as the vertebrate paleontology collections at NHMLA (leaving the collections manager and laboratory supervisor to pick up some of the slack).[21] Excavations were restricted to the summer months, and much of the excavation and fossil preparation work was done by volunteers.

At the time of my dissertation work at La Brea, the museum staff had recovered from these past cuts in some ways, but was still adjusting in others. The museum by then had funding for a full-time curator, two full-time collections managers, a full-time laboratory supervisor, a full-time laboratory assistant, and four full-time, year-round excavators. However, the benefits of this new

---

[21] A note on museum staff titles and roles:

- An NHM curator is typically the head of a museum department; they usually have a PhD related to the department's focus, and act as a sort of lead researcher. Curators set the department's research agenda, guide curatorial policies, interact with external researchers to encourage work with the collections, lead specimen collecting trips, and so on. They often oversee other staff members as well (though this is variable depending on the museum).
- Collections managers do exactly what their title implies: They manage the collection. This includes cataloging specimens, managing digital collections databases, managing the physical arrangement of specimens in the museum, and helping researchers access the collections.
- Preparation laboratory employees "prepare" fossils for use in research – that is, remove them from the sediment they are encased in. At La Brea this work involves a lot of solvents used to dissolve the oil from the bones.
- Excavators dig fossils out of the ground and record "field data" about those fossils – information about their specific location and preliminary identifications as to their taxa and element.

funding had been temporarily deferred because the three longest serving staff members (the collections manager and the laboratory supervisor who began working there in the 1970s, and the part-time curator who began in the 1990s) had all retired within the preceding five years. The laboratory supervisor and curator positions each took over two years to fill – meaning that, once again, much of the museum's research and collections management fell to the other members of the collections staff, primarily the two collections managers.[22]

I worked most closely with these two collections managers throughout this project (they have been on staff at La Brea since 2008 and 2010, respectively). I additionally interviewed two curatorial assistants, the interim laboratory supervisor, the four members of the excavation staff, and the new curator once she began her position. It is worth noting that many of these staff members started work as volunteers; one of the excavators had been working at La Brea in some capacity since 2005. The staff was eager to review their data collection methods before the arrival of the new full-time curator in late 2016; they had long felt that they were in need of revision or even replacement. However, the La Brea RMs were understandably hesitant to make changes before their new curator's arrival. Consequently, we discussed potential changes to the excavation method but did not test any of them (potential ways of trying out changes are described in future work in Chapter 6).

Resource Manager roles and priorities

Ordinarily, collections managers spend most of their time managing their fossil collection and its associated data. However, the collections managers at La Brea saw their responsibilities expand during this transition time without a curator and a laboratory supervisor. One of the collections managers described her newly expanded work responsibilities in the following way:

> In theory, my primary job is to identify, catalogue, curate, and database our collections….. But, then in addition, since we don't have any onsite curator, I play a role

[22] The museum does have a curator of birds, but he is not involved in the broader work of the museum (including collections management and excavation. They also have an interim, "off-site" curator, but he is only on-site once a week, and mostly intended to handle big-picture issues (e.g., leading the search for the full-time permanent curator, guiding exhibit design, etc).

> in helping to sort of guide the program, but not in a new way, just to sort of keep it all together. And I oversee the current Project 23 excavation…. (RLB-Coll-1)

In short, the collections managers' roles shifted from just the management of the collections and collections databases to also include the management of people and supplies.

During my visits to La Brea, I witnessed the collections managers' increased workload firsthand: their days were packed. They usually arrived before 8:00 a.m. and spent the quiet first hour or two of the morning cataloging fossils and catching up on emails. By 10:00 a.m. their shared office was flooded with visiting researchers needing access to the collections; staff and volunteers with questions about protocols or priorities for the week; museum administrators seeking feedback about new exhibit designs; technicians needing access to different parts of the lab to repair a frequently broken HVAC system (for instance, there was a large Freon leak the day before I arrived for my first visit, rendering the lab unusable for a weekend and leaving the collections staff with weeks of clean up and repairs); consultants brought in to design new solvent storage areas; colleagues from the education department who needed advice about public programs; and so on. About 80% of their time was spent working with, guiding, or otherwise managing people, and only about 20% working with and managing fossils and fossil data. Both collections managers described feeling overtaxed with these varied responsibilities; they were looking forward to a reduced workload and reduced supervisory responsibility when the new curator and lab supervisor arrived. The also hoped for further staff additions in the near future; they hoped they would be hiring a field supervisor for the excavation work within the next year.

The excavators saw their roles and job strategies shift in the curator's absence as well, but in different ways. In short, they had been trying to change some of their excavation strategies, but struggled to make these changes stick, or to understand what exactly was expected of them. One said their job was, "to protect the fossils at all cost, and excavate with the best methods, and safe practices that I can" (RLB-Exc-2); another simply said, "from my perspective, my job is just to collect" (RLB-Exc-5). However, they were at times unsure about what excavation methods, exactly, represented the "best" methods. The current group of excavators were all trained to excavate the fossils using the methods from the Shaw (1982) paper described above (and shown in Figure 4.5). However, there was growing awareness that this method held some serious drawbacks, particularly for Project 23. This method was rooted in archaeological practice

because it was assumed that Pit 91 would contain human remains; that had not been the case, either with Pit 91 or with the Project 23 fossils. This method – particularly the measurements of the bones' location – was intended to make it possible to reconstruct the structure of the deposit, and the orientation of the fossils. However, the Project 23 boxes were no longer *in situ*, and many had been found to be "disturbed," or jumbled around through the process of boxing them up and moving them to the park. Moreover, this method was extremely time consuming to use; measuring out fossils individually slowed down the excavation process considerably. Project 23 was intended to be a 5-year salvage project, yet it has been ongoing since 2008 and as of this writing, is only about one quarter completed. It was clear to the RMs that this current method needed to be altered – but it was unclear exactly how.

The interim laboratory supervisor had also been wrestling with some uncertainty around his workflows. Acting as something of an intermediary between the collections managers and the excavation staff, he supervised all the fossil preparation work that occurred in the lab, as well as the volunteers doing the fossil preparation work. Over the previous year he had focused on finishing up fossil preparation projects that were left over from his predecessor's tenure, and getting the lab "*ready for whatever type of preparation technique that needs to occur to it*." He additionally worked with the collections managers to begin revisiting their fossil preparation techniques to bring them more in line with external paleontology best practices. However, there were sometimes misunderstandings around how certain tasks should be handled.

Thus, the RMs all saw their jobs as centering around the protection and preparation of the fossils – but they all worried that their data collection was no longer driven by some sort of research design. These concerns resulted in some high hopes for their new curator; they hoped she would provide them with a new sense of direction. One excavator summarized these hopes and concerns particularly well:

> We've been talking about making a t-shirt that just says, 'When the curator comes…' [laughs]. We have a very Messiah hope with our new curator, which is unfair. But, we're like -- you know, we need to justify what we're doing, we need to have a plan. We need to be like, "Hey, we're doing this for this reason. We want this research goal accomplished. So, this is what we need for this goal to be reached." But, right now we're just accumulating data. And as you know, that is not always the most useful thing. (RLB-Exc-4).

Without research-driven collections or curation guidelines, the La Brea RMs relied on three different strategies to guide their work: 1) changing methods in small ways in reaction to specific situations; 2) collecting data via existing methods for the sake of consistency or "tradition"; or 3) turning to other museum-administrators or the museum's broader mission statement for guidance, whether research-oriented or not. Worse, all three of these strategies had to be utilized at different times, which led to occasional conflicts. With the first strategy, there were challenges making new methods stick. The excavators' training in the old methods was deeply ingrained, and they struggled to adapt, especially when the new methods were developed in a somewhat piecemeal fashion. However, the second strategy – sticking with old methods out of tradition or in the name of consistency – was equally frustrating. Excavators were left haunted by the worry that they were collecting the wrong data, or wasting their time and effort, and potentially lowering the future usability of the collections as a result.

And finally, the third strategy led to the RMs feeling like they were working in a theme park, rather than a scientific research project. The broader museum goals are as focused on public education as they are science (if not more so). As described above, the excavation site is in the middle of a public park, and is on view to the public. Over the last five years, the NHMLA's exhibit department has put more and more time and funding into developing educational materials and public programs around the excavation sites as living exhibits. The excavators were asked to be on full view to the public for the entire seven-day week. They staggered their shifts and their breaks to make this possible. However, this focus on being visible could make them start to feel their work was more "Jungly Cruise-y" (RLB-Exc-4) than scientific – there for show, not for research. One of the collections managers, when discussing the growing emphasis on the excavation-site-as-exhibit, asked, "is Project 23 just a public program? If that's the case, well then, we don't need to take any measurements, or do anything! You just have somebody fake-scavating!"[23] (RLB-CM-1). In other words, why bother doing the hard parts of scientific work if the point of it is simply the superficial performance of science?

---

[23] Fake-scavating = fake excavating. This word originated when I was on staff, if not before. Sometimes people from local TV stations or networks like the Discovery Channel want to film programs about the tar pits, and often

Resource manager values

It is important to note that many of the RMs' frustrations described above arose fundamentally because the RMs were highly invested in their work, often on a personal level. One RM, who had volunteered at La Brea as a student, came back to La Brea after working at other sites for years, partially because he had "always had in the back of my head that at some point late in my career, that I would come back" (RLB-CM-2). Several other staff members stayed at La Brea despite other opportunities elsewhere. Overall, the RMs considered the site an important natural treasure, and they considered it a privilege and honor to do the work that they did.

This personal investment could at times heighten emotions around otherwise staid topics (for instance, the decision to use one data collection method over another, whether or not curatorial workflows are sufficiently efficient, data quality). However, any conflict that arose from this erupted at least partially because everyone at the site cared about the site – about preserving it, about understanding it, and about making it well known. The RMs wanted to contribute not just to a small research community, but to scientific knowledge as a whole.

This investment in their work was accompanied by a corresponding investment in collecting high quality data. However, the RMs sometimes defined "high quality" in different ways. Several thought of it more in terms of collecting all the data they could from the excavation site – including extensive collection of measurement data, even for fossils that would not necessarily merit measuring (this workflow is described further in Chapter 5). Others, however, felt that it meant making sure to extract fossils in a way that would minimize damage to the bone.

That said, sometimes the desire to collect high-quality data conflicted with more immediate goals and constraints at the site; consequently, the RMs valued *efficiency* in their processes. They wanted to collect all the data they could – but not more than they had physical space to store, time to collect and curate, or money to fund. Excavation, preparation, and curatorial processes

---

they find that the process of real excavation is too slow or subtle for the camera, so excavators are asked to do what we called "fake-scavate" for the sake of the cameras – brushing away already loosened matrix, chiseling at grids that don't have any fossils in them, etc. This made-up word might seem silly – but it does point to a subtle and pervasive frustration at conflicts between the research and education missions of the museum.

needed to preserve as much of the physical specimens and their associated data as possible – but also as efficiently as possible. The collections managers were particularly concerned with making their collection and curatorial processes more efficient partly because they inherited a huge backlog of uncataloged or unprepared fossils that accumulated through decades of understaffing, and they did not want to make it worse for future generations. One described their ethos as, "Work as you go, and don't create a backlog. That is my goal" (RLB-CM-1).

Preventing future backlogs will likely become even more important – and challenging – in future years, because the collections and excavation projects are only expected to increase. The Los Angeles Metro is preparing to extend the Purple line down Wilshire; subway construction a mile to the east has already uncovered fossil deposits containing fauna as diverse as mammoths and sea lions, and fossil discoveries will likely only increase the closer the subway gets to La Brea. This construction has motivated quantitative estimates of how much it costs to collect, prepare and curate fossils per square foot of a deposit: conservatively, $16,000 per cubic foot (Figure 4.9). Given that there is an additional 4,500 cubic feet of sediment remaining to dig through in Project 23, streamlining processes wherever possible is of paramount importance.

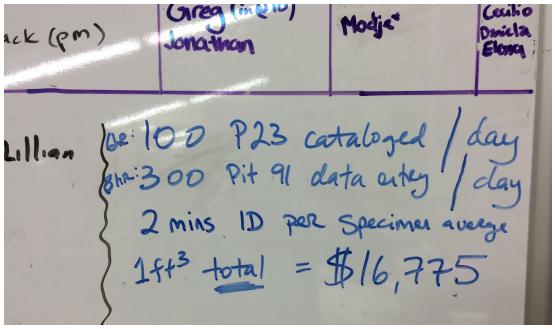


Figure 4.9. The collections managers' estimate of the rough cost of collecting, cataloging, and curating fossils per square foot of deposit.

Resource manager data needs

Where the YNP resource managers suffered from a lack of data, the RLB resource managers struggled with an overabundance. However, they were easily able to summarize what data was most important to them and their work:

Inventory. The collections managers in particular, but RMs overall, said that simply "*knowing what we have"* is of high importance. Knowing how many fossils they had and of what type (taxon and element) made it possible for them to prepare for future work, plan for future storage

needs, and understand what fossils are best fit for what research uses. They were able to meet this need somewhat by digitizing excavation field notes regularly; each excavator was required to input his or her field data (the fossil measurements and their field identifications) into a spreadsheet, which was then imported into the KE EMu database. These field identifications act as a tentative inventory of their holdings until the fossils are formally cataloged:

> One thing that's cool [with the Project 23 data] is that you can actually go in there, and you know regardless of what the identifications, you have at least a sense on what's come out, and what needs to be done, so you can plan, right? You can plan on what to do. With Pit 91 specifically, unless we do something like that, we can't even come up with a plan to attack it strategically or -- you know, in such a way. (RLB-CM-2)

Thus, there was still a need to digitize past years' field notes describing fossils that had not yet been cataloged.

Tracking process and provenance. The RMs also felt a growing need to know how fossils were prepared and treated in different solvents and processes. In the past, the lab was run by one person who simply remembered everything, and used sufficiently consistent methods that she didn't feel the need to document what was being done for every single fossil. When projects *were* documented, they were on small note cards that have subsequently been lost. Consequently, any record of fossil preparation methods are now few and far between. Since she retired, and now that the prep lab were using a broader range of methods, the RMs found that they needed to track what they were doing in a structured and stable way (e.g., in a database), and create retrospective documentation for much of the existing collection. This became particularly important given recent interest in the collection by researchers doing proteomics work; they found that certain preparation methods – and even certain storage methods – may impact protein structures in the bones:

> When I showed [the proteomics researcher] the drawer of insects, he was like "Oh, are those gelatin capsules?" I said, "Yeah." He goes, "Pig protein." So, those are going to be problematic when we try and get proteins out of the insects… And, you know, he said "Don't change your practice right now," and he said it's just important for us to know, and think about this, because [proteomics] is such a young field. (RLB-CM-1)

This same proteomics researcher also told the RMs that some of the new matrix processing methods being used may impact protein extraction, but it was unclear how. Thus, it remains important that these different treatments be documented clearly for future use.

Connections between data. Much of the La Brea data is distributed through different data systems and information structures: there are multiple catalog ledgers, multiple databases and spreadsheets, even multiple strategies for organizing physical specimens. There's additionally a great amount of relevant data published in academic journals or to repositories such as Dryad that the RMs would like to link back to the relevant specimens within the collections. The RMs did not necessarily need all this data stored at La Brea, but (like the YNP RMs) they said that they needed connections between all these data points. Ideally, this could be done through consistent use of catalog numbers in all databases and publications. (One CM described the catalog number as the main "pivot point" around which all accessory data was organized in her database.)

Geological context. Almost all of the RMs said they wished they had more geological context about their current excavation or felt that they should be collecting more data about the geological stratigraphy, using best practices for geological fieldwork as they excavated. This would entail doing things like sketching "strat columns" – drawings of the stratigraphic layers in each box. The current archeology-style excavation method makes this extremely difficult; because they excavate grid-by-grid, an entire column is rarely exposed and the "wall drawings" taken of each grid are often not detailed enough or consistent enough to be useful. New methods (taking cores from each box; excavating in larger grids; using photography) may be needed to collect this data. RMs also indicated that good photographs of the boxes in their original context would be helpful (though these can be challenging to take; the dark colors of the asphaltic sediments do not always show up well on film).

Radiocarbon dates. Radiocarbon dates – analyses of the ratio of carbon isotopes in a bone that can be interpreted to estimate the age of a bone – have long been used to date the age of the pits at La Brea. These dates are also critical to understanding how long each "pit" was open. Constraining the ages of each deposit is critically important for research use of the La Brea collections, particularly for researchers interested in questions of evolutionary biology or ecological responses to climate change (the two most prominent uses of the fossils at La Brea). In the past, radiocarbon dates were generated by researchers using the fossils for their own projects, and published in journal articles. However, the museum is increasingly getting its own radiocarbon dating done, and storing that data in its database. Further collection of these dates,

and correlation of the dates of the bones with their location within each pit, will be important for the La Brea RMs going forward.

Experience with information structures: excavation protocol concerns

As described above, La Brea has a very detailed data collection protocol that almost all of the RMs felt needed changing in some ways – however, they were not quite sure how. Notably, it was unclear if the measurement system even captured anything worth measuring. Prior attempts to plot the data *en masse* had not been entirely successful. One collection manager described a recent GIS project, in which

> [A researcher] tried to put [the measurement data] in a GIS program … she had a huge problem with just a black cloud of points, because it's too packed with stuff. So, there's no signal at all. (RLB-CM-1)

The collection manager granted that this might not be the fault of the measurement system or the software but might simply be that they did not yet know how to parse the data efficiently ("it's extremely difficult when you have 10,000 bones in a pattern of 2-meter cube to take all of that information, and tease it apart" (RLB-CM-1)), but frustrations and concern about the utility of this data remained.

The CMs tried to change the measurement system in small ways -- both to make it more efficient and safer for the bones, but also to bring it in line with broader paleontological best practices:

> I mean this is just a fossil locality. It's a very rich one. It has some interesting preservation clearly. But, there are lots of different types of fossil localities, and people who deal with peat bogs have their own little unique situation, but they're also another locality, and permafrost is another, and a cave is another, and a -- you know. So, we are no different from another really rich locality. We're unique because of all of those circumstances, and we're in the middle of a city. But, on the larger scale, in order to keep this relevant to the world, I think people need to keep up with best practices. (RLB-CM-1)

However, the changes were introduced piecemeal and the excavators were often confused about what "best practice" was actually considered best for a given situation. One said,

> It's hard for me to understand what exactly to do in all cases. So, again, it's one of those things where I'm like okay, I'm measuring everything that's big. No, am I measuring only things that are [identifiable]? What about things that I'm not sure if they're [identifiable]? (RLB-Exc-4)

Often, they fell back on "keeping with the tradition of Pit 91 just to keep consistent" (RLB-Exc-5). This was both because that was the system they were trained in and knew best, but also because they had been "raised," so to speak, to value the preservation of all data. The excavators were extremely aware that excavation is a destructive process, and they feared losing information:

> If we don't write it, it doesn't happen, which is why I think a lot of us kind of [laughter] go a little bit crazy, because we want to like preserve all of the information that we possibly can. And then, we're kind of like hit back with, you know, either "This is a salvage project," or "This is not what you need to be doing." (RLB-Exc-2)

Consequently, the measurement system was a large source of stress for the entire Resource Management staff.

The measurement system also had some interesting effects on the accessibility of the collection overall. The Pit 91 collection was not only collected in a different way from other La Brea collections, it was cataloged and stored in a different way. This is described further in the "Researcher Perspectives" section below, but the Pit 91 fossils are stored in a way that makes them extremely difficult to browse and retrieve:

> The reason why like, you know, the Hancock part of collection is heavily used is because it's well curated. Pit 91 isn't used that much, because it's curated numerically, it's not curated by taxon, and it's a struggle for people to use that collection. And not all of it -- even what's there, because it's not in the database, right? So, there's like 11,000 catalogue records from Pit 91 that are not in the database. (RLB-CM-2)

Researchers typically need to access fossils according to taxon, type, and location. However, the Pit 91 fossils are cataloged and stored in the order in which they were prepared. This means that if a researcher were looking for, say, 15 horse fossils from one area, she would first need to look up those fossils in the catalog, and then pull them out of likely 15 different drawers. The Pit 91 collection was only truly accessible with the assistance of a computer – and yet this computer system still has never really been built.

4.2.3. Researcher perspectives at La Brea

Researcher roles

I interviewed 12 researchers who worked with the collections at RLB, including six paleontologists; three paleoecologists; a researcher conducting proteomics studies with the fossils; one GIS specialist; and, coincidentally enough, one geobiologist who consulted in the

development of a new matrix processing workflow. One paleontologist and one paleoecologist were graduate students completing their dissertation research at La Brea; and the GIS specialist had recently completed her master's work at La Brea. The remaining nine researchers were all faculty members based at research institutions throughout the United States. Three participants (a paleontologist, a paleoecologist, and a paleoecology graduate student) were collaborators on an NSF-funded project about to start work at La Brea. Two other paleontologists had collaborated in the past on a large scale taphonomy study. However, it is worth noting that even the researchers who had not directly worked together knew of each other and of other researchers at the site; additionally, most of the group had co-authored with the current collections managers, or the previous collections manager, laboratory supervisor, and curator.

The researchers were also extremely complimentary of the work that the La Brea RMs did. They all spoke quite highly of the current collections staff, and recognized the challenges they faced in collecting and curating such a large collection. They all understood that the collections staff had to, at times, make strategic decisions about what to keep and what to put aside that would not necessarily be up for discussion at other sites. As one put it,

> It's this huge problem the museum has. So what do you do when you have 4000 dire wolves, you know? What do you do if you have your four-thousand-and-first dire wolf? And which ones are the most important? And there's this whole set of issues around a museum with that rich of material from a single site. (RLB-Rsch-6)

Thus, just as the La Brea resource managers wished to defer to researchers' expertise about what data to collect and prioritize, the researchers wished to defer to the collections staff about how best to curate and manage the materials.

Whereas my discussions with the YNP researchers focused around what data might be reusable, my discussions with La Brea researchers centered on learning about their individual studies, and what information they needed about a fossil or excavation to make a specimen and its associated digital data fit for use.

Researcher values

Researchers valued La Brea not just because of the sheer size of the collection (though that was certainly a factor), but also because of its nature as a *lägerstatte:* the collection represents an entire ecology from an important slice of time. La Brea captures evidence of "a time that's so

critical, where you're having pretty dramatic climate change right before humans come into the picture, and then when humans are [present] at the very end" (RLB-Rshc-9). This time window may inform understanding of the impact of climate change both with and without human presence. The paleoecologists particularly felt that La Brea was important for its fossils' taxonomic variety (e.g., animals, vertebrates, invertebrates are all represented) as well as volume:

> You just have so many different components of the ecosystem that are available, and preserved in one place. It's as close to a whole snapshot as you could ever get of the past. And it's not just one instant in time, like sometimes there's like, you know, an eruption or something that might preserve like a day in the geologic past, but it's a really large window on these ecosystems for tens and thousands of years. (RLB-Rsch-10)

Having the fossil record of an entire ecology makes possible two kinds of studies not typically feasible in paleontology: population-scale studies about different dynamics within one species, and ecosystem-scale studies about the relationships between different species. Population-scale studies can range from the broad characterization of variability within a species, or can include the investigation of phenomena that "happen at lower frequency" (RLB-Rsch-7) and just don't show up with fewer specimens. For instance, there are a large number of pathological bones within the La Brea collection – bones that show evidence of injury or disease such as arthritis, scoliosis, healed bone fractures, or tooth wear. Because the deposits are representative of entire populations of animals, the collections can be used to estimate the prevalence of these pathologies in species overall, and can be used to inform our understanding of animals' behavior. Ecosystem-scale studies focus on topics such as species interactions, food webs, and niche evolution over time. These paleoecological studies just aren't possible outside of *lägerstatten*, and La Brea is, again, a particularly rich *läggerstatte*.

La Brea's collections are also used in more typically paleontological studies of animal morphology, evolution, and taxonomy, particularly of its many rare megafauna (saber-toothed cats, dire wolves, the North American lion, several species of extinct North American horses, giant ground sloths, bison…). Work in this vein continues, though it tends to focus less on basic taxonomy and more on newer techniques such as morphometrics or isotope studies. As one researcher noted, "the preservation of the bones is wonderful," (RLB-Rsch-7) which makes them quite attractive to researchers studying fine morphological details.

Finally, researchers attributed the benefits of working at La Brea not just to its natural features, but also to the human and research infrastructure built around the site. It was not just that there's a wealth of materials available, but that there are people employed there full-time to help make them accessible.

> To me what is so special about La Brea is an active excavation with the museum tacked on to one side, and full curatorial team -- or almost a full curatorial team. And people who are working on fossil preparation, are all in the same place. Ordinarily that's not the case. (RLB-Rsch-6)

Another researcher said,

> what's interesting is that you actually have a lot of people who are working on many different aspects of the system. And because it's associated with a museum, there can be some coordination with their different efforts, right? There might be other sites where, like, there are lake sediments in Eastern North America where you can go, and anybody can go and take a sediment core from the lake. So, you might have someone take one sediment core from the lake focusing on pollen, and then you have other people take different sediment cores focusing on diatoms or other things, right? There's not necessarily any coordination between them.
>
> And at La Brea, it's not -- there's not necessarily any coordination between different people either, but because it's managed by institution, there's the potential opportunity for more coordination, right? … the La Brea Tar Pits have this institutional structure with the line of sight that should help facilitate coordination amongst different research group. (RLB-Rsch-4).

Thus the significance of the site is not just the fossil deposit, but also the institutional structure that supports research around the fossils.

Experience with information structures and parameters for use

Because the "data" at La Brea is already organized in a collection, researchers must navigate the information structures in which it is stored to use it. I am consequently combining discussion of their data needs and their experiences of protocols, practices, and models in this section, because they are too entangled to isolate (and don't make sense without one another).

Researchers described the following parameters as important for use:

Access to collections. The physical materiality of La Brea's data changes the nature of data use and reuse; researchers work first and foremost with physical specimens rather than digital data. For this, researchers simply need access to specimens and their associated catalog data (the

measurement data, their identifications). Researchers seemed quite happy with the level of access the collections staff provided for them, though several noted that they wished there were an online database that they could query on their own; researchers sometimes needed to check catalog information about the specimens they had worked with, and they disliked having to "bother" the collections managers for this, given that the collection managers were so busy.

Given the importance of having physical access to fossils, the physical organization of different fossils collections impacted their usability. By and large, the Hancock Collection is the most frequently used collection at La Brea – largely because it is simply easier to browse. A paleontologist described the difference in his work:

> The materials from Pit 91 are curated just in numeric order. As they were [prepared] in the lab… then they get their catalogue number, and back they go into collection. [But] the Hancock fossils are organized by animal, and then within each animal, by element, and then within each element, by pit. So, I could go straight to, if I want to look at humeri of horses to measure them, I can go straight to the *Equus* humeri section, and just go pit by pit by pit, and look at, and measure, and document the materials. For Pit 91, I would literally have to open every drawer to hunt up all of the materials." (RLBRsch8)

Three of the researchers I spoke with indicated that they did indeed use the physical materials from Pit 91, but wished that they were easier to access.

Information about preparation history. Researchers echoed the same sentiments described in the "La Brea Resource Manager Perspectives" section: that there was a growing need to understand how, exactly, fossils were cleaned, prepared, and repaired. This was particularly important for researchers doing molecular work (proteomics, DNA sequencing). They did not yet know enough about the effects of the preparation methods on data extraction to recommend one method over another, but they wanted to know how bones were being treated.

Consistently formatted and easily interpretable location data. All of the researchers said that knowing roughly where a fossil was taken from in a deposit was important. For some, a grid number and depth was enough; for others (particularly those interested in taphonomy), the more precise measurement data was crucial. That said, those who worked extensively with the measurement data unanimously agreed that it was difficult to work with in its existing form, and that it could likely be replaced with a simpler and more effective system.

Of the three researchers I spoke with who worked extensively with the measurement data, two were paleontologists who collaborated on a large taphonomic study of Pit 91 in the early 2000s, and one was a GIS specialist who did her master's project mapping some of the bones from Box 1 of Project 23. They all had similar problems working with this data.

First, "it's just a lot of data" (RLB-Rsch-12). There were about 45,000 records in the Pit 91 collection at the time of this researcher's study, and about 3,000 in the Box 1 collection at the time of RLB-Rsch-5's. These data are formatted idiosyncratically enough that they require significant cleaning and manipulation by hand before they can be used in any standard statistical, mapping, or GIS packages:

> The data is kind of sloppy, and it's in some old database forms… So, a lot of what I ended up having to do was a lot of data massaging, just like moving into the different programs and things like that….and it was like a mess, just moving between things. (RLB-Rsch-12).

The Pit 91 data in particular posed a challenge, in part because the measurements were all taken in inches and feet, rather than centimeters and meters. (This was originally intended for backwards compatibility and consistency with the excavations from the 1900s, whose grids and depths were all recorded in feet and inches as well. All the Project 23 measurements are now taken in the metric system.) Not only were the imperial measurements difficult to process and calculate, but RLB-Rsch-12 also found they were difficult to publish: "… when I tried to publish this stuff, like I would put stuff in feet and inches, and like journals like made me change everything [to metric]."

Researchers also struggled to import this data into standard mapping software. The Pit 91 excavation system places its "origin point" in the southeast corner of the grid (typical of archeological digs), whereas most standard GIS and mapping tools place the origin point in the southwest corner of the grid. This means that the measurement data need to be "flipped" from west to east before they can be plotted. Additionally, the data from the grids needs to be merged into one aggregate graph, rather than a series of smaller graphs.

Finally, the researchers all had a hard time using the measurements for their actual intended use – to infer the orientation and position of the bones within the deposit. The coordinates used to record the position of the bones were just fundamentally difficult to parse:

> One thing we wanted to get out of it was orientation of the bones, like so, which way they are pointing, right? Not just where they were in the pit, but also where they were pointing. And that was really hard to do sometimes, because what you would end up having is like data points for the two ends of the bones in some sort of X, Y, Z data, and then I would have to sit there and like mess with polar coordinates and figure out how to change that to information about, you know, basically strike and dip. (RLB-Rsch-12).

Despite the difficulties using this data, researchers were not in favor of stopping collection of the data altogether – rather, they preferred that the system be overhauled. Many researchers said that they wished the Hancock Collection had more reliable location data (even if just more precise grids and depths). One researcher attributed the need for this data to the destructive nature of excavation work:

> Even if nobody ever uses the data, I think it's important to record the data from Pit 91 as precisely as possible, and from Project 23, I should say that as well… Excavation is a destructive process, and once you're done excavating, the thing you're studying is gone, and so you only get the one chance. So, if those data turn out to be fairly meaningless, and you never use them, well, it slows things down. But, if those data turn out to be critical, and you didn't collect them, then you've shot yourself in the foot. I'm kind of hoping that that's not what this is leading to, is somebody saying, 'Hey, we don't need to measure everything in place, it's a waste of our time, because nobody uses the data.' (RLB-Rsch-8)

This same researcher did, however, grant that measuring bones in place from the Project 23 salvage would likely not be useful if the fossils' positions had been altered in the boxing process ("If it's already disturbed, and if the data are going to be clearly meaningless, then yeah, why record them?" (RLB-Rsch-8)).

Radiocarbon dates correlated with locations. Researchers broadly agreed that radiocarbon dates tied to precise geolocations within each deposit were critical for their work. This is related to the need for precise location data described above. One of La Brea's many unique features is that the fossils are not necessarily found in neat stratigraphic layers, as at many other sites, but rather in a jumble. This makes it difficult to infer the age of fossils through surrounding stratigraphy, or from other nearby fossils, and means that further work is needed to understand the chronological structure of the sites. The two paleontologists who worked on the Pit 91 taphonomy study described this well:

> So, the problem with the tar pits as a fossil [site] -- typically in a fossil deposit, the bones at the bottom of the deposit are older than the bones at the top, because it's a layer cake

> stratigraphy kind of thing. In the tar pits, it's unclear, or was very unclear whether that was the case, because the asphalt is kind of molten[24]. And so, the bones, all the skeletons are entirely, almost always entirely disarticulated. (RLB-Rsch-7)

Through the course of their work they found that,

> The law of superposition didn't hold, so we sometimes found older bones on top of younger bones. And that raised the question of why that might have been, were bones from different collections mixing in together, was there any sort of churn within the pit? Was it happening up at the surface with some sort of trampling and things like that? (RLB-Rsch-8)

It's still unclear just how "mixed" the fossils are; future taphonomy studies will be needed. However, it is clear that understanding the stratigraphy, depositional environment, and geologic age of each pit (and sections within the pits) will be important to everyone who does work at La Brea. What is also clear is that the radiocarbon dates are important enough that they merit special preservation, and need to be broadly accessible to all researchers. This is not necessarily the case currently; radiocarbon dates are published in journal articles, and while the La Brea collections managers have aggregated them all in a spreadsheet, they are not publicly available (without requesting the spreadsheet from the collections manager, that is).

Ability to see other researchers' data about individual specimens. In addition to radiocarbon dates, researchers said they'd generally like to see what data had been taken from individual specimens, and would like to be able to quickly look up publications in which a specimen had been mentioned. In short, they would like to see the results of their research integrated into the La Brea collections catalog.

> There have been a lot of different researchers who've gone in, and done different studies…. those data aren't -- they might be available in some off-site database… but that data's held nowhere at La Brea. Also, when different people study different -- like take measurements on fossils, a lot of times, those data, those individual specimen data aren't published… Well how cool would it be if I looked at those same ones, that [another

[24] When I was an excavator at La Brea, I spent a lot of time speaking to school groups. This past experience compels me to make a small correction here: "Molten" is not accurate– the asphalt isn't heated, it's just very mobile and viscous. Thanks to the movie "Volcano," members of the general public often ask if there really is a volcano beneath the La Brea Tar Pits. There is no volcano. The asphalt is not lava, and it is not molten.

> researcher] had done all these measurements of, and was able to, you know, use those data, right? (RLB-Rsch-9)

Aggregating this data would certainly be challenging – both because it's distributed in hundreds of publications, and because in the past, paper journals simply have not had the capacity to publish all the data from large studies.

It is worth noting that researchers broadly supported the idea of data publication or sharing; they all said they would be happy to share their data with the collections staff, and most of them published their data to repositories such as Dryad or Neotoma. Additionally, one team of researchers described collaborating with the La Brea RMs through the use of Google Sheets. The researchers were given a can of bulk matrix to sort through; La Brea RMs would assign catalog numbers to specimens as they were identified. This made it possible for all stakeholders to meet their needs at the same time; researchers were able to do their work, and RMs were able to catalog fossils with stable persistent identifiers.

4.2.4 La Brea Case Summary

Summary of La Brea resource manager perspectives

In summary, the La Brea resource managers collected much of the data that they needed for their work – but they struggled in getting that data into a usable and accessible format. They primarily valued data that helped them manage the fossil collections and plan for future excavations and collections management needs. They additionally valued data that helped them contextualize their fossil collections in time and space – particularly geological data, and radiocarbon dates associated with specific fossils and regions of the deposits.

La Brea resource managers additionally valued efficiency in the data collection and curatorial processes. This included efficient use of a range of resources, including time, money, and space. However, efficiency must be balanced against the long-term well-being of the fossils and their associated data.

La Brea resource managers interacted with information structures at a number of different manifestations: via the various excavation and data collection standards they used, and in the form of the different information systems they used in their work. These standards, practices/traditions, and systems have a complex relationship that needs to be taken into account for any changes to curatorial processes or other data practices.

Summary of La Brea researcher perspectives

La Brea researchers valued La Brea as a site because of the size, diversity, and breadth of its collections and the range of studies that they support. They noted that many kinds of research are possible at La Brea that simply are not possible at other paleontological localities, particularly ecosystem- or population-scale studies, and studies of phenomena that happen at lower frequency. Researchers also valued La Brea as a site for its human and curatorial infrastructure; the presence of a full-time data collection and curation staff makes the site much more appealing to work at.

La Brea researchers needed information about fossils' original depositional environment for their work. They additionally needed radiocarbon dates from a range of fossils at each deposit. The location and radiocarbon data were used to infer the structure and age of the deposits, which are crucial to numerous other analyses. La Brea researchers additionally valued having access to information about fossils' preparation history; however, this was not as critical for their work as the geolocation and radiocarbon data.

Finally, researchers said that they wished they had access to the digital collections database, and additionally wished for access to other researchers' data. In some cases, they did not necessarily want to reuse this data, but felt that it would help them contextualize their own results, or verify past findings.

4.3 CHAPTER SUMMARY

In this chapter I have presented case study narratives describing the perspectives of researchers and resource managers at each of my study sites. I contextualized these perspectives in key historical and scientific background on each of my sites. I roughly organized these subsections around four key themes that emerged from my analysis: stakeholders' roles, valued aspects of the sites and priorities, key data needs, and interaction with information structures. These themes are summarized in Table 4.1. In the next chapter, I will discuss my participants' data needs in further detail, and present the minimum information frameworks designed to address those needs.

	YNP Resource Managers	YNP Researchers	La Brea Resource Managers	La Brea Researchers
Roles	Site mgmt., researcher mgmt., some data mgmt.	Data collection, analysis, publishing, reporting	Data collection, data mgmt., research mgmt., site mgmt.	Some further data collection, analysis, publishing
Values & priorities	Ability to strategically plan for future research, awareness of research projects, efficiency, not having to police people	Accessibility of the site, ability to complete work independently without excessive oversight, precision	Efficiency, avoiding backlog, protecting fossils & other resources	Fossil collection breadth, diversity and accessibility
Key Data	Awareness of research activities, "larger context" of research activities (methods, data management plans), connections between silos	Genetic data, basic geochemistry about hot spring environment, photographs, information about methods	Collections inventory, process and provenance information, geological context, radiocarbon dates	Radiocarbon dates; reliable and consistent location information, information about prep history, access to other researchers' published data
Interaction with information structures	Constrained by use of IRMA and ICMS databases, and by NPS reporting guidelines	Wary of data standards; concerned about ability to integrate data	Constrained by legacy information organization systems; feel there are shortcomings to data collection protocol	Access to fossils somewhat constrained by legacy information organization systems; willing to share data as requested

Table 4.1. Summary of key case study themes

5. SITE-BASED INFORMATION FRAMEWORKS

By taking participatory approaches with both of my case studies, I sought to answer my research questions through the process of developing and contributing protocols or best practices (or revising existing protocols or best practices) for the collection and curation of data from scientifically significant sites. In this chapter I present the results of this collaborative work: namely, minimum information frameworks developed at each of my sites. I also present the results of workflow analysis that contributed to the development and evaluation of these frameworks. Finally, I present a more generalized model of site-based information, and compare this model to several existing ontologies/systems that use site-based data representation.

5.1 DEVELOPING MINIMUM INFORMATION FRAMEWORKS

The development of (and terminology describing) minimum information frameworks is rooted in work conducted through the SBDC project, in which we (the SBDC team) worked with site stakeholders to develop "a framework of principles and processes that helps to articulate and support upstream and downstream processes as a general model for site-based data curation" (Palmer et al., 2017). The term "framework" came to refer to a minimum information framework[25] developed through work with YNP stakeholders: a high-level information model describing the minimum elements needed to effectively curate data for reuse at the site. I likewise developed a minimum information framework (MIF) for La Brea; these MIFs represent the core of this chapter.

The term "minimum information framework" is intended as a nod toward and invocation of the many "minimum information" reporting standards in the biosciences (e.g. Brazma et al., 2001; efforts towards synthesis discussed in Taylor et al., 2008). Minimum information standards take a "checklist" approach to reporting, which aims to provide researchers with guidelines about

[25] I recognize that my use of the term *framework* could be confusing here, given that I am using it to refer to high-level information models. I use this terminology throughout this dissertation for the sake of consistency with prior work and recent journal publications. Given that a framework can be defined as "a basic structure underlying a system, concept, or text," I feel that this terminology is still broadly appropriate.

what information should be reported rather than prescribe how data should be collected or created. We felt this was an appropriate data reporting philosophy given some participants' (particularly the YNP Researchers') concerns that data standards could potentially constrain their choice of methods artificially. I additionally found that framing discussions around what "information" they needed for their work, rather than data or metadata, helped focus conversations with participants by avoiding potentially distracting conversations defining or debating the distinction between data and metadata.

As outlined in section 3.3, I use methods adapted from systems analysis to augment my framework analysis and development. This approach is also rooted in work conducted through the SBDC project. Specifically, we used process modeling techniques to make the nuanced tasks in complex research workflows more explicit, and information modeling techniques to reveal the relationships between specific research artifacts and broader classes of information.

Process models were created with key collaborators in each case. In the YNP case, this work was conducted with geobiologist Bruce Fouke, the co-PI of the SBDC project, who allowed us access to ten years of his field and research data from YNP. At La Brea, this work was conducted with the two collections managers who served as my primary informants and collaborators in the case. Though Fouke is part of the Researchers stakeholder group, and the collections managers are part of the Resource Managers group, both process models focus on the *data collection and fieldwork processes* (which is possible because of the unique work arrangement at La Brea, in which Resource Managers are responsible for data collection).

[likely need to tie to this research questions – the process of developing these frameworks reveals what aspects of sites stakeholders value (RQ1); and how stakeholders reflect their site structures through data structures and curatorial processes (RQ2). Further, analysis of data collection, curation, and analysis workflows reveals how stakeholders negotiate their respective needs of data collections at scientifically significant sites (RQ4a).]

5.2 Minimum information framework for geobiology at YNP

As described in chapter 4, the YNP stakeholders identified several categories of data that were important to the reuse and curation of geobiology data from the YNP hot springs, initially through focus groups and activities at the Spring 2013 workshop. Resource managers were

broadly interested in data that would give them awareness of current research activities, "larger context" of research activities (methods, data management plans), and connections between silos. Researchers, on the other hand wanted somewhat more succinct categories of information: genetic data; basic geochemistry; photographs; methodological context; and environmental context.

A preliminary version of a more detailed list of important information elements was drafted by two of the geobiologists immediately following the workshop, and then revised through analysis of the workshop outcomes (see Thomer et al., 2014) and through consultation with key stakeholders throughout the remainder of the SBDC project. This list was formalized into a Uniform Modeling Language diagram for journal publication (Palmer et al., 2017), and is presented below. The text describing the MIF elements, the relationships between classes, and the encoding guidelines are also from Palmer et al., 2017.

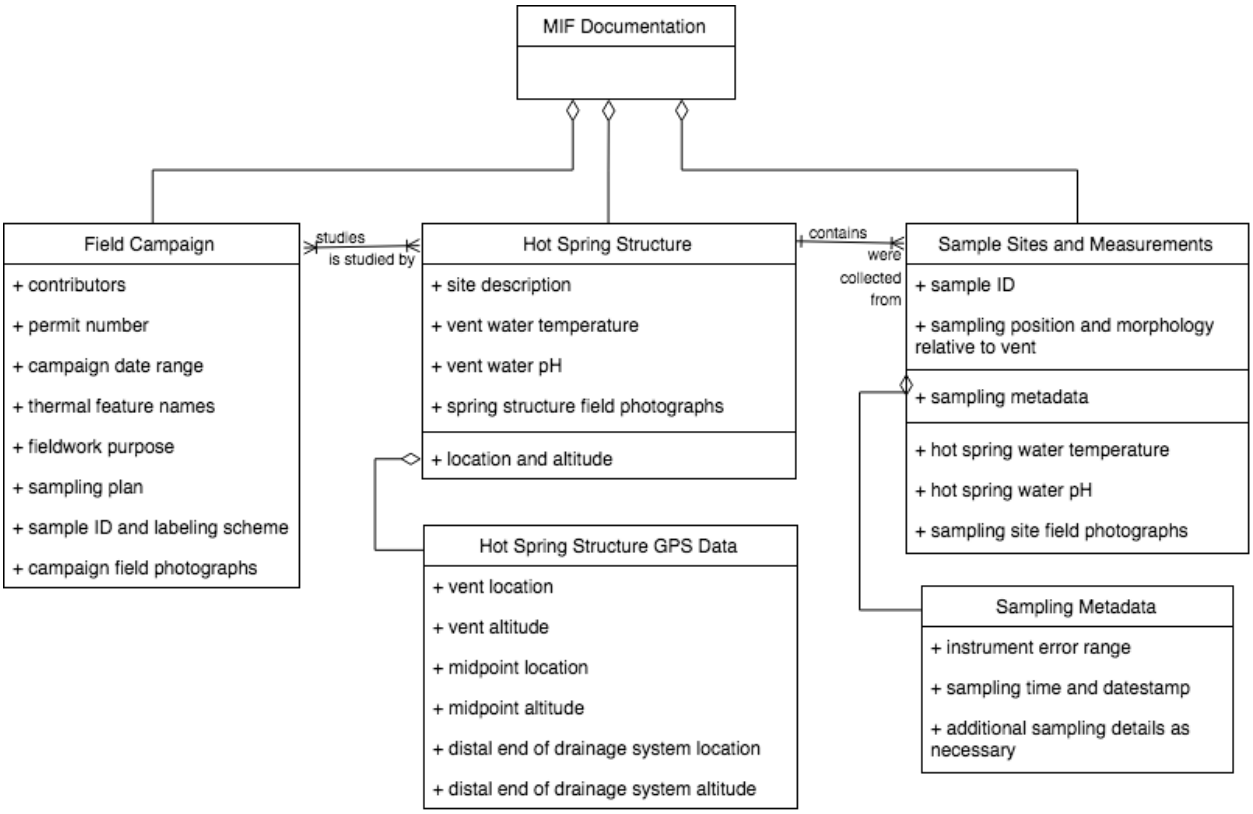


Figure 5.1. Class Diagram showing the three classes of the Minimum Information Framework (MIF): Field Campaign, Hot Spring Structure, and Observations from Sample Sites.

The MIF is composed of three classes of information (Figure 5.1):

1. The Field Campaign class. This class provides basic information about the project's goals and mission, the springs being studied, and the people involved in a project. Elements include:

- *fieldwork purpose,* a description of the overall goals and motivating hypotheses of a projects;
- *sampling plan* for data collection;
- explanation of the *sample ID and labeling schema* (e.g., explanation of any codes used to label samples);
- *thermal feature names,* the names of the hot springs being sampled (specific geolocations of the springs are required as well but documented under a different class). For sites at YNP, use of the "official" location names and ID numbers found in the NPS Thermal Inventory is recommended if known, though we note that corrections may be needed depending on changes in thermal feature activity over time;
- *NPS permit numbers*;
- *names* of all project *contributors,* including the full name(s) of PI/investigators and all members of the field party; and
- *date range* of the project.

Additionally, large-scale photography of the entire hot spring system should be included here.

2. The Hot Spring Structure class. This class includes information that describes and characterizes each hot spring within the study. Special focus is paid to the vent of the hot spring, which serves as the triangulation point for all of the outflow drainage system. Elements include:

- The *temperature* and *pH* of the water at the vent; these data function as a heuristic characterization of the hot spring's microbial ecology (see Fouke 2011);
- a *site description* in free text (accompanied by sketches if necessary). These should detail the overall site and condition of the vent and sampling sites, as well as information describing the primary flow path in a range of methods (e.g., sketches, free text description, estimated size of spring, etc.); and
- the *location and altitude* of the *vent*, the *mid-point* of the drainage system, and the *end-point* of the drainage system. This array of points would allow the geometry and flow directions of the entire system to be identified and reconstructed.

Detailed *photographs of the entire spring system* are included to clearly illustrate the spatial hydrologic continuity between the vent and the outflow drainage channel.

3. The Sample Sites and Measurements class. This class includes information about each site of sample collection. Elements include:

- *sample ID,* the label assigned to the sample or measurement, critical for capturing the provenance of future analyses;
- *sampling position,* which should describe the position of each sampling site in terms of the distance and bearing to the vent, and the *morphology* of the lithographic facies in which the sampling site is located (e.g., the pond, apron, or channel of the hot spring drainage system; see Fouke, 2011). We note that a description of distance and bearing is necessary in lieu of a simple GPS location because the distance between sampling sites (often a few centimeters to meters) in these springs is frequently too small to be recorded accurately by GPS; and
- any relevant instrument-specific *sampling metadata, which* should also be recorded (e.g., error ranges for thermometers, date and timestamps of measurements). The sample site's water *temperature* and *pH* should be collected (preferably in triplicate) along with each sample.

Detailed *photographs* at a range of scales should be included, to clearly illustrate the spatial relationships and hydrologic continuity between each sample collection site and its position in the vent and the outflow drainage channel and collection sites therein. All photographs should be taken at a range of scales for each class of information (mm-cm-m length), preferably with embedded geolocation and timestamp information.

Relationships between classes

There is a many-to-many relationship between Field Campaign and Hot Spring Structure classes (many field campaigns can study many hot springs and many hot springs can be studied in many field campaigns), and a one-to-many relationship between Hot Spring Structure and Sample Sites and Measurements. These relationships are meant to represent the real-life spatial relationships between samples and sites (see Figure 5.2).

The GPS data is split out as a subclass of Hot Spring Structure location and altitude data because of their prevalence in earth science datasets and the existence of numerous broadly adopted standards in recording and sharing GPS coordinates. Sampling Metadata is split out as a subclass of the Sample Sites and Measurements class because of its great variability; it is anticipated that researchers seeking to apply the MIF to their work would need to customize this subclass to their study and instrumentation. Researchers may wish to include information on sampling technique and experimental design, measurement units and uncertainty, and instrument detection limits in this subclass.

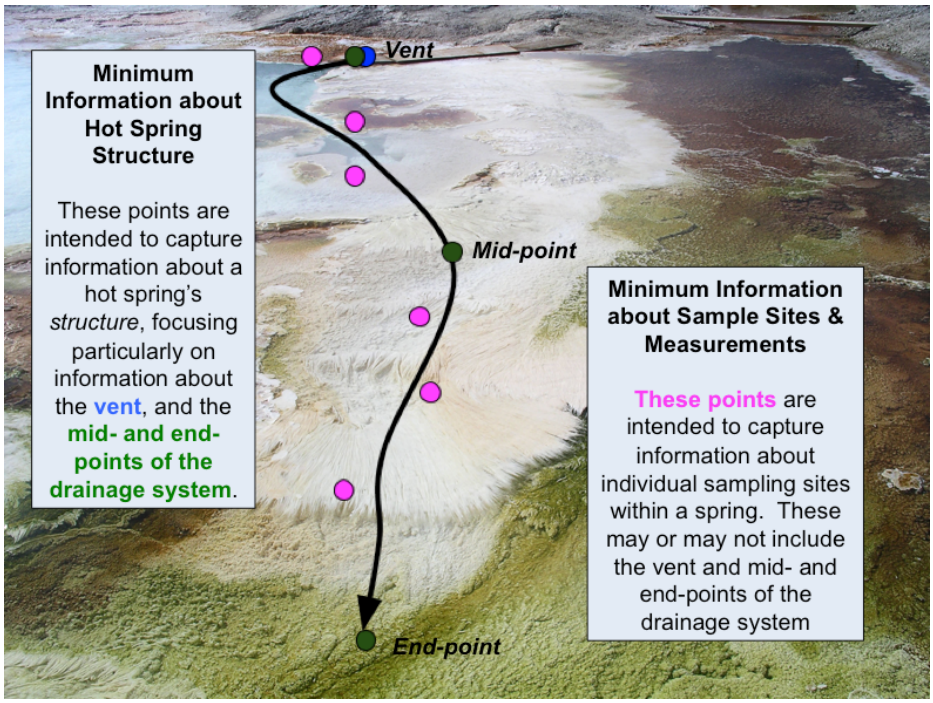


Figure 5.2. Illustration of geospatial relationship between information collected in the Hot Spring Structure class and the Sample Sites and Measurements class. Figure from Palmer et al., 2017.

Encoding guidelines

The MIF is intended to guide general documentation practices for sample and data collection for geobiology; further work will be needed to develop data capture mechanisms. As such, the UML class diagram shown in Figure 5.1 should act as a starting point for developing documentation and information management practices appropriate for a given project. The classes and attributes should be refined based on a researcher's sampling plan and instrumentation. Additionally, research communities will need to coordinate among themselves to develop systematic encoding standards for MIF elements. That said, we have developed initial recommendations as follows.

Contributors should include the full name(s) of PI/investigators and all members of the field party. *Thermal feature names* lists the names of locations where sampling was conducted. For sites at YNP, use of the "official" location names and ID numbers found in the "NPS Thermal Inventory" (an official list of geothermal feature names maintained by the NPS) is recommended if known (though we note that corrections may be needed depending on changes in thermal feature activity over time). GPS locations should be recorded in UTM if possible. Altitude measurements should be represented in meters, and derived from topographical maps or separate GPS systems. *Photographs* should be taken at a range of scales for each class of information (mm-cm-m length), preferably with embedded geolocation and timestamp information. *Site descriptions* should detail the overall site and condition of the vent and sampling sites, as well as information describing the primary flow path in a range of methods (e.g., sketches, free text description, estimated size of spring, etc.)

Controlled vocabularies need to be applied when possible. As noted there are official names for YNP geological features (e.g., Angel Terrace), but sub-feature terms, such as the names of the facies (e.g., pond, apron, channel, etc.) along the spring drainage system, should also be consistently applied. This may involve development of a local controlled vocabulary. Other controlled vocabularies may need to be adapted or extended. For example, terms from the Biodiversity Collections Ontology (Walls et al., 2014) may be appropriate for description of sampling plans and locations, and terms from the growing National Environmental Methods Index ("National Environmental Methods Index," n.d.) may be useful in describing sampling methods. For some sampling methods, a simple description of the kind of instrumentation used may suffice, such as "by paper" or "by instrument" for pH measurement.

Assessing the efficacy and feasibility of the MIFs involved first a) understanding current reporting mechanisms, b) analyzing current data collection workflows, and then c) how and where information elements are or could be collected, and thereby identifying potential points of future upstream data curatorial intervention.

Evaluation through EarthChem templates

The MIF was also compared with an existing earth science data publication standard, the EarthChem "Vent Fluids" template. This template was identified from a list of relevant initiatives and platforms developed through stakeholder analysis and a survey of the literature.

As described in Gordon et al., 2014, we attempted to migrate one of Fouke's water chemistry datasets into the template. We found that the EarthChem template could be adapted for geobiology work with fairly minimal additions of geobiology-specific terms (for instance, descriptions of the facies of rock the water sample was taken from). However, we additionally found that further identifiers were needed to link field samples to other derived data, such as genetic sequences. This work with the EarthChem templates verified that the MIF elements could be captured with only moderate alterations to existing infrastructures.

EVALUATION THROUGH WORKFLOW ANALYSIS

As stated in the beginning of this chapter, one of the goals of the SBDC project was to "develop and articulate upstream and downstream processes for site-based data curation." The MIF only provides us with one piece of this puzzle. Although it enumerates the data elements minimally necessary for data curation and reuse, it does not describe how or when those data elements are or should be collected. A workflow analysis was not included in previous work with Fouke; consequently, in this section I turn to methods from systems analysis to better document and explicate key stakeholders' research workflows – specifically, the fieldwork processes of our key collaborator, SBDC co-PI Bruce Fouke.

To understand Fouke's typical research workflows and outputs, we first conducted an extensive inventory of co-PI Fouke's research hard drive. Fouke maintains well-organized collections of all data from his field work at YNP, as well as archives of his later "downstream" data products. After an initial broad survey of his files, we selected two years of data to study as exemplars of his field processes and created detailed "artifact inventories" of all of these files. Next, we worked with Fouke and his research assistant to document the relationships between each of these files, and in doing so, began to model his typical research workflows and data output.

YNP activity diagram

Fouke's geobiology data work can be roughly divided into three key phases: Planning, Fieldwork, and Processing & Analysis. The Planning stage is presented below as an example of the kind of activity diagrams created in consultation with Fouke (Figure 5.3). Initially, we focused on illustrating his field processes as they were at the time, rather than as they ideally would be after implementation of the MIF. However, we then reviewed his workflow to try to identify points at which MIF elements were already collected – or, if not, points at which they

could or should have been collected. Thus, Figure 5.3 is color-coded to highlight processes in his Planning work that might produce MIF elements.

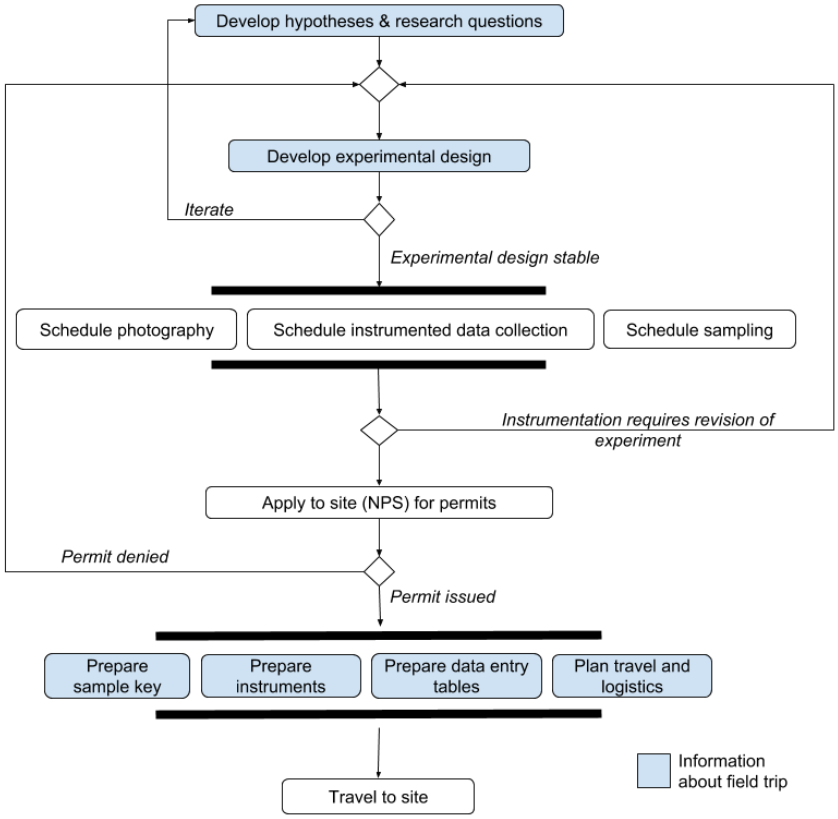


Figure 5.3. The activity diagram is constructed following notation outlined in Dennis, Wixom & Teagarden (2012). Each rectangle represents an individual process that is part of the larger activity. The control flow arrows connecting the rectangles represent the sequencing of the processes. Generally speaking, each process is a discrete set of actions that requires 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 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, "apply to site admin (NPS) for permits" is followed by a decision point. If the permit is approved, the planning and preparation processes begin. If the application is rejected, then the control flow returns to the earlier planning processes that precede the permit application.

Figure 5.4 shows Fouke's Fieldwork phase of research split into two columns. This activity diagram is also color-coded to highlight potential points of curatorial intervention; additionally, I explicitly call out processes where data collectors should take extra steps to collect data recommended by the MIF.

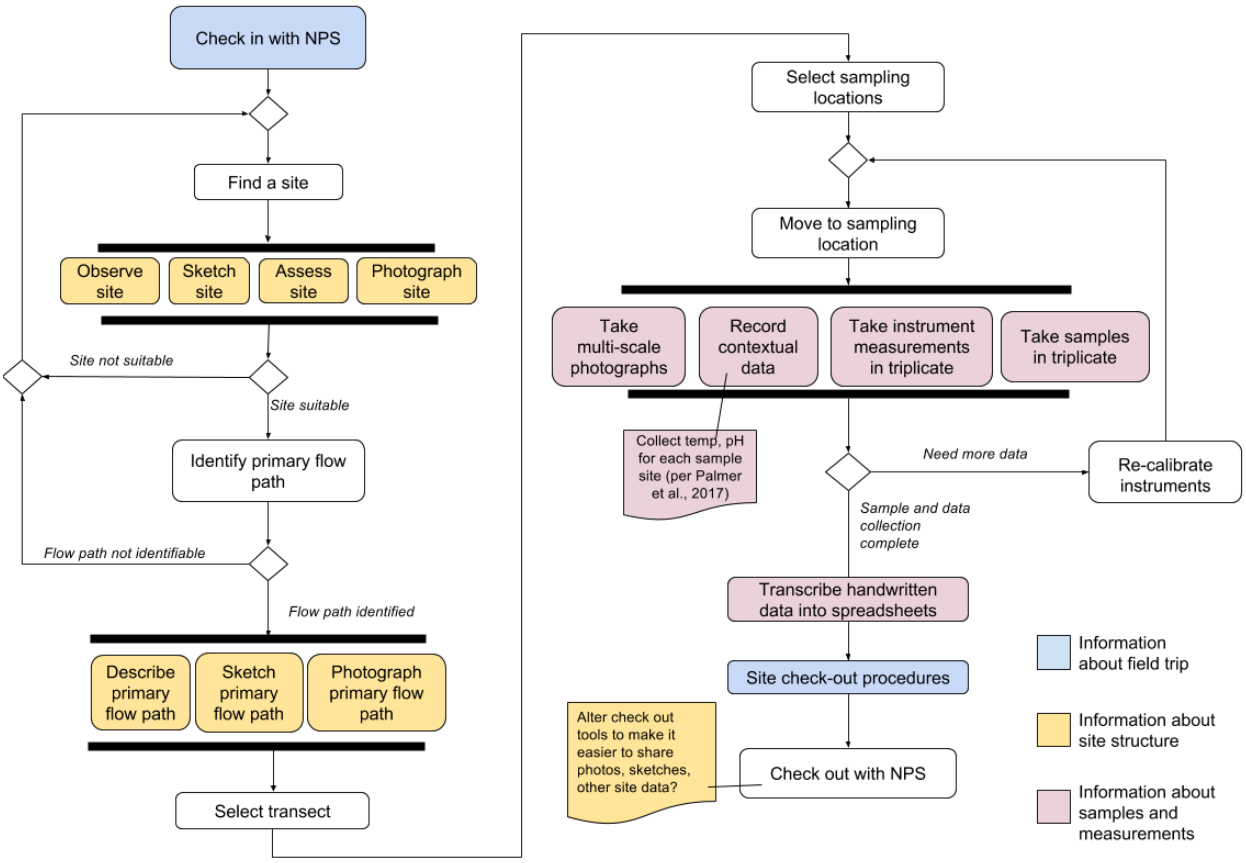


Figure 5.4. The fieldwork stage of research. The "document" shapes (squares with wavy bottoms) indicate recommended or potential curatorial interventions. The pink callout recommends that researchers collect data according to the guidelines recommended in Palmer et al., 2017. The yellow call out suggests that existing site check-out protocols/mechanisms be altered to support reporting of information about site structures back to the NPS – specifically, the photos and sketches taken by data collectors in the field.

Through annotation of Fouke's workflow, we were able to "test out" implementation of the MIF – at least conceptually. This allowed us to verify the importance and relevance of MIF elements in geobiology research, as well as identify and make explicit gaps in reporting infrastructure. One of the prevailing topics of conversation (and areas of concern) while developing the MIF was just how feasible it would be to collect and aggregate all of the data elements we identified as being minimally necessary for data reuse and preservation. YNP stakeholders were well aware that a data standard is not necessarily helpful if it's not one that can reasonably be adhered to. However, workflow analysis allowed us to assess the feasibility of the MIF and gain a better understanding of how it might be implemented in the future to encourage greater adoption of the standard.

We found that, by and large, Fouke already collects the information elements recommended by the MIF, and few changes would need to be made to his workflow to adhere to our recommendations. However, our workflow analysis did underscore the need for further reporting mechanisms. While Fouke collects a large amount of information about his field site's structures, and about his individual field trips, not all of this is reported back to YNP, or necessarily reported through his journal publications. Additionally, this information is collected all throughout different stages of his work, and is consequently at risk of being lost or misplaced throughout the months (and sometimes, years) that take place between the planning and analysis stages of research. These workflow and infrastructure gaps are discussed further in chapter 6.

This workflow analysis additionally led to the development of a method of *research process modeling*, which uses several approaches from systems analysis to create workflows, inventories, and provenance documentation to create well-annotated research objects suitable for publication and archiving. Early iterations of this work have been presented at the annual meetings of the American Geophysical Union (Thomer, Baker, Jett, Gordon, & Palmer, 2014; Wickett, Thomer, Baker, DiLauro, & Asangba, 2013), and a publication describing this approach has been accepted pending revisions to JASIST.

Evaluation through work with Fouke's class at YNP[26]

I additionally tested the feasibility of the MIF in practice by serving as the field assistant in Fouke's Fall 2015 Introduction to Biocomplexity course. In this course, students take a field trip to YNP to learn the fundamentals of geobiology fieldwork; I served as the course's field assistant. I created an MIF-based template for the students' data collection. I additionally gave several class lectures explaining the MIF, how to use the template, and general overviews of data curation best practices.

I observed the students' work in the field and reviewed their completed data collection sheets. The MIF functioned well in supporting structured description of data at the point of collection in the field. Some students reported that they enjoyed using the template, and none found it to

[26] This subsection is adapted from Palmer et al., 2017

impede their project work. That said, some had difficulty recording the precise locations of the vents and the bearing of each sample site relative to the vent. Students worked in the main area of Mammoth Hot Springs, which includes upwards of 20 individual vents and several overlapping spring systems. Because this is a delicate and publicly viewable area, they were restricted to the "boardwalk" paths and unable to reach certain segments of the springs to take more precise geolocations. The complexity of the spring system and these access limitations consequently limited their ability to collect all the data we asked of them.

Thus, through this trial run we found that even in localities as accessible as YNP, there are still challenges in gaining sufficient access to a field site for recording certain key contextual elements. In large hot springs, researchers may not be able to precisely locate key site structures, such as the vent or the mid- and end-point of the drainage channel, and they may not be able to precisely measure the position of sampling sites relative to the vent. In such cases, estimates of distances are a reasonable alternative for supporting later reconstructions of the hot spring structure. In these cases, supplemental photography (including recent satellite imagery) and field sketches should be encouraged.

5.3 Minimum information framework for paleontology at La Brea

The development of the La Brea information framework was somewhat more straightforward than that of YNP, in part because of my extensive prior work developing the YNP, but also because of La Brea's existing data collection standard. The La Brea stakeholders were less concerned about the impact of standardizing their data collection work, because they had been working with a data collection standard for decades. Additionally, the La Brea stakeholders were already invested in re-engineering their data collection protocols and curatorial processes prior to my dissertation work at the site.

Like the YNP MIF, the La Brea MIF enumerates the information that stakeholders identified as being minimally necessary for effective data curation and reuse. Researchers and resource managers both identified the following data elements as important to their work: the ages of taphonomically informative fossils tied to their locations within a deposit; the broader geologic context of a deposit; and their fossils' preparation histories. Resource Managers additionally needed a rough inventory of their collections prior to their formal curation.

The Resource Managers also noted several recurring issues regarding the data collection method and curatorial processes at La Brea:[27]

- The grid system is rooted in the southeast corner of the deposit; this makes the data difficult to map in common software. Plotting the origin in the southwest corner would make analysis much easier. The data would likely still need manipulation prior to import, but not as much.
- The measurements are too time consuming to take, particularly for a salvage project that needs to be completed within certain time constraints.
- The measurements don't provide researchers with the data they need; a "slope," orientation, and basic geolocation would potentially be more effective.
- Not all fossils need to be measured out – just the ones that might be taphonomically informative. This would likely include "long bones" from animals' limbs such as femurs, humuri, and so on.

I sought to integrate these needed changes in my design of the La Brea MIF. Additionally, with the La Brea MIF I sought to formally encode information elements that are important, known by some site workers, but often not written down – people's full names, fossil preparation methods, and so on. Finally, encoding guidelines for the La Brea MIF are more well-developed than YNP's, because the information system for La Brea is more established.

Like the YNP MIF, the La Brea MIF is composed of three classes of information: Information about an Excavation Project; Information about a Fossil Deposit [the site]; and Information about Fossils and Soil Samples (Figure 5.5).

[27] A full outline of my recommendations and discussion of potential changes for La Brea is included in Appendix D.

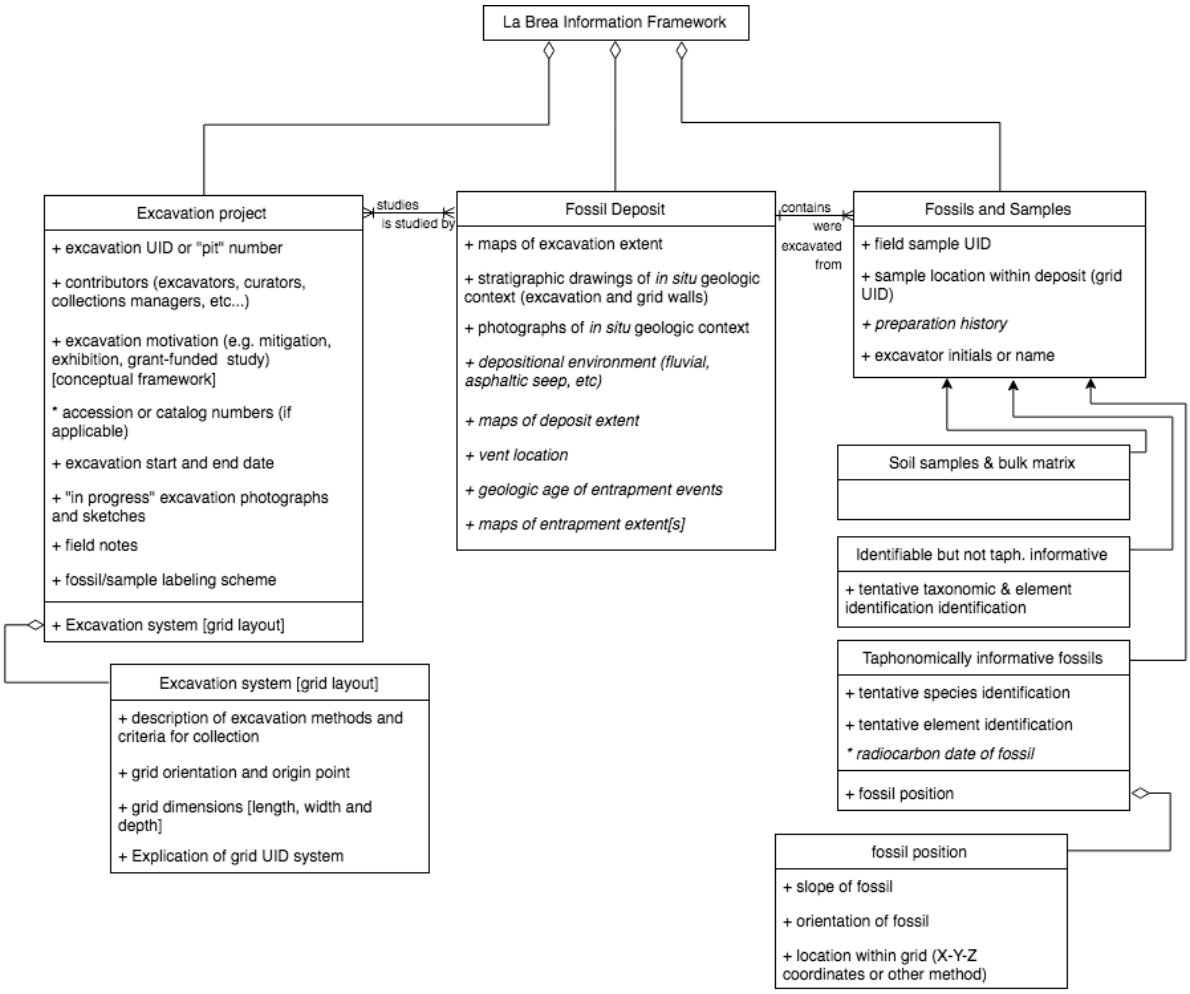


Figure 5.5. The La Brea Information Framework. These classes have several unique subclasses and attributes not found in the YNP MIF. The La Brea MIF also includes several elements that may not be applicable to all excavations, as well as several elements that could only be input long after an excavation's completion.

UID = unique identifier (e.g., a grid number or a specimen number). +Elements with pluses: required. Elements in italics: would likely need to be inferred and documented iteratively, over the course of excavation and research activities. *Elements with asterisks: may not be needed or feasible for all instances.

1. Information about an Excavation Project. This class includes the following elements:

- *Excavation UID or "Pit" number.* A unique identifier assigned to the excavation project.
- *Contributors (excavators, curators, collections managers, etc.)*. These are the full names of La Brea staff and volunteers involved in the excavation and curation of the project, as well as any outside researchers that contributed to the development of the excavation project or method.

- *Excavation motivation (e.g. mitigation, exhibition, grant-funded study)*. An explanation of the goals and motivations for the excavation.
- *Accession or catalog numbers (if applicable).* The excavations are sometimes given individual accession numbers by the registrar at NHMLA; they may also have ranges of catalog numbers reserved for the fossils and samples that result from the excavation.
- *Excavation start and end date(s).* This should include a description of whether the excavation was seasonal (only in the summers) or ongoing (year-round).
- *Annotated "in progress" excavation photographs and sketches*. Photographs and sketches need to be taken of the excavation as it progresses; they must be annotated with the date, "pit" number, relevant grid number, and other contextualizing data.
- *Field notes.* Daily narrative records describing excavation conditions and other miscellaneous details as needed.
- *Fossil/sample labeling scheme.* Description of the method for assigning identifiers to samples and fossils.
- *Subclass: Information about the Excavation System.* This subclass includes elements describing the method of excavation for a project:
    - a narrative *description of excavation methods and criteria for collection*, such as an internal report or a formal publication;
    - descriptions of the gridding system, including the *grid orientation and origin point,* the *grid dimensions* (i.e., length, width, and depth of the grids), and an *explication of the grid UID system* (an explanation of how identifiers are assigned to grids and if they contain any semantic meaning).

2. Information about a Fossil Deposit. This class is intended to document the physical structure of the fossil deposits of La Brea. In the past, some of this information may be informally or anecdotally known about a deposit, but that information is rarely stored in a structured manner and therefore is at risk of being lost, or at the very least, cannot be used to organize or curate the collections. It includes three information elements already collected at La Brea:

- *Maps of the excavation extent*. These are hand-drawn and satellite maps detailing the bounds of an excavation – or, in cases such as Project 23 in which the excavations have been moved *en bloc*, the original bounds and orientation of the removed fossil block.

- *Stratigraphic drawings of* in situ *geologic context (excavation and grid walls).* These are sketches of the geologic layers and composition of the deposit as it is excavated. In the Pit 91 and P23 projects, these have consisted of "wall" and "floor" drawings – sketches of the strata exposed at the bottom and on the sides of each grid after it is excavated. In a mitigation context in which the deposits are being moved *en bloc* this category should also include drawings of the *in situ* context surrounding the removed block of sediment (several La Brea stakeholders noted that they wished they had access to this information in the P23 project).
- *Photographs of the* in situ *geologic context.* Similarly, photographs are needed of the grid walls, floors, and original depositional environment. However, unlike the YNP case, these aren't as valuable as well-rendered sketches; often the detail needed to make geologic interpretations cannot be easily rendered in a photograph (especially when the sediments are so dark).

This class additionally includes five information elements that would need to be inferred once excavation is completed and a certain amount of analysis has been done on the fossils. These elements are not necessarily collected or stored consistently at La Brea. These elements include:

- *Depositional environment (fluvial, asphaltic seep, etc.).* A characterization of the nature of the deposit. In some cases, this may be immediately apparent, but in others it may need to be revealed over the course of excavation. In this latter case, excavation methods may need to be altered to "fit" the deposit environment (for instance, if a deposit was believed to be too mixed to collect positional data, but is revealed to contain a partially articulated skeleton, the excavators may need to begin collecting positional data).
- *Maps of deposit extent.* This is a map of the bounds of the fossil *deposit* – rather than the excavated earth. These would need to be inferred from "in progress" wall and floor maps or other imaging used throughout the excavation. This information is needed to help contextualize the fossil collections and inform future work with the specimens.
- *Maps of entrapment extent(s).* A deposit may contain skeletons from several entrapment events – periods of time in which animals became trapped in the seep. These time spans are identified through radiocarbon dating of individual bones; the spatial extents of them are inferred by cross-referencing radiocarbon dates with stratigraphic and positional data.

Mapping these boundaries will help contextualize the fossil collections and inform future research. Even understanding whether the deposit is too "mixed" to identify specific bounds of entrapment events is informative in and of itself; it helps bound the kind of research that can and should be reasonably conducted with the specimens.

- *Geologic age of entrapment event(s).* As described above, this is inferred through cross-referencing the radiocarbon dates from individual specimens and stratigraphic data.

3. Information about Fossils and Samples. In my revision of the La Brea data collection protocol, I propose that the La Brea RMs collect fossils in three different ways, depending on their identifiability and taphonomic informativeness. Thus, I model "Fossils and Samples" as a superclass that includes three classes. The superclass includes minimal information elements inherited by all three classes of fossils (a *field sample UID,* a *sample location* or *grid number UID, excavator UID,* and *preparation history)*.
An excavation project may collect just one of these classes of fossils, or multiple classes, or all three. These classes include:

a. Soil Samples and Bulk Matrix. This is the simplest class – samples of sediment and small fossils or fossil fragments that are collected in bulk, and therefore do not merit field identifications or detailed positional data. This class simply inherits the basic information included in the "Fossils and Samples" superclass (*field sample UID, grid UID, excavator UID,* and preparation history).

b. Identifiable but not Taphonomically Informative Fossils. These are fossils that are complete enough to merit a field identification (needed for later inventory and curatorial work) but that are either too small or too disturbed (coming from a deposit with little stratigraphic integrity) to merit the collection of specific positional data. In addition to inheriting the information elements from the "Fossils and Samples" superclass, this class of information includes:
    - *Tentative taxonomic identification:* a preliminary taxonomic identification, down to species level if possible.
    - *Tentative element identification:* a preliminary anatomical identification (e.g., femur, ulna, vertebra) including side (right/left, proximal/distal) or maturity (juvenile, sub-adult) if possible.

c. Taphonomically Informative Fossils. These are fossils that have been assessed as potentially informing the future interpretation of a deposit's taphonomy and depositional structure. The definition of a "taphonomically informative fossil" will likely change from project to project, and even may change over time. For instance, in early Pit 91 excavations, all fossils and fossil fragments over one quarter inch in size were considered taphonomically informative; whereas in the beginning stages of the P23-1 excavation, only identifiable "3-point" fossils (Shaw, 1982) were. The goal of this class is not necessarily to define what fossils should be considered taphonomically informative, but rather, to define the information that ought to be collected if fossils *are* deemed informative.

The taphonomically informative class includes two of the same elements of the "identifiable" fossil class (tentative taxonomic and element identifications), but also includes *information about a fossil's position (modeled here as a subclass)*. Fossil location measurement methods may change from one excavation to the next and should be described in the "*description of excavation methods and criteria for collection*" element in the "Excavation System Information" subclass. The "Information about a Fossil's Position" subclass should be tailored to the method being used in each specific excavation.

Here I propose that future excavations begin collecting the *slope of the fossil, the orientation of the fossil,* and *the fossil's location within the grid,* in lieu of the current three-point measurement. In the future, though, La Brea RMs may use image-based methods such as LIDAR for documenting the fossils' positions. Adoption of these methods and technologies would require re-evaluation of this subclass to account for the need to tie individual fossils to the image output.

EVALUATION THROUGH WORKFLOW ANALYSIS

In the YNP case, we first evaluated the MIF through comparison to an external reporting standard. At La Brea, though, I moved directly to workflow analysis. Because La Brea already uses an in-house excavation standard, comparison to an external standard would not be as

informative. However, I do compare the La Brea MIF to other paleontological data structures in the following chapter by means of discussing broader workflow issues in site-based data curation.

I first modeled La Brea's present excavation workflow, which is rooted in the Shaw 1982 protocol (Figure 5.6).

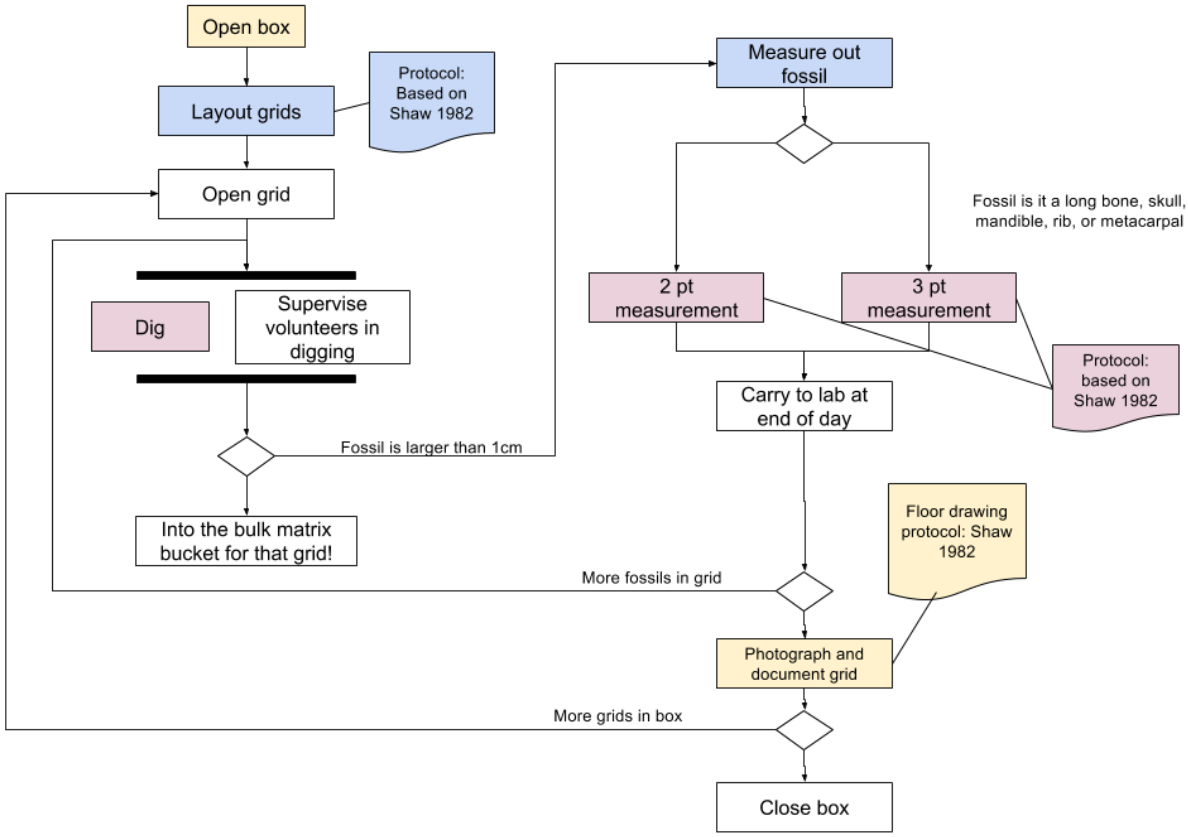


Figure 5.6. La Brea's present day workflow.

I then revised this workflow to represent the recommended changes developed with the La Brea stakeholders (Figure 5.7; our new protocol is referred to as "La Brea 2017").

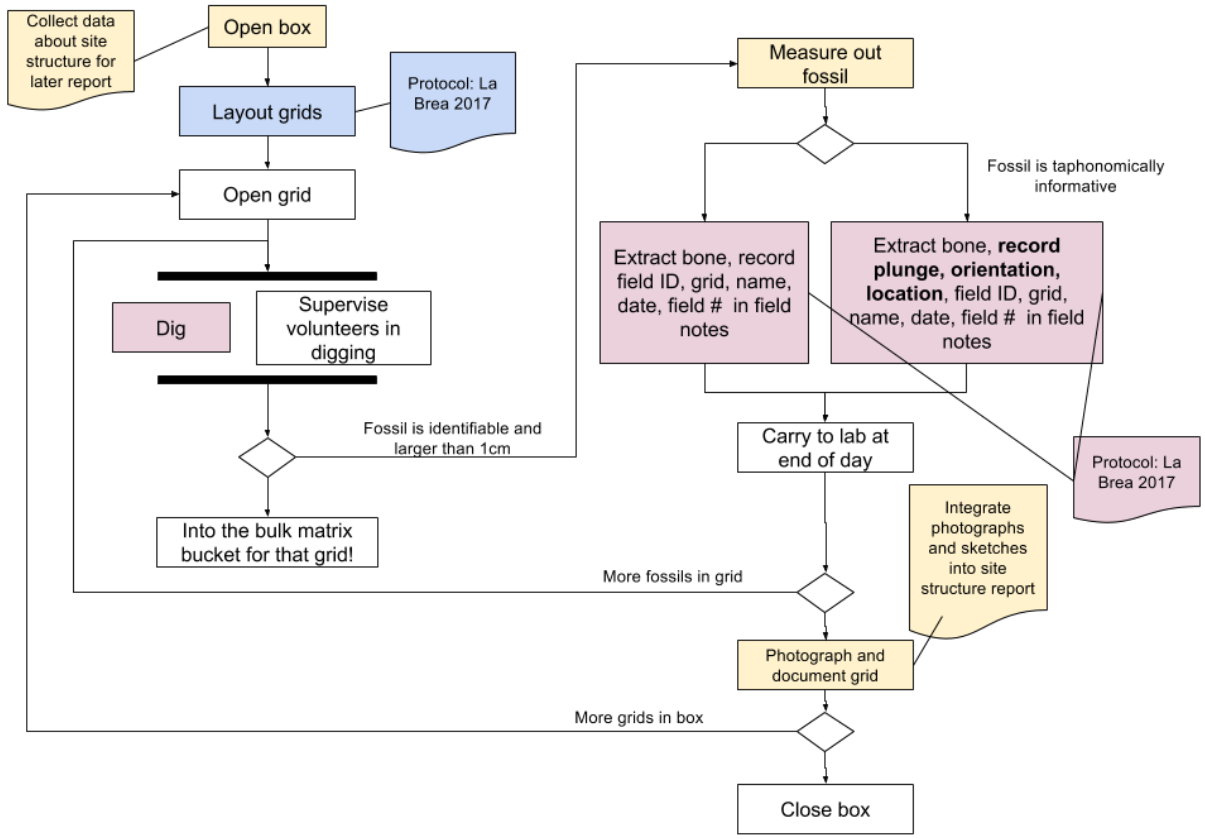


Figure 5.7. La Brea's revised workflow, reflecting proposed changes to their excavation protocol.

As with the YNP MIF, annotating the La Brea data collection workflow allowed me to verify the relevance of MIF elements, and to clarify the point in the data collection workflow where they are collected. In this case, though, it also verified the feasibility of the new data collection protocol, while also identifying areas for future work. In Figure 5.7 I've called out two curatorial interventions not present in the original workflow in Figure 5.6: the collection of data about the site's structure during the "open box" process, and the integration of photographs and sketches into the "site structure report." The RMs already collect this information about site structure in an *ad hoc* manner, but further work is needed to determine the best format for it, and the best method of ensuring that it is collected.

The re-engineered and annotated workflow also helps underscore that the proposed changes to the data collection protocol do not alter the high-level configuration of work processes at La Brea. Rather, they change some of the lower-level data collection and information recording

strategies. These changes could be modeled at a lower level of granularity in the future, if necessary.

Evaluation through comparison to other information systems

Though La Brea maintains its own in-house collections database, the data that researchers derive from the La Brea collections is often published through paleontology databases such as Neotoma, the Paleobiology Database, or through domain-agnostic databases such as Dryad. The researchers' data contains several elements central to the La Brea MIF, including radiocarbon dates of the fossils, and potentially more detailed maps describing the geological context. I assessed both Neotoma and the Paleobiology Database to discover a) whether they captured these key elements, and b) if they provided an API or other mechanism that could potentially be used to publish or harvest data about La Brea in the future.

Neotoma. Neotoma is a paleoecological database for "recent" fossil sites up to five million years in age. Neotoma is described as "a centralized database with virtual constituent databases, e.g., the North American Pollen Database or FAUNMAP" (Grimm, 2008). While Neotoma does publish descriptions of geological context and radiocarbon dates and also provides access to this data through an API, the geological context is likely not at the granularity needed for curation or analysis. For instance, the locations of the Rancho La Brea fossils currently in Neotoma are only identified down to "Pit" – not to grid, depth, or surrounding stratigraphy.

Paleobiology Database. The Paleobiology Database is a database of fossil occurrences harvested by hand from the paleontological literature and input into a centralized database. The geological and temporal classifications are even less fine-grained than Neotoma's; the few La Brea specimens in the database are all classified as "Rancholabrean" in age, which is anywhere from five to 50,000 years old. And the only geological context provider is the pit number, and very broad characterization of the stratigraphy (for instance, "tar; poorly lithified tar").

The gaps between the La Brea MIF and these paleontology databases underscore a gap in broader data reporting workflows identified in the YNP case. Though researchers often produce data that would inform curatorial work, it is rarely reported in a manner or granularity that would be needed for reuse.

5.4 TOWARDS AN INFORMATION FRAMEWORK FOR SITE-BASED DATA CURATION

Despite the differences in the sites, their administration, the domains of study, and data collection methods in these cases, the La Brea and YNP information frameworks can still be modeled with a similar tripartite structure (Figure 5.8). The classes can be generalized into information about:

- Collecting Event: the people, processes, methods, and strategies used to collect data;
- Site Structure: the key natural features at a site at the time of data collection, the relationships between those features, and their relationship to sampling locations; and
- Sample Sites and Measurements: the individual observations, measurements, samples and specimens collected from a site, and their orientation or location within a site.

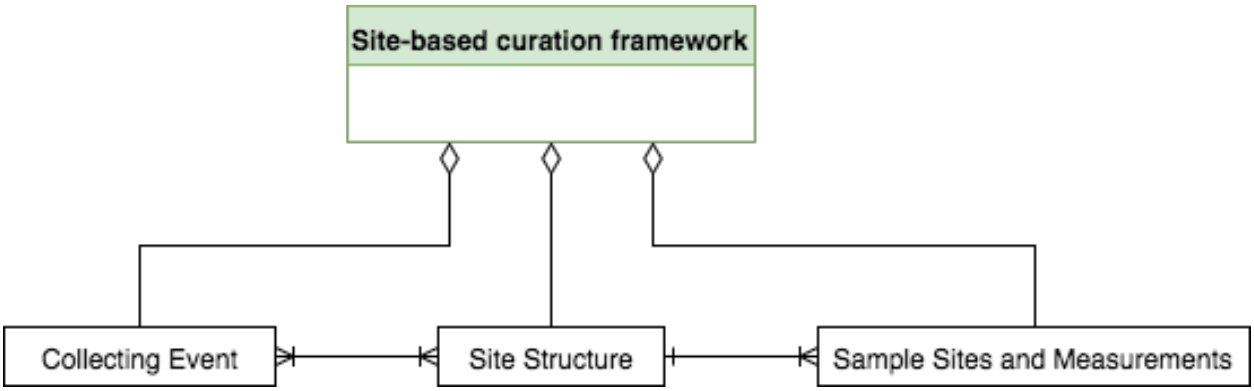


Figure 5.8. General site-based data curation information framework.

The "Site Structure" class is a key contribution of this model. Explicit and detailed representation of a site's natural structure through key parameters correlated with geolocations seems critical to curation and future integration or reuse of data collected from scientifically significant sites.

In the YNP case, information about the site structure includes information that makes it possible to reconstruct a hot spring's physical extent, orientation, and arrangement in space (via the location of the vent, mid-point, and end-point), as well as the spring's water chemistry (via the temperature and pH of the vent, mid-point, and end-point). In the La Brea case, site structure is represented through information that makes it possible to reconstruct the extent and geologic age of a fossil deposit – not just an excavation site.

It is unclear how many existing ontologies or site-based information structures do or do not necessarily prioritize information about site structure; however, based on the review of the data reporting platforms above (EarthChem, Neotoma, and the Paleobiology Database), it seems that some site structure elements are at least minimally present, albeit not explicitly called out. Additionally, there are slightly more domain-neutral ontologies that could be used in the future to make this information more overt. The BioCollections Ontology (BCO), for instance, was explicitly developed to link specimens to their sites of origin and their collecting events (Walls et al., 2014). That said, though the ontology does provide site-based terms, they number only six out of 117; the focus of the ontology is clearly on documenting collecting events and the specimens and samples themselves. BCO's site terms can of course be augmented through terms from other ontologies (for instance ENVO, the Environment Ontology, provides upwards of 2,159 classes and 18,791 axioms describing "representations of habitats, environmental processes, anthropogenic environments, and entities relevant to environmental health initiatives" (Buttigieg et al., 2016)). Yet these axioms focus on the intra-class relationships between environmental entities – not inter-class relationships between site structure, specimens, and collecting events.

5.5 CHAPTER SUMMARY

In this chapter I have presented the minimum information frameworks developed at each of my sites, and compared them to data collection workflow models to assess the feasibility of their implementation. I have additionally presented a general information framework for site-based data curation, containing three basic information classes that apply to both of my study sites. I compared these information frameworks to data collecting workflows at each of my sites, and additionally compared them to existing information systems, standards and ontologies relevant to research at scientifically significant sites. I find that workflow analysis is an effective method of evaluating the completeness and efficacy of minimum information frameworks, as as well as an effective way of identifying potential points for curatorial intervention.

6. Discussion and Conclusions

In this chapter, I discuss and analyze findings from previous chapters. I start with key observations from my intercase and intersubcase analysis; then refine a definition of scientifically significant sites and site-based data curation; and further explore the kinds of reuse my research has revealed through the lens of Coombs' Theory of Data. I tie my findings to my research questions throughout these sections. I conclude with a discussion of the effectiveness of my methods and the limitations of this study; and present my propositions and proposed directions for future work.

6.1 Key differences in curatorial infrastructure and process

Differences in sites' infrastructure and administrative priorities

La Brea as a site has a much more robust curatorial infrastructure than YNP, possibly because La Brea has more of a curatorial mandate. The Tar Pits are protected primarily for the sake of the fossil collections, whereas YNP was set aside for a broader range of uses and stakeholders. La Brea RMs also have far more data, and far more curatorial responsibilities. However, one could ask which comes first - the collection or the collection infrastructure? The YNP RMs indicated that they would certainly be interested in curating researchers' data, however they do not (or did not, at the time of my work with them) have a dedicated storage system for it.

La Brea's curatorial infrastructure may also benefit from La Brea's relative organizational independence; though the La Brea Museum is overseen by NHMLA, their parent-child relationship doesn't come with the same bureaucratic overhead as YNP's relationship to NPS. For instance, while La Brea was given the option to migrate its collections database to the KE EMu system used by several other NHMLA departments, it was not required to, and La Brea RMs retain some leeway to use other systems (Excel, Google Drive, legacy databases) if necessary. The YNP RMs, however, are required to use the two core NPS data systems (IRMA and ICMS) and must also adhere to numerous other NPS guidelines. La Brea therefore has some ability to experiment with different curatorial processes and systems, whereas the YNP RMs feel more that they must wait to follow the NPS' lead.

Both the La Brea and YNP RMs discussed their desire for curatorial leadership from "experts" – either the researchers visiting the sites, or someone higher in their site's administration. The YNP

RMs were waiting for leadership or direction from others in NPS, especially regarding data infrastructure development and permitting requirements, whereas the La Brea researchers were waiting for their new curator to finally begin her job. RMs at both sites repeatedly said that they wanted their data curation policies (and collection policies, in the case of La Brea) guided by researchers' research questions, not arbitrary decisions.

DIFFERENCES IN RESOURCE MONITORING AT EACH SITE

YNP resources are much wilder and much more varied than those at La Brea. Though this present case study focuses on the geobiology research in the hot springs at YNP, many other kinds of scientific studies are conducted at the park – even within the hot springs. Additionally, scientific research is only one of the many broader uses of the site overall. As reviewed in Chapter 3, the park was founded primarily as a site for recreation, and the NPS tends to prioritize the needs of recreational visitors over those seeking to do research. Thus, the YNP RMs need to monitor diverse uses of a broader range of natural resources than the La Brea RMs.

At La Brea, the resource most in need of ongoing monitoring might be staff members' time and physical storage space within the museum, rather than the natural sites themselves. The La Brea RMs experience the same diverse demands of their sites, in which the site must be made available both to researchers and to the general public, but La Brea RMs are more likely to be asked to directly participate in improving recreational visitors' experience of the site. For instance, excavators are asked to give talks throughout their day, and are required to do as much of their work in public view as possible; other collections staff must give tours and manage excavator and volunteer schedules so that that the excavation and laboratory areas on public view are always well-staffed. Thus, the RMs must account for diverse site needs in their schedules and workflows. The La Brea RMs' interest in efficiency further reflects their need to manage time as a resource. La Brea's data collections are so massive that they must ensure they are doing their work as quickly as possible to avoid getting overwhelmed by a curatorial backlog. This backlog takes up curatorial time as well as physical space in the museum; processed and curated materials are easier to store than unprocessed. This is not to say that YNP RMs don't need to manage time and storage space as well (their search for a dedicated "space" for data storage reflects the latter need); these concerns just were not as present in our conversations.

Thus, these case studies underscore that site managers need not only to manage natural resources, they must manage infrastructural resources as well. A full list of resources under RMs' management might include: the integrity and accessibility of the sites themselves for diverse uses in research; the integrity and accessibility of the sites themselves for education and public enjoyment; the data collections associated with the sites; the storage for the data collections associated with the site (both physical and digital); and finally, the RMs' time and attention. These factors will be important to consider for future data curation research and workflow analysis.

Specimens, samples and identifiers

Physical samples – that is, physical, material pieces of the natural world – are collected at both La Brea and YNP. However, these "samples" fall under different jurisdictions at each of the sites. At YNP, geobiologists are able to avoid the strict regulations that typically govern any researcher collecting physical samples at the park by taking advantage of a loophole in said regulations: They collect physical (rock and water) samples small enough that they don't need to be accessioned into NPS collections via the ICMS database. At La Brea, though, the museum famously preserves and catalogs just about everything from the site, down to small fragments of insect remains and microscopic flecks of plant material. These different curatorial strategies impact the long-term management of the specimens, as well as their ability to be linked to other data products.

At YNP, there are both benefits and drawbacks to researchers' sidestepping the museum accessions process. Benefits include a higher level of independence for the researchers, and less administrative work for the RMs; drawbacks include a loss of potentially important data curation infrastructure and reporting mechanisms. If these smaller water and rock samples were accessioned into ICMS, they would be assigned a unique persistent identifier which could be published with downstream data products (genetic sequences, isotope data), and thereby provide a "link" back to contextual data about the site and field project. Our evaluation of the EarthChem Vent Fluids template showed that these links between field samples and derivative data products are necessary for site-based data work, and often missing from current information systems (such as EarthChem). However, researchers and resource managers alike expressed concerns that this sort of accessioning policy would be nearly impossible to undertake; the formal NPS

accessioning process is time intensive, and could delay researchers' work to the point that they would simply avoid coming to YNP and instead try to find other places to conduct their research. There is a third-party platform that may make it possible for researchers to register their samples without burdening the NPS or becoming beholden to their specific infrastructure: the System for Earth Sample Registration (SESAR; http://www.geosamples.org/) allows researchers to assign International GeoSample Numbers to their samples, which can then be used in later derived data products. However, SESAR is a relatively new platform and further work is needed to explore how best to integrate it in geobiology workflows.

Conversely, at La Brea collections staff are working to change their own cataloging processes so as to *avoid* assigning catalog numbers to very small samples, such as those aforementioned microscopic flecks of plant material. In the past every small fossil attached to a larger fossil would be cataloged individually; now, La Brea RMs are binning them into "bulk matrix" cans for later bulk cataloging. The La Brea RMs are additionally trying new ways of cataloging specimens (especially microfossils) while researchers are simultaneously working with them, using simple tools such as Google Sheets as a shared workspace in which RMs can assign catalog numbers to specimens and researchers can enter species identifications and morphological measurements. This concurrent effort makes it easier for the researchers to publish data along with the permanent catalog number (rather than a temporary ID that is not tracked by the museum), thereby making it possible to later associate derivative data with specific specimens. Thus, while Google Sheets certainly shouldn't function as a permanent storage place for collections catalogs, it does help align curatorial and research workflows that in a way that is simply not possible with relational collections databases. Further work is needed to explore how to integrate technologies like these, which facilitate collaborative or simultaneous data work, to further align curatorial and research workflows.

6.2 Defining scientifically significant sites & site-based data curation

At the outset of this study, I defined a scientifically significant site as one that had attracted enough sustained interest from researchers to merit government protection and the curation and maintenance of dedicated collections. This study confirmed my definition of a scientifically significant site, and, further, confirmed that scientifically significant sites are unique loci of science and data curation. Additionally, this study underscores that it is not just a site's natural

features that make a site significant; it's the infrastructure around them built to support research. As one researcher working on a pollen project at La Brea described it,

> What's interesting is that you actually have a lot of people who are working on many different aspects of the system. And because it's associated with a museum, there can be some coordination with their different efforts, right? There might be other sites where like there are lake sediments in Eastern North America where you can go, and you know -- anybody can go and take a sediment core from the lake… So, there's not necessarily any coordination between them. And at La Brea, it's not -- there's not necessarily any coordination between different people either, but because it's managed by institution, there's the potential opportunity for more coordination, right? (RLB-Rsch-4)

I would expand on this, and argue that even if resource managers are not coordinating research groups, they are coordinating internally to *support* research groups. That is, they are invested in developing services to support scientific endeavors at the park, and acting as intermediaries between researchers, other site administrators, and sometimes data collections. In my first research question, I asked, "what aspects of a site are most important from the perspective of researchers vs. resource managers?" While both groups of stakeholders see value in the various natural features in and around sites, they also see much of the value at scientifically significant sites in their research and resource management infrastructure. Thus, these sites seem to function as variations or subsets of Latour's "centres of calculation" (Latour, 1987): centers of coordination, collections management, and curation. In future work, I hope to unpack how the colocation of the actual calculation "center" – that is, a museum collection or a resource management office – and the natural site itself impacts the kind of calculation and coordination work that is done.

"Information about site-structure" as core to site-based data curation

A key motivation in this dissertation was understanding the roles played by data collection protocols and data curation processes in site-based research and data collections. I did this in part by developing minimum information frameworks and documenting and annotating researchers' workflows (presented in Chapter 5). The resulting three-part general information framework for site-based data makes explicit the data points researchers need to represent their study sites and that resource managers need to coordinate research activities and curate data. These frameworks additionally address my second and third research questions: How do data collection protocols and curation processes represent study sites? And how do protocols and processes appear in the datasets derived by sites?

As shown in Chapter 5, the general information framework for site-based data includes three classes of information: information about a field project, about site structure, and about specimens and measurements. The "Information about Site Structure" class is a unique contribution of this work, and possibly a unique feature of scientifically significant sites and their information structures. In non-site-based sciences, minimum information needs would more likely consist simply of two classes: information about a project, and information about samples and observations. However, in site-based sciences, researchers need to root their study designs and data collection protocols in an understanding of how the site's phenomena function, how different parameters are related to one another, how different physical natural features are related to one another. Information about site structure is also critical to systems-wide data integration and reuse. Site-based resource managers similarly need to understand this structure to inform their resource management processes, as well as their information organization strategies; additionally, even if they don't need the information about site structure for their personal work, they would likely need to facilitate researchers' access to it for further integrative research. Thus, models and representations of the site's structure are vital conceptual infrastructure for researchers and resource managers.

The importance of the "Information about Site Structure" class has several immediate implications for site-based data curation. First, it reveals a need for further expressivity in our data standards regarding information about site structure. My analysis of external data standards showed that though site structure can indeed be described through existing ontologies and data standards, it is not often called out as a class unto itself – or, if it is, it's usually much less developed than other classes of information. This points to a need for further work making existing information classes about site structure more explicit and more detailed.

Second, it makes clear a need to further align research and resource management workflows. RMs can only use or provide access to information about a site if they themselves have access. Unfortunately, though, information is often created through downstream analytical processes from which RMs already struggle to reclaim data. For instance, at La Brea radiocarbon dates are essential to inferring the site's structure, but these can only be generated through later research analyses. At YNP, temperature and pH data can be collected in the field, but will likely need some post-processing to put them in an easily reportable format.

Thus, there needs to be a feedback loop through which resource managers can retrieve important data products – either through data reporting, data harvesting, or ongoing collaboration between researchers and resource managers. Some of the successes at La Brea may be informative here; their use of Google Sheets as a collaborative platform for data work could be adopted easily elsewhere. Additionally, the La Brea RMs may receive more data from their researchers because of the high rate of collaboration between La Brea RMs and researchers. I speculate that this collaboration could have been facilitated in part because, despite their recent turnover, La Brea has had very few changes in curatorial staff. YNP on the other hand has seen fairly high turnover just within the last five years. It is easier to maintain relationships with visiting researchers when the curatorial staff is more consistent.

Finally, it is worth noting that the "Information about Site Structure" class was possibly easier to elucidate in the first place because I avoided discussing the dichotomy between data and metadata with my participants. As described in Chapter 5, I tried to use the term "information" when speaking about their data needs, because in early conversations I found that the term "metadata" had potential to distract or confuse. I believe that the three-part structure of my information frameworks was easier to infer than it would have been had I prejudiced my conversations with participants with the two-part dichotomy between data and metadata. Where the "Collecting Event" class contains information that LIS researchers would typically consider metadata (descriptions of the provenance, "aboutness," "who" and "how" of a dataset's production/creation), and the "Sample Sites and Measurements" class contains information that we would typically consider data (the individual observations that comprise the core of scientific analysis), the "Site Structure" class contains information that could easily be considered either data or metadata.

Thus, discussing data and metadata needs in terms of information classes may help re-introduce nuance into our discussions of scientific data, particularly with domain researchers with little prior experience in LIS. While the notion that "one person's data is another's metadata" is certainly worth discussing in an LIS setting, I found that this relativity distracted participants from the matter at hand: understanding what broad classes of information were needed to make site-based data reusable. "Information" was a sufficiently broad but still easily understood

concept that allowed us to discuss data curation needs without getting lost in complex discussions.

6.3 Kinds of Reuse and Coombs' Theory of Data

Data reuse was an underlying concern of this study, given that site-based data curation is fundamentally meant to support the reuse of data collections from scientifically significant sites. Throughout this work, participants described several kinds of reuse:

Direct reuse, possibly leading to further integrative reuse: La Brea researchers described several kinds of data that might be directly reusable with little cleaning or statistical manipulation, such as individual fossils' morphological measurements, radiocarbon dates, and other isotopic data. These data could be reused directly or integrated into a larger study.

Retrospective correlative reuse: reanalyzing a collection of datasets to correlate parameters that were not necessarily the focus of the original studies. This is similar to integrative reuse, but may require more repurposing of the data for analyses than its collector intended. At YNP, this kind of reuse came up in discussions of key parameters to include in the MIF; most researchers agreed that temperature, pH, and geological context would support this kind of reuse (though other parameters were suggested as well).

Reuse for reproducibility: at La Brea, researchers described wanting access to others' data about individual specimens so they could verify or reproduce their measurements or conclusions about specimens. For instance, if one researcher described bones as having evidence of certain kinds of pathologies, another researcher may want to review those specimens as well as the original researcher's descriptions of pathology to verify the diagnosis.

Reference reuse: comparing a newly collected dataset to a collection of other curated data, and thereby situating the new dataset within an existing corpus. At YNP, reference reuse came out of discussions of genetic sequence data and the GenBank repository: researchers use pattern matching algorithms to compare sequences to GenBank and identify them taxonomically. The sequences in GenBank are thus reused not as inputs in the research project, but as comparative points of reference. At La Brea, researchers might not necessarily use computational algorithms to make comparisons, but similarly described wanting to compare, for instance, morphological measurement data to other researchers' findings. In both cases, reference reuse can also include

reviewing existing datasets to gain inspiration for future research, acting as a very in-depth literature review.

Many of these kinds of reuse are mentioned in one form or another in other studies. However, I want to note that all data "use" at La Brea is fundamentally rooted in "reuse" of fossils that were excavated, documented, prepared, and curated by museum staff. This is distinct from "direct reuse" in that specimen-based research generally results in the creation of additional facts (measurements, chemical data) about the specimens; it is also distinct from specimens being used solely for "reference reuse" because their specimen records are being improved through the process of reusing them. This kind of reuse could be described as *generative reuse*. Though new data points about the specimens are generated in pursuit of varied research questions, they can be integrated back into the museum's catalog provided they are published with a catalog number correlating new data to a specific specimen.

Generative reuse is possible at La Brea largely because of the materiality of La Brea's data; physical specimens can be analyzed over and over again as new analytical methods are developed. This reanalysis is simply less possible with immaterial observational data. However, generative reuse is also facilitated through La Brea's cataloging system and data curation workflow. Each specimen is cataloged with a unique identifier; if later data points and relationships between data points are published along with identifiers, La Brea RMs can associate downstream data products with their specimen records. Generative reuse is therefore dependent not just on action by the researchers, but also the resource managers, who must somehow pull downstream data products back into the collection. In this way, generative reuse is the result of coordination or collaboration between resource managers and researchers; working to foster this kind of collaboration may help researchers and resource managers balance their needs for common data stores (RQ4).

In this way, generative reuse represents a way in which NHM curatorial processes may be extended to site-based data collections (whether physical or not). As Strasser has argued, some of the most robust data repositories in biology – GenBank and the Protein DataBank – can be regarded as adaptations of the natural history tradition of collections-based research. He refers to these databases as "the experimenters' museum" and argues that

> These databases undoubtedly represent an outcome of the experimental tradition, but at the same time they belong to a way of knowing that is perhaps best described as "natural historical," in that it rests on the collection and comparison of natural facts, often across many species. (Strasser, 2011)

These databases additionally depend on a tradition of generative reuse and curation, in which researchers and curators work together to annotate and generate additional data about their specimens. Reuse is effectively built into their collections development policy. Further articulating how reuse has and continues to function to improve collections will enrich notions of the modern natural historical tradition of work (as described by Strasser), and inform further alignment of curatorial and research workflows both in and out of natural history collections.

The categories of reuse described above – particularly this notion of generative reuse – deepen existing notions of data reuse. However, further work is needed to better define what data reuse actually entails. Despite our field's enthusiasm for encouraging data reuse, the concept is somewhat underdeveloped. As noted above, some of the categories described here have been mentioned in prior work, but there do not seem to have been major efforts the creation of a typology of kinds of reuse until quite recently. Recent work by Pasquetto, Randles, & Borgman (2017) is one example of this; the authors draw a distinction between "independent reuse" or reuse for replication ("Reproducing a study is an example of independent reuse of a dataset") and "data integration" (in which datasets are "compared and integrated for a single analysis study, a meta-analysis, parameter modeling, or other purposes"). This work complements that of earlier work by researchers such as Cliff Lynch, who sorted types of reuse into two different core categories: "reexamination for a compilation or a metaanalysis" in conjunction with similar data, for purposes that are not too different from the original, and reuse of data "outside the disciplinary frameworks within which they were collected" (as quoted in Weidman, Arrison, & National Research Council (U.S.), 2010 pg 7). My work shows that there are further distinctions to be made in this vein. The nuanced differences between varying methods and motivations of data use will require very different tools and infrastructures; better defining the tasks and goals we're referring to by saying "data reuse" will be a critical first step in developing them.

Reuse and Coombs' "Theory of Data"

Understanding modes of reuse may help refine our understanding of what makes data reusable. As shown in Chapter 4, researchers identified several types of highly reusable data. When

analyzed through the lens of Coombs' Theory of Data (introduced in Chapter 3), the relationship between data collection protocols, information structures, and usability becomes clearer (RQ3). The highly reusable categories of data can be summarized in the following four categories:

Physical samples and specimens: Whether rock samples at YNP or fossils from La Brea, physical samples are highly reusable, because, as detailed above, their material properties can be reanalyzed over and over again.

Photographs: Particularly in the YNP case, stakeholders were adamant about the importance of photography in documenting site conditions and thereby contextualizing site data. Photographs provide visual representation of a deposit or specimen's original geologic context. At both YNP and La Brea, photographs may not be used directly as data, but more as metadata that would help make an associated dataset reusable. However, YNP researchers did note that photographs that resulted from analytical work – micrographs and forms of crystallography – would be highly reusable.

Geologic maps and sketches: La Brea stakeholders – particularly the RMs – said that geologic maps (drawings of the stratigraphic layers within and surrounding the deposit) were reusable as well as key to supporting the reuse of other data. In some ways, sketches can be more useful than photographs at La Brea because the dark brown, oil-soaked sediments simply do not photograph well. When made by a trained geologist, sketches make clear subtle differences between soil types and depositional layers. However, it's important to note that sketches are a product of individual interpretation: whoever is creating the map must make many classifications of different soil types and colors. There are standards available to support this work – for instance, Munsell color charts provide standardized terminology for describing subtle differences in the color of sediments, and the American Society for Testing and Materials International (ASTM) publishes a standard for soil classification for engineering purposes that is often used by researchers – but the implementation of these standards is up to the creator of the map.

Data collected using well-established methods: Though the specific parameters vary from site to site, researchers seem more comfortable using data collected via methods with little variability – methods that were more normalized, so to speak. At YNP this included temperature and pH readings; at La Brea this included information such as morphological measurements.

Coombs' Theory of Data can be used to show the relationships between these categories of data and later perceptions of usability. As reviewed in Chapter 3, Coombs identifies three phases of data collection:

> Phase 1: observation and collection, in which the researcher decides which of the universe of potential observations she will collect;
>
> Phase 2: in which the researcher transforms, interprets, identifies, and labels her observations, thus rendering them into data;
>
> Phase 3: in which the researcher detects relations, order, and structure, following the data and the model used for analysis.

Essentially, Coombs models the scientific method as a series of decisions and interpretations: deciding what data to collect from the "universe of potential observations," how to classify it, and how to make connections between those classifications.

The categories of data outlined above can be mapped to Coombs' model according to the rough number of interpretive decisions involved in creating or collecting the data (Figure 6.1):

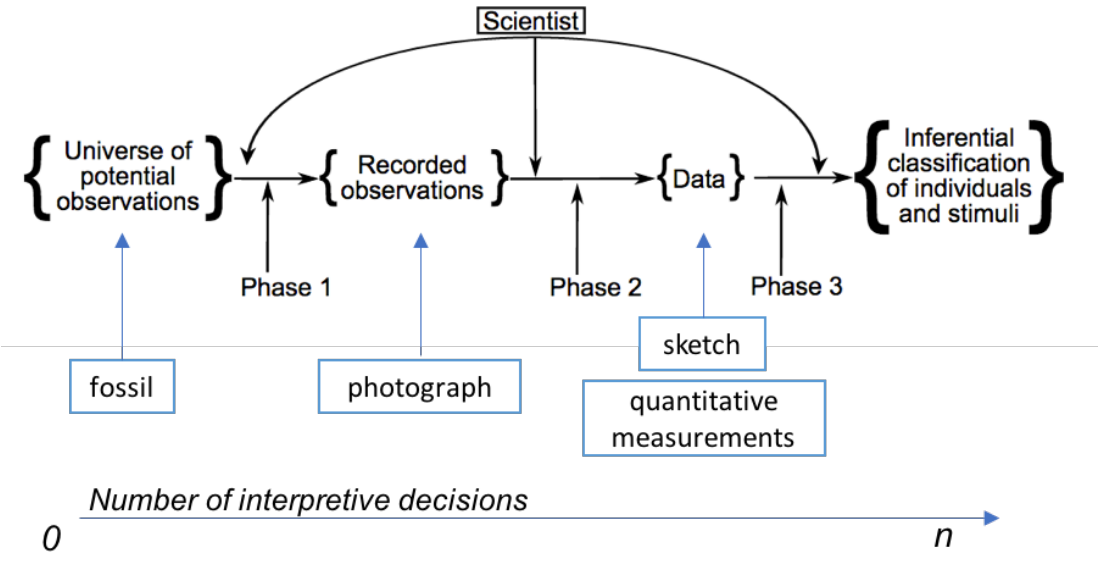


Figure 6.1. Reusable data types mapped to Coombs' Theory of Data.

Fossils and physical samples retain much of their original materiality, and thereby represent an almost unaltered extraction from the universe of potential observations; as such, they can be reanalyzed – or reinterpreted – repeatedly. Photographs occupy a middle space between "potential observations" and "recorded observations": they are certainly a record of some aspect of the universe, but one which could be reanalyzed. Quantitative data and sketches are both created through significant interpretation and classification work on the part of a scientist (depending on the degree of refinement of the dataset or geologic map).

I argue that viewing data reuse through the lens of Coombs' model helps reveal several common assessments that researchers undertake before reusing data: How many decisions and interpretations by other people went into making this data? Can I tell what those decisions and interpretations were? Do I trust whoever it was that made those decisions and interpretations? Would I have done the same thing? When researchers balk at reusing one another's data, it may, in part, be attributable to an innate understanding that there are analytical processes and decisions that are made during each of these three phases that are not necessarily explicit. These interpretations and decisions manifest in the data collection protocols. Additionally, when researchers say they need methods metadata or other kinds of context to assess whether data are fit for use or not, they are similarly appealing to a need to understand the varied decisions and interpretations that went into making a dataset. Thus, site context can be viewed as placing a dataset back in its originating universe of potential observations, whereas methodological context helps tell the story of how the data came to be from there. The amount of contextual metadata associated with a dataset might be thought of as one axis for determining the potential for reuse; the number of interpretations made in the creation of a data object might be another (Figure 6.2).

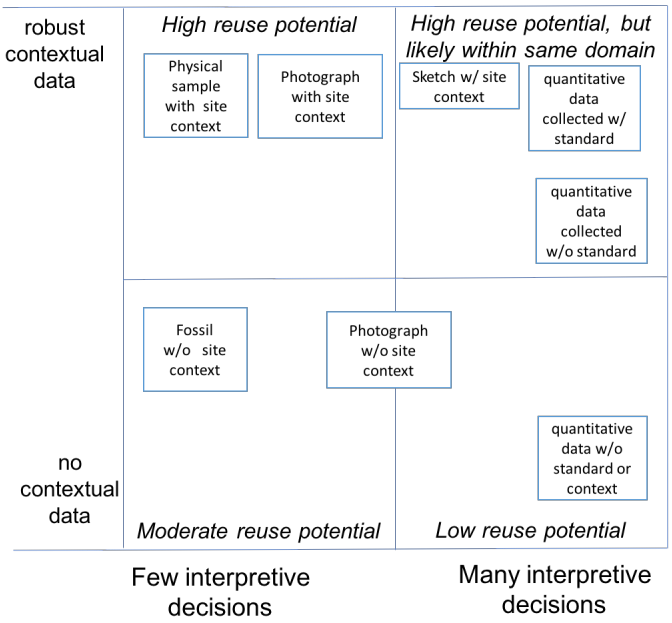


Figure 6.2. Plotting the reusability of datasets.

There are many additional dimensions to data use and reuse that would of course need to be considered to assess a dataset's potential for reuse – for instance, the fitness for the specific *kinds* of reuse outlined in section 6.2, and the fitness for use by a certain domain or user community (no matter how well documented the fossil, it likely won't be useful to a sociologist or a physicist). However, the criteria listed above may inform the creation of simple data quality metrics within site-based or domain repositories, or may contribute to further development of notions of analytic potential (see Palmer et al., 2011).

Coombs' Theory of Data is an effective lens for this work in part because his idea of a "universe of potential observations" is a fitting framing for a scientifically significant site. Though coming from psychology, his theory is both material and socially constructed -- that is, he assumes the existence of an external, observable world; that we make decisions in interpreting it; and that those decisions impact data structure. Framing site-based data in this way makes decisions that impact data structure and usability much more visible. Additionally, the findings from the application of this theory may extend to other kinds of data.

6.4 Effectiveness of Methods & Limitations of Study

For each of my cases, I used slightly different methods of participatory engagement; this resulted in some informative unevenness in my data. In some ways, the data from La Brea are more

robust due to my years as an excavator there and my continued collaborations with collections staff at the site. However, in other ways the YNP data are richer because the two groups of stakeholders were put in direct conversation with one another over the course of the two-day kickoff workshop. A focus group or workshop would have been helpful at La Brea, particularly to verify some of the MIF elements. However, despite the slight unevenness of my case studies, I now have more confidence in our methodology from the SBDC workshop; though I did not have the opportunity to develop as much of an ethnographic or experiential understanding at YNP as I have at La Brea, I do nevertheless think we fostered a meaningful conversation, and gathered important data.

Overall, I found that the participatory methods applied in this study are particularly appropriate for this work. The process of creating the MIFs guided and structured discussions with stakeholders about their needs, and made it easier to talk about abstract topics such as information models and workflows. Thus, the benefit of creating something like an information framework isn't just the new information framework – it's the discussions and community building that occur around the process of creating the information framework. My development of the information frameworks and use of systems analysis techniques through participatory action research approaches demonstrate a useful way of interacting with domain communities in a way that is more collaborative than other methods (e.g. qualitative interviews, non-participant ethnography), while still producing research of interest and relevance to the information science community.

LIMITATIONS OF THE CASES

These cases and the conclusions drawn from them contain the usual limitations of case studies; though rich, these findings are not necessarily predictive of curatorial or research arrangements of other sites. The frameworks are also potentially limited in applicability to their domains of study, and to specific sites; further work will be needed to explore how applicable my findings are to other scientifically significant sites.

These cases are additionally limited in that they are both very geological sites: the methods and epistemology of geobiology and paleontology are both largely rooted in geology as a discipline. Consequently, both domains are reliant on specimen collection, and both need considerable information about geological context and stratigraphic deposition. Thus, the findings from this

work may have limited applicability to other scientifically significant sites with a less geological focus (for instance, ecological sites), which may have very different data collection strategies and ways of knowing.

6.5 PROPOSITIONS

Throughout this dissertation I organized my data analysis around the revision of a set of *propositions*: short statements describing my expected findings, based in my and others' prior work. My original propositions were presented in Chapter 3; I used these statements to guide the coding of my data, and revised them throughout my analysis. Here I present the revised propositions and brief discussions of the revisions, organized according to the Research Questions each proposition addresses. The revised statements additionally function as succinct summaries of key findings from this dissertation.

RQ1: What aspects of a site are most important from the perspective of researchers vs. curators? How do these differ among and between sites?

Originally, my proposition addressing this question stated,

> *Original P1-1 Different stakeholders value the aspects of the site that help them do their jobs. These will differ depending on the natural features of a site and the goals of the stakeholders' work.*

I found that this proposition largely held true; however, I revised it as follows, to include greater detail about the perspectives of each of my stakeholder groups:

> *Final P1-1: Different stakeholders value different aspects of a site.*
>
> > *P1-1a: Researchers primarily value sites for their potential in research, and consequently value the specific samples and parameters they need for their research. Researchers additionally value the research, resource management, and curatorial infrastructure that facilitates their ongoing research.*
> >
> > *P1-1b: Resource managers value their sites' usefulness for diverse purposes, and therefore value data that other stakeholder groups (e.g. researchers, the general*

public) find useful, in addition to valuing information that facilitates the long-term management of the site and its associated data collections.

I additionally developed two propositions to describe how valued aspects of a site differ between sites, and how those values become apparent:

Final P1-2: Valued aspects of a site differ between sites depending on the natural features of a site, the breadth of research and other activities conducted at a site, and the administrative structure managing the site.

Final P1-3: Valued aspects of a site are reflected by stakeholders' decisions regarding what data to collect and preserve, and by decisions regarding what curatorial processes to prioritize over others.

The revised propositions addressing RQ1 emphasize that the "aspects" of a site valued by stakeholders are often intangible; research and resource management infrastructure are as important as unique natural phenomena. Additionally, they reflect the language around values that I adopted over the course of this work. I found that asking my participants to describe what they valued about a site was more effective than asking what aspects they found important. This further paves the way for future work considering how to reflect those values in the design of information frameworks, infrastructures and reporting processes.

RQ2: How do the data collection protocols followed by researchers represent study sites? How do curation processes followed by natural history museum curators represent sites?

I originally drafted three propositions in response to this question. I found that Proposition P2-1 held true throughout my work, and I did not revise it substantially:

Final P2-1: Scientifically significant sites have unique natural features and idiosyncratic administrative arrangements which need to be accounted for in any data collection and curation protocols, practices and processes at the site.

Proposition P2-2 also largely held true, however I revised it to include further detail about the site structure's role in shaping data collection protocols:

Original P2-2: Researchers develop data collection protocols and practices based on their project goals, hypotheses, access to resources, access to a site, and on prior work in their field. These factors lead them to select certain data points for collection over others.

Final P2-2: Researchers develop data collection protocols based on their project goals, hypotheses, access to resources, and access to a site. These protocols are also rooted in their state-of-the-art understanding of a site's structure and the drivers underlying the natural phenomena at a site. These factors lead them to select certain data points for collection over others, and guide them in relating data points to one another. Thus, data collection protocols represent not just the site, but the researchers' interpretation of the site, and the methods used to make that interpretation.

The final version of P2-2 emphasizes that data collection protocols represent a scientist's interpretation of a site and its features, as filtered through the lenses of a researcher's theories, methods and analysis. While this interpretive role was implied somewhat through my original version of P2-2, application of Coombs' Theory of Data brought it into sharper focus. As Coombs' argues, any data collection from the natural world is fundamentally rooted in data *selection* from a universe of potential observations and specimens. Thus, data collection protocols are records of a scientists' selection and decision-making process. In the future, I hope to further explore how data collection protocols are used to mediate between the natural world and the researchers' studying it.

Proposition P2-3 held true but was similarly revised to include additional details about how curatorial processes represent sites' administrative structures in addition to natural phenomena.

Original P2-3: Resource managers develop data curation processes and practices based on their curatorial mandates, access to resources, access to a site and to data collected at a site, and best practices at other similar sites. These factors lead them to prioritize curation of different aspects of a dataset over others.

Final P2-3: Resource managers develop data curation processes based on their curatorial mandates, access to resources, access to a site, access to data collected at a site, and best practices at other similar sites. These factors lead them to prioritize

curation of different aspects of a dataset over others. Their curatorial processes therefore represent their sites' administrative structures as much as their sites' natural structures.

Where my revisions to P2-2 were greatly informed by analysis through Coombs' theory of data, my revisions to P2-3 potentially reveal a shortcoming of his theory. Coombs' theory simply does not account for the work of curating and managing data collections over time, let alone the impact that administrative structures have on the selection, collection and curation of data. This points to a need for a theory of not just data, but of data collections. In future work, I hope to connect Coombs' theory of data with existing literature in LIS on the curation and development of data collections.

RQ3: How do these protocols and practices appear in the information models associated with datasets created by researchers and curators?

I originally drafted the following two propositions in response to this research question:

Original P3-1: Data collection protocols and curatorial processes are often not explicitly described in a dataset's metadata, and this makes datasets more difficult to curate or reuse.

Original P3-2: Data collection protocols and curatorial processes structure a dataset's underlying data model. These heterogeneous data models impact later attempts to aggregate, curate or reuse data.

These propositions are rooted in prior work on "small science" data curation (e.g. Cragin et al., 2010; Heidorn, 2008). Though I did not necessarily find that they were false over the course of my work, I found that issues of protocol and process representation and heterogeneity simply weren't as central to my work as originally anticipated. Rather, I found that data structure certainly impacted later usability, but perhaps not as strongly as I thought it might. I consequently lightly revised these propositions as follows, so as to soften my claims about the role of data structures in site-based data:

Final P3-1: Data collection protocols and curatorial processes are often not explicitly described in a dataset's metadata, and this impacts stakeholders' ability to curate or reuse data.

Final P3-2: Data collection protocols and curatorial processes structure the specific presence and absence of elements in a dataset, as well as the relationship between those elements. The heterogeneity of data models impacts later attempts to aggregate, curate, or reuse data.

RQ4: How can researchers and curators develop data collection protocols and curatorial processes that balance their respective needs of data from a scientifically significant site?

I drafted three propositions in response to this research question. The first concerned the data needs of different stakeholder groups; I did not substantially revise this proposition:

Final P4-1: Researchers and resource managers need different data and/or metadata about a site for their work.

The second proposition concerned the role of site-based information frameworks. The original proposition stated:

Original P4-2: Site-based data information frameworks make it possible to manage or mitigate heterogeneous data structures. Specifically, collections organized around aspects of sites' structure may be more usable over time.

As in the proposition regarding RQ3, I found that heterogeneous data structures played less of a role than I had anticipated. That said, I did find that site-based information frameworks made data collections more broadly useable. I revised this proposition accordingly:

Final P4-2: Site-based data information frameworks facilitate the management of data for diverse stakeholder groups over time. Specifically, collections organized around aspects of sites' structure may be more usable over time.

My third proposition in response to RQ3 concerned how stakeholders balance their needs of site-based data collections. The original proposition stated:

Original P4-3: Researchers and resource managers can balance their needs from a site through collaboration on research projects.

This held true; however, I additionally found that that stakeholders came to a common understanding of data collections through coordinated data cleaning and processing work, and I revised the proposition as follows:

> *Final P4-3: Researchers and resource managers can balance their needs from a site through collaboration on research projects, and/or coordination in data cleaning and data processing work.*

The revisions to P4-3 point to further potential connections between the work presented in this dissertation and research in the field of computer-supported cooperative work. Going forward, I hope to investigate the function of collaborative data entry, cleaning and management platforms such as databases and Google Spreadsheets in data curation over time.

RQ4a: What aspects of traditional natural history work practices might inform site-based standards development?

I originally drafted one proposition in response to this research question, which I did not substantially revise. This proposition is largely rooted in the work of Bruno Strasser:

> *Final P4a-1: There are existing site-based data collection and curation practices, such as the curatorial processes used by museums, community-developed data collection standards used by natural historians, or the databasing practices that led to the development of GenBank and the Protein DataBank, that are applicable to site-based data curation processes outside of museums.*

I additionally drafted the following proposition regarding generative reuse:

> *Final P4a-2: Natural history museums facilitate a kind of generative reuse, in which collections are improved through use over time; this is an important concept for development of data collections, whether site-based or not.*

This notion of "generative reuse" was discussed in detail in preceding sections of this chapter. I plan to continue work on this concept in the future.

Discussion of role of propositions in case study research

It is worth noting that not all case studies use or report their propositions. I do so here because I found them useful in my data collection and analysis, and therefore felt they merited reporting. During data collection, they were critical in the development of my interview protocols, and helped focus my interview questions. During data analysis, they helped focus my coding methods and quotation selection. Furthermore, over the course of their revision, I came to feel that they represented succinct responses to my research questions. I speculate that they may be useful in further work in site-based data curation research, whether by me or by other researchers. In future work, I will likely start with the revised propositions as presented in this section, and continue revising them going forward; other researchers in this field may find them useful in starting their work as well. On-going revision of propositions could facilitate the comparison of additional case studies to the cases presented herein, and thereby lay groundwork for further development of best practices and eventual theory development.

6.6 Directions for future work & Concluding remarks

The frameworks presented in this dissertation are intended to act as a guide for the development of future data collection protocols and information systems at each of my sites. In the future, I hope to continue working with resource managers at both of sites to implement the data collection protocols recommended through the minimum information frameworks. At YNP, resource management staff have already begun asking researchers to collect data according to our recommendations; however, they are interested in developing a more detailed data collection template similar to the EarthChem vent fluids template for use in the field (Erik Oberg, pers. comm., Jan 2017). This implementation could perhaps build on early SBDC work in which the Sustainable Environment through Actionable Data (SEAD) platform acts as a repository (Gordon et al., 2014); in our prior work with SEAD we explored the feasibility of photo-based browsing of data repositories and the use of site-specific metadata. I hope to expand on this prototyping work with the SEAD repository in the future.

At La Brea, I hope to continue my collaborations with collections staff and put my recommended changes to their excavation protocol into action. My research at La Brea has additionally pointed to several areas for other projects going forward. For instance, there is still a substantial collection of legacy data at La Brea that needs curation and cleaning for use in modern software

and in modern analyses. Migrating and transforming the existing collections data into a format usable by GIS systems could be of huge usefulness to the staff at La Brea, and could also answer questions about how best to integrate data cleaning into a curatorial workflow. Additionally, through my case study I gathered a partial history of the various information organization and data collection strategies used at the site; I would like to complete this history by interviewing previous staff members from the museum. This work would provide insight to ways that data curation structures are migrated and maintained over long periods of time.

This work has shown that the concept of data reuse is underdeveloped despite its prevalence in LIS literature. I identified several forms of data reuse through my case studies. Further work is needed to test the applicability of these kinds of reuse to other domains and sites. Better defining and scoping kinds of data reuse will be critical to the future development of data repositories and data curation processes.

Finally, this work has touched on issues of computer-supported cooperative work and distributed data work that merit further investigation. The stakeholders at my sites use a diverse range of software in their work, and have come up with several novel methods of circumventing their infrastructural constraints. Though it was out of scope for this study, I would like to investigate these strategies further in the future.

Concluding remarks

Scientifically significant sites are important centers for data curation, research coordination, and collaboration. Their stakeholders have distinct needs of information systems and data structures; these needs are underserved by existing best practices in data curation. Site-based datasets have unique potential for reuse, in that they are often collected with the same goal of describing the same sites, systems, or phenomena; however, there are challenges inherent in integrating these datasets because they are heterogeneously structured and idiosyncratically created. The use of standardized data collecting protocols could mitigate this heterogeneity, but it could also unnecessarily constrict researchers' study designs, independence, and creativity.

The research presented in this dissertation investigated stakeholders' needs for data curation, access, and use at scientifically significant sites. It confirmed the scientifically significant sites have unique characteristics worthy of study and support. It contributes further insight into the

values and priorities of two key groups of stakeholders, resource managers and researchers. It additionally contributes Minimum Information Frameworks for stakeholders at each of my sites: high-level information models that articulate what data elements are likely needed to facilitate reuse. These information frameworks are both organized around three classes of information: information about a Collecting Event; a Site Structure; and about Sample Sites and Measurements. Though rooted in just two case studies of scientifically significant sites, it is possible that this three-part structure may apply to other sites, and could be used to inform future infrastructure and standards development at a range of localities.

REFERENCES

Adair, J. R. (1997). The bioprospecting question: should the United States charge biotechnology companies for the commercial use of public wild genetic resources. *Ecology LQ*, *24*, 131.

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

Bates, J., Goodale, P., & Lin, Y. (2014). Data Journeys as an approach for exploring the socio-cultural shaping of (big) data: the case of climate science in the United Kingdom.

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, *35*(Database issue), D301-303. https://doi.org/10.1093/nar/gkl971

Biederman, P. W. (1993a, January 16). 25 Laid Off at Natural History Museum. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1993-01-16/local/me-1212_1_natural-history-museum

Biederman, P. W. (1993b, September 2). Museum Caught in Budgetary Morass : Cutbacks: Staff at the Page, famous for its La Brea Tar Pits, finds that layoffs have put a lot of research on hold and dramatically increased the workload of those who remain. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1993-09-02/news/we-30778_1_la-brea-tar-pits

Bishop, A. P., Van House, N. A., & Buttenfield, B. P. (Eds.). (2003). *Digital library use: social practice in design and evaluation*. Cambridge, Mass: MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information \ldots*, 1–40.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1–2), 17–30. https://doi.org/10.1007/s00799-007-0022-9

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., … Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, *29*(4), 365–371. https://doi.org/10.1038/ng1201-365

Brunt, J. W., & Michener, W. K. (2009). The Resource Discovery Initiative for Field Stations: Enhancing Data Management at North American Biological Field Stations. *BioScience*, *59*(6), 482–487. https://doi.org/10.1025/bio.2009.59.6.6

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

Carlson, S. (2006, June 23). Lost in a Sea of Science Data. *The Chronicle of Higher Education*, *52*(42), A35.

Cassell, C., & Johnson, P. (2006). Action research: Explaining the diversity. *Human Relations*, *59*(6), 783–814. https://doi.org/10.1177/0018726706067080

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. (2014). Enhancing metadata for research methods in data curation. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4. https://doi.org/10.1002/meet.2014.14505101103

Chao, T. C. (2015). *Methods metadata: curating scientific research data for reuse* (Doctoral Dissertation). University of Illinois at Urbana-Champaign, Urbana, IL. Retrieved from http://hdl.handle.net/2142/88180

Coombs, C. H. (1964). *A Theory of Data*. New York: John Wiley & Sons.

Cragin, M. H., Chao, T. C., & Palmer, C. L. (2011). Units of Evidence for Analyzing Subdisciplinary Difference in Data Practice Studies. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 441–442). New York, NY, USA: ACM. https://doi.org/10.1145/1998076.1998175

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *368*(1926), 4023–38. https://doi.org/10.1098/rsta.2010.0165

Current Research. (2015, June 24). Retrieved April 5, 2017, from https://tarpits.org/research-collections/current-research

Curtis, B., Kellner, M., & Over, J. (1992). Process modeling. *Communications of the ACM - Special Issue on Analysis and Modeling in Software Development*, *35*(9), 75–90. https://doi.org/10.1145/130994.130998

Darch, P. T. (2014). Ship Space to Database: Motivations to Manage Research Data for the Deep Subseafloor Biosphere. In *Proceedings of the American Society for Information Science and Technology* (Vol. 51). Seattle, WA.

Delson, E., Harcourt-Smith, W. E. H., Frost, S. R., & Norris, C. A. (2007). Databases, data access, and data sharing in paleoanthropology: First steps. *Evolutionary Anthropology: Issues, News, and Reviews*, *16*(5), 161–163. https://doi.org/10.1002/evan.20141

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.

Dilek, Y., Furnes, H., & Muehlenbachs, K. (Eds.). (2008). *Links between geological processes, microbial activities & evolution of life: microbes and geology*. [Dordrecht] ; London: Springer.

Douglass, K., Allard, S., Tenopir, C., Wu, L., & Frame, M. (2014). Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. *Journal of the Association for Information Science and Technology*, *65*(2), 251–262. https://doi.org/10.1002/asi.22988

Edwards, P. N., Mayernik, M. S., Batcheller, a. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

El Shafie, S. (2016). Bones in the bell tower. Retrieved April 5, 2017, from http://berkeleysciencereview.com/article/bones-bell-tower/

Faniel, I. M., Barrera-Gomez, J., Kriesberg, A., & Yakel, E. (2013). A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists. *IConference*, 797–800. https://doi.org/10.9776/13391

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

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *The Bulletin of the Ecological Society of America*, *86*(3), 158–168. https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Feldman, P. (1995, July 18). Ex-Official at County Museum Denies Embezzling $2 Million : Crime: The former deputy director and two aides are charged with stealing from the financially strapped Museum of Natural History. All three remain in jail. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1995-07-18/local/me-25067_1_natural-history

Flicka, user: (2007). *Deutsch: Castle Geysir bei Ausbruch. Wyoming, USA*. Retrieved from https://commons.wikimedia.org/wiki/File:Yellowstone_Castle_Geysir.jpg

Fore, J., Wiechers, I. R., & Cook-Deegan, R. (2006). The effects of business practices, licensing, and intellectual property on development and dissemination of the polymerase chain reaction: case study. *Journal of Biomedical Discovery and Collaboration*, *1*(1), 7.

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., Farmer, J. D., Marais, D. J. D., Pratt, L., Sturchio, N. C., Burns, P. C., & Discipulo, M. K. (2000). Depositional Facies and Aqueous-Solid Geochemistry of Travertine-Depositing Hot Springs (Angel Terrace, Mammoth Hot Springs, Yellowstone

National Park, U.S.A.). *Journal of Sedimentary Research*, *70*(3), 565–585. https://doi.org/10.1306/2DC40929-0E47-11D7-8643000102C1865D

Fouke, B. W., Palmer, C. L., Thomer, A. K., Dilauro, T., Gordon, S. C., & Hendrix, C. L. (2014, October). *Fostering Interdisciplinary Science through Data Curation: Geobiology at Yellowstone National Park as Exemplar.* Presented at the 12th Biennial Scientific Conference on the Greater Yellowstone Ecosystem, Mammoth, Wyoming.

Frank, R. D., Kriesberg, A., & Yakel, E. (2015). Looting Hoards of Gold and Poaching Spotted Owls: Data Confidentiality Among Archaeologists & Zoologists. *Proceedings of the American Society for Information Science and Technology*.

Fry, J. (2006). Scholarly research and information practices: a domain analytic approach. *Information Processing & Management*, *42*(1), 299–316. https://doi.org/10.1016/j.ipm.2004.09.004

General Conditions for Scientific Research and Collecting Permit - Yellowstone National Park (U.S. National Park Service). (n.d.). Retrieved March 20, 2017, from https://www.nps.gov/yell/learn/nature/npsconditions.htm

Glionna, J. M. (1995, July 13). 3 Charged in Embezzlement at L.A. Museum. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1995-07-13/news/mn-23539_1_museum-foundation

Gordon, S. C., Thomer, A. K., Dilauro, T., Jett, J. G., Fouke, B. W., & Palmer, C. L. (2014). Site Based Data Curation : Developing a Data Portal for Geobiologists at Yellowstone National Park. *Proceedings of the American Society for Information Science and Technology*, *51*, 1–4. https://doi.org/10.1002/meet.2014.14505101167

Greater Yellowstone Ecosystem - Yellowstone National Park (U.S. National Park Service). (n.d.). Retrieved March 20, 2017, from https://www.nps.gov/yell/learn/nature/ecosystem.htm

Greene, E. (2011). Why Keep A Field Notebook. In M. R. Canfield (Ed.), *Field notes on science & nature*. Cambridge, Mass: Harvard University Press.

Grimm, E. (2008). *Neotoma: An Ecosystem Database for the Pliocene, Pleistocene, and Holocene* (llinois State Museum Scientific Papers No. E Series 1). Retrieved from https://www.neotomadb.org/uploads/NeotomaManual.pdf

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., … Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/10.1890/120103

Harris, J. (2007, June). Bones from the Tar Pits. *Natural History Magazine*. Retrieved from http://www.naturalhistorymag.com/htmlsite/master.html?http://www.naturalhistorymag.com/htmlsite/0607/0607_feature.html

Hayes, G. R. (2011). The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction*, *18*(3), 1–20. https://doi.org/10.1145/1993060.1993065

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Hey, A. J. G. (Ed.). (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

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

History (U.S. National Park Service). (n.d.). Retrieved March 18, 2017, from https://www.nps.gov/aboutus/history.htm

hot spring geology. (2016). In *Encyclopedia Britannica*. Retrieved from https://www.britannica.com/science/hot-spring

Hou, C.-Y., Thompson, C. A., & Palmer, C. L. (2014). Profiling open digital repositories in the atmospheric and climate sciences: An initial survey: Profiling Open Digital Repositories in the Atmospheric and Climate Sciences: An Initial Survey. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4. https://doi.org/10.1002/meet.2014.14505101121

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., … Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature*, *455*(7209), 47–50. https://doi.org/10.1038/455047a

Karasti, H., & Baker, K. S. (2008). Digital Data Practics and the Long Term Ecological Research Program Growing Global. *International Journal of Digital Curation*, *3*(2), 42–58. https://doi.org/10.2218/ijdc.v3i2.57

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological

Research (LTER) Network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321–358. https://doi.org/10.1007/s10606-006-9023-2

Kent, W. (1978). *Data and reality: basic assumptions in data processing reconsidered*. Amsterdam ; New York : New York: North-Holland Pub. Co. ; sole distributors for the U.S.A. and Canada Elsevier/North-Holland.

Khoo, M., & Rosenberg, G. (2015). Historical Considerations in Biodiversity Informatics. In *iConference Proceedings* (pp. 1–15). Retrieved from https://www.ideals.illinois.edu/handle/2142/73440%5Cnhdl.handle.net/2142/73440

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology*, *45*, 247–294. https://doi.org/10.1002/aris.2011.1440450113

Kubota, K. (2013). CARD-FISH for Environmental Microorganisms: Technical Advancement and Future Applications. *Microbes and Environments*, *28*(1), 3–12. https://doi.org/10.1264/jsme2.ME12107

La Brea Tar Pits FAQs. (2015, June 24). Retrieved April 5, 2017, from https://tarpits.org/la-brea-tar-pits/faqs

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Lau, F. (1999). Toward a Framework for Action Research in Information Systems Studies. *Information Technology & People*, *12*(2), 148–176. https://doi.org/10.1108/09593849910267206

Lewin, K. (1943). Forces behind food habits and methods of change. *Bulletin of the National Resource Council*, (108), 35–65.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). New York, NY, USA: ACM. https://doi.org/10.1145/1940761.1940818

Mayernik, M. S., Wallis, J. C., & Borgman, C. L. (2013). Unearthing the Infrastructure: Humans and Sensors in Field-Based Scientific Research. *Computer Supported Cooperative Work (CSCW)*, *22*(1), 65–101. https://doi.org/10.1007/s10606-012-9178-y

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, *7*(1), 330–342. https://doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Millerand, F., & Baker, K. S. (2010). Who are the users? Who are the developers? Webs of users and developers in the development process of a technical standard. *Information Systems Journal*, *20*(2), 137–161. https://doi.org/10.1111/j.1365-2575.2009.00338.x

Mogk, D. (n.d.). Field Notes. Retrieved from http://www.minsocam.org/msa/Monographs/Mngrph_03/MG003_047-052.pdf

Mounce, R. (2014). Open Data and Palaeontology. In S. A. Moore (Ed.), *Issues in Open Research Data* (p. 151). London: Ubiquity Press. Retrieved from http://dx.doi.org/10.5334/ban.j

National Environmental Methods Index. (n.d.). Retrieved November 24, 2016, from https://www.nemi.gov/home/

National Science and Technology Council. (2009). *Harnessing the power of digital data for science and society.* (Report of Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council.). Science and Technology Council.

National Science and Technology Council, Interagency Working Group on Scientific Collections (Ed.). (2009). *Scientific collections: mission-critical infrastructure for federal science agencies: a report of the Interagency Working Group on Scientific Collections (IWGSC)*. Washington DC: Office of Science and Technology Policy.

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century* (No. NSB-05-40). National Science Foundation. Retrieved from https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

NPS Archeology Program: Antiquities Act of 1906. (2016, March 15). Retrieved March 19, 2017, from https://www.nps.gov/archeology/tools/laws/antact.htm

NPS: Explore Nature » Benefits Sharing. (2017, January 4). Retrieved March 21, 2017, from https://www.nature.nps.gov/benefitssharing/

Paleontological Laws and Policy. (2015, April 28). Retrieved March 20, 2017, from https://www.blm.gov/wo/st/en/prog/more/CRM/paleontology/paleontological_regulations.print.html

Palmer, C. L., & Cragin, M. H. (2008). Scholarship and disciplinary practices. *Annual Review of Information Science and Technology*, *42*(1), 163–212. https://doi.org/10.1002/aris.2008.1440420112

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

Palmer, C. L., Thomer, A. K., Baker, K. S., Wickett, K. M., Varvel, V., Choudhury, G. S., … Rodman, A. (2013). Building a Framework for Site-Based Data Curation. In *Proceedings of ASIST 2013* (pp. 1–4). https://doi.org/10.1002/meet.14505001144

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801174/full

Pasquetto, I., Randles, B., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*(0). https://doi.org/10.5334/dsj-2017-008

Perry, D. E., Sim, S. E., & Easterbrook, S. M. (2004). Case studies for software engineers. In *26th International Conference on Software Engineering, ICSE 2004* (Vol. 26, pp. 736–738). Edinburgh, Scotland. https://doi.org/10.1109/ICSE.2004.1317512

Peters, M., & Robinson, V. (1984). The Origins and Status of Action Research. *The Journal of Applied Behavioral Science*, *20*(2), 113–124. https://doi.org/10.1177/002188638402000203

Prideaux, G. J., Long, J. a, Ayliffe, L. K., Hellstrom, J. C., Pillans, B., Boles, W. E., … Warburton, N. M. (2007). An arid-adapted middle Pleistocene vertebrate fauna from south-central Australia. *Nature*, *445*(7126), 422–425. https://doi.org/10.1038/nature05471

Renear, A. H., & Dubin, D. (2008). Three of the four FRBR group 1 entity types are roles, not types. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–19. https://doi.org/10.1002/meet.1450440248

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4.

Sacchi, S., & Wickett, K. M. (2012). Taking modeling seriously [in digital curation]. *IPres: Research Challenges in Digital Preservation*, 14–16.

Sepkoski, D. (2012). Towards "A Natural History of Data": Evolving Practices and Epistemologies of Data in Paleontology, 1800–2000. *Journal of the History of Biology*, *46*(3), 401–44. https://doi.org/10.1007/s10739-012-9336-6

Shaw, C. A. (1982). Techniques Used in Excavation, Preparation, and Curation of Fossils From Rancho La Brea. *Curator: The Museum Journal*, *25*(1), 63–77. https://doi.org/10.1111/j.2151-6952.1982.tb00583.x

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

Steinhardt, S. B., & Jackson, S. J. (2014). Reconciling rhythms: plans and temporal alignment in collaborative scientific work. *CSCW'14*, 134–145. https://doi.org/10.1145/2531602.2531736

Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation Work: Cultivating Vision in Collective Practice. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 443–453. https://doi.org/10.1145/2675133.2675298

Strasser, B. J. (2011). The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine. *ISIS*, *102*(1), 60–96.

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 85–7. https://doi.org/10.1016/j.shpsc.2011.10.009

Swan, A., & Brown, S. (2008). The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. *A Report to the Joint Information Systems Committee (JISC), July*, 34.

Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., … Wiemann, S. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, *26*(8), 889–896. https://doi.org/10.1038/nbt.1411

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thomer, A. K. (2009). *Project 23 at Rancho La Brea*. Presented at the Western Association of Vertebrate Zoology, Holbrook, AZ.

Thomer, A. K., Baker, K. S., Jett, J. G., Gordon, S. C., & Palmer, C. L. (2014). Linking Geobiology Fieldwork and Data Curation Through Workflow Documentation. *AGU Fall Meeting Abstracts*, *41*. Retrieved from http://adsabs.harvard.edu/abs/2014AGUFMIN41A3644T

Thomer, A. K., Baker, K. S., Sacchi, S., & Dubin, D. (2012). Completeness, Coverage & Equivalence in Scientific Data Records. *Proceedings of the American Society for Information Science and Technology*, 1–4. https://doi.org/10.1002/meet.14504901331

Thomer, A. K., & Farrell, A. (2011). Field note digitization at Rancho La Brea -- Preliminary Case Study and Framework for Future Work. Presented at the Society for the Preservation of Natural History Collections, San Francisco, CA.

Thomer, A. K., Gordon, S. C., Dilauro, T., Baker, K. S., Jett, J. G., Palmer, C. L., & Fouke, B. W. (2014, October). *Aggregating and Integrating Geobiological Data from Yellowstone National Park: A Prototype Data Portal.* Presented at the 12th Biennial Scientific Conference on the Greater Yellowstone Ecosystem, Mammoth, Wyoming.

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

Thomer, A. K., Thara, T. S., & Wilson, M. D. (2009). *Excavation and 3-Dimensional Data Visualization at the La Brea Tar Pits*. Presented at the Computer Applications and Quantitative Methods in Archeology Conference, Williamsburg, VA.

Thomer, A. K., & Twidale, M. B. (2014). How Databases Learn. *IConference 2014 Proceedings*, 1–4. https://doi.org/10.9776/14409

Thomer, A. K., Vaidya, G., Guralnick, R., Bloom, D., & Russell, L. (2012). From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. *ZooKeys*, *209*, 235–253. https://doi.org/10.3897/zookeys.209.3247

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A., & Thiers, B. (2012). Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys*, *209*, 103–113. https://doi.org/10.3897/zookeys.209.3125

Vaughan, D. (1999). The Role of the Organization in the Production of Techno-Scientific Knowledge. *Social Studies of Science*, *29*(6), 913–943. https://doi.org/10.1177/030631299029006005

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

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

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

Weidman, S., Arrison, T. S., & National Research Council (U.S.) (Eds.). (2010). *Steps toward large-scale data integration in the sciences: summary of a workshop*. Washington, D.C: National Academies Press.

Wickett, K. M., Renear, A. H., & Furner, J. (2011). Are collections sets? *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10. https://doi.org/10.1002/meet.2011.14504801145

Wickett, K. M., Sacchi, S., Dubin, D., & Renear, A. H. (2012). Identifying Content and Levels of Representation in Scientific Data. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901199

Wickett, K. M., Thomer, A. K., Baker, K. S., DiLauro, T., & Asangba, A. E. (2013). How Workflow Documentation Facilitates Curation Planning. *AGU Fall Meeting Abstracts*, *40*. Retrieved from https://www.ideals.illinois.edu/handle/2142/50249

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

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.

Wood, M. (2000). Are National Park Resources for Sale: Edmonds Institute v. Babbitt. *Pub. Land & Resources L. Rev.*, *21*, 201.

Yarmey, L., & Baker, K. S. (2013). Towards Standardization: A Participatory Framework for Scientific Standard-Making. *International Journal of Digital Curation*, *8*(1), 157–172. https://doi.org/10.2218/ijdc.v8i1.252

Yellowstone Permit Conditions - Yellowstone National Park (U.S. National Park Service). (2016). Retrieved March 20, 2017, from https://www.nps.gov/yell/learn/nature/ynpconditions.htm

Yin, R. K. (2009). *Case study research: Design and methods.* (4th ed.). SAGE Publications Ltd.

Yin, R. K. (2012). Part I. Starting points. In *Applications of Case Study Research* (pp. 3–20). SAGE.

Zimmerman, A. S. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16. https://doi.org/10.1007/s00799-007-0015-8

Zimmerman, A. S. (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. https://doi.org/10.1177/0162243907306704

APPENDIX A. INTERVIEW PROTOCOLS AND IRB FORMS

2013 YNP WORKSHOP FOCUS GROUP DISCUSSION GUIDE

Below are the discussion questions that guided two focus groups conducted on April 17, 2013. One focus group included resource managers from YNP; the other, researchers who conduct work within YNP. Please note that not all questions were asked due to time constraints with the groups.

Resource manager focus group questions

1) What short-term needs do you have for researchers' data? Long term?

What about for geobiology specifically?

What would be the next domain to approach after geobiology?

2) Are there important lessons from how artifacts and their associated digital data are being kept in the archives?

What if you have data without a document or an artifact – how does that fit into your system?

What are you learning now at the HRC about how to manage the data that researchers' do provide you?

3) YNP has been referred to as a model of best practice for other park service. We'd like to think that if we can come up with some good practices about digital data that other parts of the NPS might pay attention. How do you think that might be facilitated? Are there professional communities within the park service?

What would the barriers to dissemination be?

4) We've talked a bit in the last two days about strategic planning for science. I wonder if you've thought any more about our discussion on this, and how that might play into short or long term goals?

What do you need to know about work by independent researchers at YNP for strategic planning for science? Yesterday there was some mention of understanding the spatial overlap of sampling – is there more you want to say about this? Are there other things

that you could understand better with a coordinated view of research going on in the park?

5) How can current guidelines and processes best be leveraged?

What parts of the permitting processed can be modified or augmented? what are the possibilities with IARs and permitting? These are our two points of intervention: permitting and publication.

6) If we were going to provide you with materials or recommendations to provide researchers getting permits, what would those look like? Should it be a strong recommendation that they take geolocation?

What kinds of outreach could be useful in helping scientists?

7) Who needs to be involved beyond who has been in the room? How do we engage them?

8) If you had to pick one big win for this project, in terms of your own interests and professional roles at YNP, what would it be?

9) What else haven't we talked about that you think that we should have?

Researcher focus group questions

1) As was raised yesterday, all YNP data may not be equally valuable for re-use purposes.

What bodies of data that might be potentially high impact, or particularly re-usable by others?
Are there data that would be particularly complementary to what is currently being provided through the RCN?

2) We know geo / temporal elements are vital to your science – the where and when of data collection. What aspects of "how" are also critical?

3) What types of relationships need to be made explicit among data sets?

Example include NSIDC – historical photos and current atmospheric for climate; Projects; Physical, biological, chemical?

4) To us, understanding site based curation at YNP should have implications for other scientifically significant locations, such as coral reefs, deep earth observatory, etc. Do you see any unique aspects about YNP that would not generalize?

5) How do you think policies or guidelines to support curation are likely to be received by the larger community of YNP researchers? Such as expectations about core metadata and submitting data to a repository for public access?

6) If you had to pick one big win for this project, in terms of your own interests and the science you conduct at YNP, what would it be?

7) What else haven't we talked about that you think we should be considering?

Potential issues include: Semantic web; Larger universe of networked data; International interoperability.

LA BREA ACTION RESEARCH CONSENT FORM

To Whom It May Concern:

You are invited to participate in a research study. The goal of this study is to assess the role of data collection standards and protocols at scientifically significant sites, and to create a prototype of a new data collection standard for *possible* use at the La Brea Tar Pits. This research will be used in Andrea Thomer's dissertation (under the supervision of Professor Michael Twidale), and may be published in other academic outlets.

About the project

Scientifically significant sites, such as national parks and research localities managed by museums, are important points for data management and curation. However the majority of data curation research has focused on the needs of academic institutions and individual researchers – not necessarily the needs of site-based resource managers, curators and other staff, or the needs of researchers reusing data from a specific site.

This study seeks to contribute to site-based data curation best practices by exploring how data collection protocols impact later data use, and how site resource managers and visiting researchers negotiate and navigate their different needs of data. I am particularly interested in how researchers and managers represent study sites in their data collection protocols and curation processes, and how data collection practices impact later data structures.

What participation entails

Over the next 6 months, I will be working with staff at La Brea to understand their data collection practices, and to understand what they like and dislike of the existing data collection and curation protocols. Through this work, we will create a prototype of a new data collection protocol for possible use at the site. I invite you to participate in this work.

If you consent, we will work over the next week to create a first draft of a new excavation protocol. Then I will begin drafting a propotype of a new protocol, possibly consulting with you from time to time over the week for feedback, as you are available. With your consent, I may

take notes about our work together, and use them in my research; I may also take photographs of your excavation site or work environment.

After I leave La Brea, I will interview researchers who use data from La Brea to understand what information they need the most. These interviews will inform further refinement of the protocol. I may consult with you by email to discuss any changes. I may refer to any email correspondence we have about this project in my analysis, and quote from it in my dissertation and resulting publications.

Finally, I will return to La Brea in late summer or early fall, to present the revised draft of the protocol, and collaborate further with you and others to make additional changes as necessary.

Confidentiality and the right to end participation

Your participation is entirely voluntary. You may withdraw your consent at any time by emailing Andrea Thomer (thomer2@illinois.edu).

Your name will not be used in my dissertation or any resulting publications; rather, you will be given a pseudonym instead, or referred to by a generalized description of your job (e.g. "a member of the curatorial staff" or "a staff member involved in excavation activities"). However, because I will be naming the La Brea Tar Pits as my study site explicitly, there is still a very minimal chance that someone in your field or workplace may be able to identify who said what. To ensure that you are comfortable with how you and your words are portrayed, I would be happy to give you the opportunity to review and suggest revisions to any quotes or descriptions of your work in my publications. If you do not wish to take the time to review these quotes, you are free to pass on this option as well.

What are the risks of participation in this project?

There are no risks involved in participating in this study other than those involved in everyday life. We will be discussing aspects of your job that you likely already discuss and consider in your day-to-day work. I will only be working on-site for two non-concurrent weeks with the support of La Brea administrators, and I will not disrupt your work. You are free to participate as much or as little as you wish. The excavation protocol we draft will only be a prototype for possible use, and you and your colleagues are free to reject or act on any part of it as you wish!

Will my study-related information be kept completely confidential?

In general, we will not tell anyone any information about you. When this research is discussed or published, no one will know that you were in the study. However, laws and university rules might require us to disclose information about you. For example, if required by laws or University Policy, study information which identifies you and the consent form signed by you may be seen or copied by the following people or groups:

- The university committee and office that reviews and approves research studies, the Institutional Review Board (IRB) and Office for Protection of Research Subjects;
- University and state auditors, and Departments of the university responsible for oversight of research;
- Federal government regulatory agencies such as the Office of Human Research Protections in the Department of Health and Human Services;

You may retain a copy of this consent form for your records. If you have any questions / comments about this study or are interested in the results, please direct your inquiry to Janet Eke. If you have any questions about your rights as a participant in this study, please contact the University of Illinois' Institutional Review Board at 217.333.2670 or via email at irb@illinois.edu.

Andrea Thomer – thomer2@illinois.edu
Professor Michael B. Twidale – twidale@illinois.edu
Graduate School of Library and Information Science
501 E. Daniel St.
Champaign, IL 61820

By signing this document, you verify that you have read and understood this consent form, are at least 18 years of age.

Please indicate below whether and in what ways you are willing to participate in this project:

☐ Yes ☐ No I allow researchers to photograph my work site

☐ Yes ☐ No I allow researchers to use non-identifiable photographs of my work site in the publications or presentations

☐ Yes ☐ No I allow the researchers to use short, pseudonymized quotes from work discussions and emails in publications and presentations

☐ Yes ☐ No I would like to review any quotes derived from my emails or other interactions prior to their publication

Signature _____ **Date** _____

LA BREA INTERVIEW CONSENT FORM

To Whom It May Concern:

You are invited to participate in a research study. The goal of this study is to assess the role of data collection standards and protocols at scientifically significant sites, and to create a prototype of a new data collection standard for *possible* use at the La Brea Tar Pits. This research will be used in Andrea Thomer's dissertation (under the supervision of Professor Michael Twidale), and may be published in other academic outlets.

About the project

Scientifically significant sites, such as national parks and research localities managed by museums, are important points for data management and curation. However the majority of data curation research has focused on the needs of academic institutions and individual researchers – not necessarily the needs of site-based resource managers, curators and other staff, or the needs of researchers reusing data from a specific site.

This study seeks to contribute to site-based data curation best practices by exploring how data collection protocols impact later data use, and how site resource managers and visiting researchers negotiate and navigate their different needs of data. I am particularly interested in how researchers and managers represent study sites in their data collection protocols and curation processes, and how data collection practices impact later data structures.

What participation entails

Over the next 6 months, I will be working with staff at La Brea to understand their data collection practices, and to understand what they like and dislike of the existing data collection and curation protocols. Through this work, we will create a prototype of a new data collection protocol for possible use at the site. I would like to interview you about your work at La Brea, so that we might develop a protocol that takes into account the varied needs excavation and collections staff have of the La Brea collections and their associated data.

This interview will approximately 30-60 minutes. I will ask you questions such as:

- What aspects of the existing excavation protocol do you like? Which do you dislike?
- What aspects of existing curatorial practices do you like? Which do you dislike?

- What is the most difficult part of your job with regards to data?

Your answers to these questions will inform our development of a prototype of a new excavation protocol. At the end of the interview, I may ask you if I can contact you again in approximately 6 months to review the prototype; you are free to decline, either now or in the future.

Confidentiality and the right to end participation

Your participation is entirely voluntary. You may withdraw your consent at any time by emailing Andrea Thomer (thomer2@illinois.edu).

Your name will not be used in my dissertation or any resulting publications; rather, you will be given a pseudonym, or referred to by a generalized description of your job (e.g. "a member of the curatorial staff" or "a staff member involved in excavation activities"). However, because I will be naming the La Brea Tar Pits as my study site explicitly, there is still a very minimal chance that someone in your field or workplace may be able to identify what you said. To ensure that you are comfortable with how you and your words are portrayed, I would be happy to let you review and suggest revisions to any quotes or descriptions of your work in my publications. If you do not with to take the time to review these quotes, you are free to pass on this option as well.

I will not share any identifiable quotes from this interview with any of your coworkers. However, I may share pseudonymized quotes if necessary for our development of the excavation protocol.

What are the risks of participation in this project?

There are no risks involved in participating in this study other than those involved in everyday life. We will be discussing aspects of your job that you likely already discuss and consider in your day-to-day work.

Will my study-related information be kept completely confidential?

In general, we will not tell anyone any information about you. When this research is discussed or published, no one will know that you were in the study. However, laws and university rules might require us to disclose information about you. For example, if required by laws or

University Policy, study information which identifies you and the consent form signed by you may be seen or copied by the following people or groups:

- The university committee and office that reviews and approves research studies, the Institutional Review Board (IRB) and Office for Protection of Research Subjects;
- University and state auditors, and Departments of the university responsible for oversight of research;
- Federal government regulatory agencies such as the Office of Human Research Protections in the Department of Health and Human Services;

You may retain a copy of this consent form for your records. If you have any questions / comments about this study or are interested in the results, please direct your inquiry to Michael Twidale. If you have any questions about your rights as a participant in this study, please contact the University of Illinois' Institutional Review Board at 217.333.2670 or via email at irb@illinois.edu.

Andrea Thomer – thomer2@illinois.edu
Professor Michael B. Twidale – twidale@illinois.edu
Graduate School of Library and Information Science
501 E. Daniel St.
Champaign, IL 61820

By signing this document, you verify that you have read and understood this consent form, are at least 18 years of age.

Please indicate below whether and in what ways you are willing to participate in this project:

☐ Yes ☐ No I consent to an interview lasting between 30-60 minutes, and I allow the researchers to use pseudonymized quotes from interviews in publications and presentations

☐ Yes ☐ No I consent to audio recording of this interview.

☐ Yes ☐ No I consent to being contacted for a "follow up" interview after this initial interview is concluded. I understand that I may decline a second interview even if I consent to being contacted for one.

☐ Yes ☐ No I would like to review any quotes attributed to my pseudonym before they are published. I understand that this may entail an additional time commitment from me.

Signature _____ **Date** _____

INTERVIEW PROTOCOLS

La Brea - Resource Managers

These first questions are trying to find out a little bit about your work:

1) Could you tell me a bit about your work at La Brea?
   [follow up/detailed questions below if necessary]
   a. What is your job title and how long have you had that job? [maybe keep this short]
   b. What are your primary priorities or goals in your work?
   c. Has your job/work changed over time?
2) How do you see your work as fitting within the broader work of the museum?

About your work with data:

2) In your view, what is most significant about Rancho La Brea as a research site?
3) With keeping that "most significant" feature in mind, of the data you record/manage, which is the most vital?
   a. How does the data you record represent the site overall?
   b. [maybe skip] In your view, what is the primary value of the measurement data collected with each fossil? In other words, why do you think it's important to collect the measurement data? [be sensitive that this could be leading]
   c. [maybe skip] In your view, how does this system support the use and curation of fossils over time?
4) Think of your work with La Brea data and metadata over time:
   a. What data do you or most often refer back to? What data have you seen other researchers or staff members refer back to?
      i. How often do you need to know/refer back to the specific measurements of individual fossils?
      ii. How often do you need to know/ refer back to the grid from which a fossil was excavated?
      iii. How often do you need to know/refer back to the pit/box from which a fossil was excavated?
      iv. How often do you need to know who excavated a fossil?
      v. How often do you need to know when a fossil was excavated
5) Over time, have there been cases where you wish you had *more* information about:
   a. a fossil?
   b. A fossil's preparation history?
   c. an excavation season?
   d. an excavation site?
   e. An excavator or other museum staff member?

f. Anything else?

6) What is the most difficult part of your job with regards to data collection or management?
    a. [do you ever make trade offs, or hard choices in your work --]
7) In addition to the La Brea excavation protocol, what best practices or standards do you use in your work?
    a. Any preparation standards?
    b. Any data formatting or recording standards?
    c. Any mapping, illustration, or photography best practices or standards?
8) *[maybe skip] Who do you learn from? Whose your community of practice?*
    a. Who are your closest collaborators at La Brea? At LACM?
    b. Who are your closest collaborators in paleontology or geology overall?
        i. How often do you communicate with other people in your field, how often, and what about?
        ii. Do you attend any meetings in your field (e.g. SVP, the prep conference).
    c. Do you often work with researchers visiting La Brea? If so, could you tell me a bit about this?

About the future:

9) If you could change anything about the excavation protocol, what would it be?
10) Is there anything about the current excavation protocol that you absolutely would *not* like to see changed?
11) What motivates your work? What makes you most excited or passionate about your job?

La Brea - Visiting Researchers

Questions about work over all, and background and work at La Brea:

1) Could you tell me a bit about your work at La Brea?
   [follow up/detailed questions if necessary]
   a. What is your job title/home institution, and how long have you been there?
   b. How long have you worked at/with the La Brea Tar Pits or its specimens?
      i. Do you study a particular organism? Particular morphological feature? Something else?
   c. What are your primary research questions addressed in your work?
   d. What's the most important aspect of La Brea as a site?

About your work with specimens and data:

2) Can you tell me about your work with physical *specimens* and *fossils* from La Brea?
   a. How do you decide what fossils to use? Do you use any additional physical materials (e.g. soil samples, oil samples) in your work?
      i. How many fossils/samples do you work with at a time?
   b. <u>What data about a fossil do you *need* in order to do your work? About each fossil's locality?</u>
   c. What data do you collect/measure from each fossil you work with (e.g. measurements, radiocarbon dates, images, something else)?
      i. How do you analyze that data?
   d. What data from La Brea do you use in your analysis, or publish with your analysis (e.g. specimen numbers, locality descriptions, anything else?)?
      i. Do you publish your data anywhere (e.g. via it to a repository such as Figshare, Dryad, an institutional repository)? Why or why not?
      ii. Do you share your data with La Brea collections staff? Why or why not?
3) La Brea has a very detailed data collection and excavation method:
   a. How much do you know about this method?
   b. Have you ever felt that the excavation and data collection method has impacted your work? If so, how?
4) La Brea also has a fairly unique fossil preparation method:
   a. How much do you know about these techniques?
   b. Have you ever felt that the preparation methods have impacted your work? If so, how?
5) Do you use any of the metadata collected/created through/about the excavation or preparation methods? Do you use any information about the:
   a. depth or specific location (grid) the fossil was recovered from?
   b. excavation site conditions?
   c. excavator who dug up the fossil?

    d. other fossils excavated on that day or in that field season?
    e. steps taken to prepare the fossil (clean off the tar)?
    f. steps taken to curate the fossil's data?
    g. Do you ever consult field notes describing a fossil?
6) Is there any information you *wish* you had about the fossils/specimens you work with?
7) In many fields, there are best practices or data standards that researchers rely on to format their data. Are there any best practices or standards you use in your work?
    a. Any data formatting or recording standards?
    b. Any mapping, illustration, or photography best practices or standards?
    c. Do you ever have any problems migrating La Brea data or metadata into these formats?

About your collaborations at La Brea:

8) Who are your closest collaborators at La Brea?
9) Do you consult with curatorial or excavation staff? If so, what do you discuss?
10) What keeps you coming back to La Brea? Why do you continue to work there?
    a. What's the most important aspect of La Brea as a site?

YNP – Visiting or External researchers (geobiology/astrobiology)

[These interviews are semi-structured and I may alter questions to better suit the specific background of the person I am interviewing; thus the questions below are representative rather than strictly prescriptive of the kinds of questions I will be asking my participants]

[FOR PHONE/SKYPE INTERVIEWS: begin with the following text:

Thank you for taking the time to participate in this research. I wanted to begin by once again verifying your permission to record this interview. You are free to (a) discontinue participation in the study at any time, (b) request that the audio recorder be turned off at any time, (c) request to speak "off the record" at any time, and (d) pass on any question you do not want to answer. Do you consent to this interview?]

1) I'd like to start by hearing about your work overall - could you tell me about your work at JPL/in geobiology/astrobiology? [follow up/detailed questions below if necessary]
    a. What is your primary field of study (e.g. biology, chemistry….)
    b. What is your job title/home institution, and how long have you been there?
    c. At what specific sites/localities do you most frequently work?
    d. Do you study a particular organism? Particular geological feature?
2) We're interested in understanding how researchers decide what parameters to collect at their study sites. Thinking back to your most recent project, how did you decide what data to collect?
    a. Did you use any data collecting standards in your work?
    b. Did you create a data management or curation plan prior to your work? Did that affect your study design?
    c. Are there particular features of your study sites that are important to describe, or collect data about? Do these important features differ between sites?
3) In our work at YNP, we found that researchers needed to know the temperature and ph of the water from which each sample was collected, in order to integrate data across sites – are these parameters relevant to your work?
    a. If so, please describe how
    b. If not, are there key parameters that *are* critical to data integration at your study sites?
4) Do you work with physical samples or specimens?
    a. If so, how do you acquire your samples/specimens?
    b. How do you decide what samples/specimens to work with?
    c. What data do you collect from each sample/specimen you work with (e.g. measurements, radiocarbon dates, images, something else)?
    e. How do you analyze that data?

5) Do you use data collected by other researchers in your work?
    a. If so, how do you assess whether or not that data is appropriate for your research? Are there specific parameters you need? Or data collected in a specific way?
    b. If not, why?

APPENDIX B. LIST OF YNP PERMITTING CONDITIONS

Yellowstone Permit Conditions 2016 – from
https://www.nps.gov/yell/learn/nature/ynpconditions.htm

1.You are responsible for the research-related activities of your staff. Please ensure that field staff adhere to all conditions of your permit. Field staff must possess a copy of your permit at all times while in the field.

2.You are required to post, via the Internet, your research trip itineraries no later than the Sunday prior to your trip. The following website has been established to facilitate this process: https://irma.nps.gov/rci/. Once working in the park, report all emergencies by calling 911.

3. You are required to have a Safety Plan on file that addresses the range of activities you will encounter while working in Yellowstone National Park. All personnel who enter Yellowstone to conduct work under this research permit must review their Safety Plan prior to beginning work. At minimum, a safety plan shall cover a) training requirements and documentation that personnel have received training appropriate to the work being conducted (examples: bear safety, bear spray use, thermal area safety, fording streams); b) work party size (note: travel in groups of 3 or more is recommended when hiking in bear country); c) safety equipment (examples: bear spray for each party member, rain gear, heat resistant gloves, extendable pole for sample collection in thermal areas); d) trip itinerary (example: provide supervisor with an itinerary specific to daily activities and travel patterns); e) worker check-in. Note: in addition to completing the online researcher check-in per Yellowstone Condition #2, it is advised that all field personnel designate an emergency contact (e.g., supervisor, co-worker) whom they will check-in with at the end of each field day or session. This designated emergency contact will know the trip details and will contact emergency services (911) and the Research Permit Office (307-344-2239) in the event field staff fail to make contact at the end of the day or reporting period.

4.While conducting permitted field research activities (unless otherwise authorized by the Superintendent), park researchers are prohibited from possessing firearms. Researchers are also prohibited from bringing firearms in government buildings and in government vehicles (cars, boats, aircraft).

5.Unless otherwise authorized on your permit, you must carry out all of your activities out of public view. If you have obtained special permission to work in public view, it will be noted on your permit-specific conditions. Please consult these conditions for further guidance.

6.If you are approved to collect specimens (either to be permanently retained or destroyed through analysis), please contact the Yellowstone Curator's Office (307-344-2565) to report your collections annually. Specimens must be tracked and an inventory provided to Yellowstone (count, type, and location) by February 28th following the permit year as part of the reporting process. Prior to collecting specimens, a repository form must be completed and on file. Any permanently retained specimens must bear accession and catalog numbers, and include the required metadata per the National Park Service's catalog system (discipline-specific classification information, scientific/object name, locality, current specimen location, collector, collection number, collector date, and identifier).

7.All equipment left in the field including plot markers must be specifically authorized in advance. Label all equipment with your name, phone number, and the words "Research Study #XXXX." If you are authorized to place equipment or plot markers in Yellowstone, you will be required to GPS their locations.

8.Your research permit does not authorize you to enter closed or restricted areas in Yellowstone. Examples of restricted areas include most service roads, carcass dump sites, bear management areas, thermal areas, some bird nesting areas, wolf den sites, and trout spawning areas.

9.Cultural resources must not be adversely impacted by your research activities. Ground disturbance (e.g., digging) must be specifically authorized in advance. Report any archeological findings (artifacts, historical trash, rock cairns) to the Research Permit Office.

10. The Permittee agrees to notify the Chief of Resources of Yellowstone National Park (YNP) of every subject discovery or invention that relates in any respect to research results derived from YNP research studies or use of any research specimens or other materials collected from YNP, or that may be patentable or otherwise protected under the intellectual property (IP) laws of the United States or other jurisdiction. Notification must occur within sixty (60) days of the time that an inventor or other agent of the Permittee reports such a subject discovery or invention to the person(s) responsible for patent or other proprietary rights matters in the Permittee's

organization. Additionally, the Permittee agrees to notify the Chief of Resources of Yellowstone within thirty (30) days of filing any patent application or other IP claim in the United States or other country that relates in any respect to research results or other discoveries or inventions derived from YNP research studies or any research specimens or other materials collected from YNP. For purposes of this paragraph, the term "subject discovery or invention" means any discovery or invention related to or derived from YNP research studies, or research specimens or other materials collected from YNP. All invention disclosures shall be marked as confidential under 35 U.S.C. Section 205.

11. All filming associated with this research permit must be reviewed and approved in advance by the park's Film Permit Office. Depending on the type of filming a Film Permit may be required. Filming of certain research activities may be used for education in the classroom setting or on private educational platforms which are password protected and the footage must clearly state that the research was conducted under a Yellowstone National Park NPS Research Permit. Any use, including social media, websites, newspapers, periodicals etc. of photos or videos from within closed areas or of research taking place in closed areas is prohibited without prior NPS approval. For further information contact Rachel Cudmore @ e-mail us (307) 344-2722, or Tammy Wert @ e-mail us (307) 344-2115.

12. Each year, investigators are required to submit copies of journal articles, theses, and dissertations that resulted from park research activities to the Research Permit Office. Documents should be submitted in PDF format, with the exception of theses/dissertations, which should be sent as bound documents.

APPENDIX C. NPS REPORTING FORMS AND DATA INFRASTRUCTURES

INVESTIGATOR ANNUAL REPORT

Investigator's Annual Reports (IARs) are *"mandatory year-end reports required from all Principal Investigators (PI's), who engage in science and resource management activities in the various parks. A wide range of technical disciplines are represented. Parks are listed alphabetically and a search utility is provided for retrieval of research summaries according to park, subject or names of the PI's."*[28] IARs from all parks and national monuments from approximately 1990 to present are publically available in the NPS Researcher Permit and Reporting System.[29] Data that can be reported in IARs include:

- Investigator Names, Addresses, and other contact information
- Project Name, Permit Number and Study Number
- Project Discipline (e.g. Geology, Biology, Education, etc)
- Project Status (e.g. continuing, completed)
- Project Objectives and Results
- Funding amount and source
- And an indicator of whether physical collections have been collected, and if so, where they are being stored.

Researchers complete their IARs with varying levels of thoroughness and detail, and NPS resource managers noted that they sometimes struggle to get researchers to submit them punctually by the annual NPS regulatory deadline of March 31. YNP resource managers ask for field surveys (Appendix 9) to be completed along with and IAR.

[28] http://www.nps.gov/glac/naturescience/for-researchers.htm
[29] science.nature.nps.gov/research/ac/ResearchIndex

THE YNP HOT SPRING "FIELD SURVEY" FORM

Yellowstone National Park Field Survey

1) Principal Investigator:

2) Study Number:

3) Dates in Yellowstone:

Location	UTM (NAD 83) (Easting)	UTM (NAD 83) (Northing)	# Days at this location	#Researchers at this location	Total researcher-days (item D x item E)	Specimen type & quantity from location	Equipment left at location	# Days Equipment Left in Field	Comments
Example 1: Roaring Mountain	521645.22	4958260.17	30	2	60	2, 15ml H20, 5 fist-sized rocks	2 dataloggers	35	
Example 2: Tantalus Creek	523024.1	4952604.18	4	1	4	None	None	na	Observations only at this site

YNP Resource Managers ask researchers who work at the hot springs to submit the above field survey form along with their IARs

PERMITTING AND DATA MANAGEMENT INFRASTRUCTURE AT YELLOWSTONE NATIONAL PARK

[this subsection was previously published in Thomer, Palmer, et al., 2014]

In interacting with scientific researchers, YNP personnel oversee a permitting process for research on site, maintain a tracking system for physical specimens, and manage site visits. Thus, the research office at Yellowstone serves two purposes: 1) permit administration and 2) resource protection in compliance with the National Environmental Policy Act (NEPA). The aim of this appendix is to describe existing permitting and data reporting policies, standards and computer systems, as well as to consider the context within which a site-based tracking system for digital research data and metadata may eventually be created. Any new application would best be designed to capitalize upon and coordinate with existing systems.

YNP makes use of NPS-wide digital information systems as well as on-site, stand alone, YNP-specific desktop applications and data collections (often collected and maintained by park personnel). One of the NPS-wide systems in use is IRMA, the Integrated Resource Management Application. IRMA brings together many formerly separate NPS-wide databases and reporting applications, such as: the Research Permit and Reporting System (RPRS), which manages permit applications and Investigator Annual Reports (IARs; see Appendix 11); NPSpecies, a database of species checklists for each park; and the NPS DataStore, a repository for publications, maps, theses, and some datasets related to the parks. Though IRMA does provide public, web-based access to its holdings, many of the resources in it are for internal use and viewing only. Additionally, while IARs are publically searchable and downloadable, permits and permit applications are for internal use only. We note that IRMA is a work in progress; functionality is being added on an ongoing basis. IRMA (via the DataStore) has historically been primarily used to store and aggregate manuscripts and digital documents, however, now that the RPRS system has been integrated into IRMA, it appears that researchers may be able to upload datasets along with their IARs. We are exploring whether there are file size and type constraints, as well as whether IRMAs search capabilities will be sufficient for researchers' needs.

Additionally, because it is NPS policy to collect, protect, preserve, provide access to, and use objects, specimens, and archival and manuscript collections to aid understanding and advance knowledge, many NPS data reporting systems and requirements concern physical

specimens and their associated data. The Heritage & Research Center (HRC) houses YNP's museum collection,[30] archives, research library,[31] and herbarium,[32] and accession and catalog numbers are tracked by an NPS-wide database, the Interior Collections Management System (ICMS). Specimens may be only collected by researchers issued a Scientific Research and Collecting permit, and though researchers may be granted permission to store the specimens at their home institution (typically through a 10-year, renewable loan agreement), these specimens remain the property of the NPS permanently and must bear NPS labels must be accessioned and cataloged in the NPS National Catalog (via ICMS). Regardless of where the specimens are housed, collectors must report specimen-related metadata to park collections managers, and participate in annual specimen inventories. This metadata includes the specimens' taxonomic identifications, number of specimens collected, location collected with UTM coordinates and date/time collected, method of preservation or preparation (e.g., herbarium sheet, preserved in alcohol/formalin, tanned and mounted, dried and boxed, destroyed through analysis, etc.), and current location.

NPS digital systems a range of tracking numbers to manage permits and documents associated with researchers and their projects. While some practices are NPS-wide, others are more park specific and ad hoc. Though some of the identifiers described below are used NPS-wide, some of the practices surrounding them are unique to YNP:

1) Application Number - The research permit application process is typically initiated via an email exchange of forms. Upon receipt of an application for a new project or a renewal of an existing project, an application number is issued.
2) Permit Number – In communications at the park, the permit number is the main project identifier used by park coordinators. Permit numbers are issued sequentially to applicants upon review and approval of permit applications. The permit number also appears on IARs. At the time of the workshop, the research permitting was managed by a stand-alone application. This process has since been migrated and is now carried out by IRMA.
3) Investigator Annual Report number - A permit is valid for one year but may be renewed via submission of a new application. An investigator is required to fill out an Annual Report (IAR) that is identified by the permit number.

---

[30] http://www.nps.gov/yell/historyculture/museum.htm
[31] http://www.nps.gov/yell/historyculture/collections.htm
[32] http://www.nps.gov/archive/yell/nature/plants/index.htm

4) Trip Number – There may be one or many trips to the park planned as part of a single project. For each trip, the researcher is required to fill out a web form about their trip (date, time, location) for safety reasons. Upon submission, a trip number is assigned. The trip may be seen either on the researcher check-in or on the researcher page.

The permit number and trip number represent potential mechanisms for organizing and linking external researchers' data via new digital data applications. That is, digital data logs, catalogs and collections eventually may be developed to manage the data of external researchers regardless of where development occurs - internal and/or external to the park.

The trip number is not publically available at this time. If the trip number were made public, it could potentially be used as an identifier for data relating to each park visit. Design discussions would be needed to determine whether the trip number or the permit number granularity is more appropriate for data identification.

APPENDIX D. RECOMMENDED CHANGES TO LA BREA EXCAVATION PROTOCOL

CHANGES TO THE ON-GOING P23 EXCAVATIONS AND FUTURE MITIGATIONS

FUTURE EXCAVATIONS SHOULD BE PLOTTED FROM THE SW CORNER

Researchers consistently had issues working with measurement data because of the way the grids are laid out in excavations. The "origin" point of the grids is in the southeast corner, but in most general-purpose mapping programs place the origin in the southwest corner of the plot. This means that the measurement data needs to be extensively reformatted before it can be analyzed.

Additionally, La Brea could begin reformatting and storing measurement data for its researchers. This option is discussed in the "In-house data curation" section.

EXPLORE DIFFERENT METHODS OF DOCUMENTING EACH DEPOSIT'S GEOLOGY

Several staff members felt the current method of documenting the deposits' geological context was ineffective. For the last 10-20 years, grid "wall" and "floor" sketches have been drawn by multiple people using a range of notations over long periods of time; this makes the sketches difficult to use and potentially unreliable. Additionally, the "wall" and "floor" sizes are more appropriate for an archaeological dig, not paleontology, and they are often too small to really capture the geological context of the deposit.

There's consequently a need to explore different methods of documenting deposit structure. The following methods were suggested (I've added comments about their potential efficacy):

- **Continue sketching walls and floors in their current dimensions, but assigning just one staff geologist to sketch them using standardized and consistent notation**. If going this route, I would recommend that the sketches be integrated into complete diagrams of the entire deposit floor or wall, as soon as possible. Integrating the sketches while the excavation is still on-going will make it easier to check any potentially uncertain areas with on the remaining deposit, and will reduce the risk that staff members will move on (taking their institutional memory with them) before the data can be integrated.
- **Taking core samples from *in situ* deposits**, even if they run the risk of damaging the bones. With the Project 23 deposits, this may additionally be a good way to determine

whether the deposit is still roughly intact, or if it is disturbed. Though fossils could be damaged in taking the core, this information loss may be more than made up for in the information gained in understanding the deposits' stratigraphy (or amount of disturbance).

- **Using new methods of photography** to document the excavation as it progresses; e.g. 3D scanning, LIDAR, etc. These kinds of methods may make it unnecessary to draw the walls and floor; however, it would be necessary to try them out before committing to their use.
- **Taking sides off still-intact boxes, to draw the stratigraphy of the entire deposit before it is excavated**. This was done for deposit 5B and worked fairly well. I would caution that additional safety measures should be taken if this were repeated in the future. For one thing, the deposits could collapse; for another, chiseling overhead can cause repetitive motion injuries (as happened with my shoulder while excavating 5B).

To summarize: whatever the method, grid walls and floors should be refined and integrated into whole-deposit sketches as soon as possible. One person with a background in geology should likely draw all of the sketches for the sake of consistency.

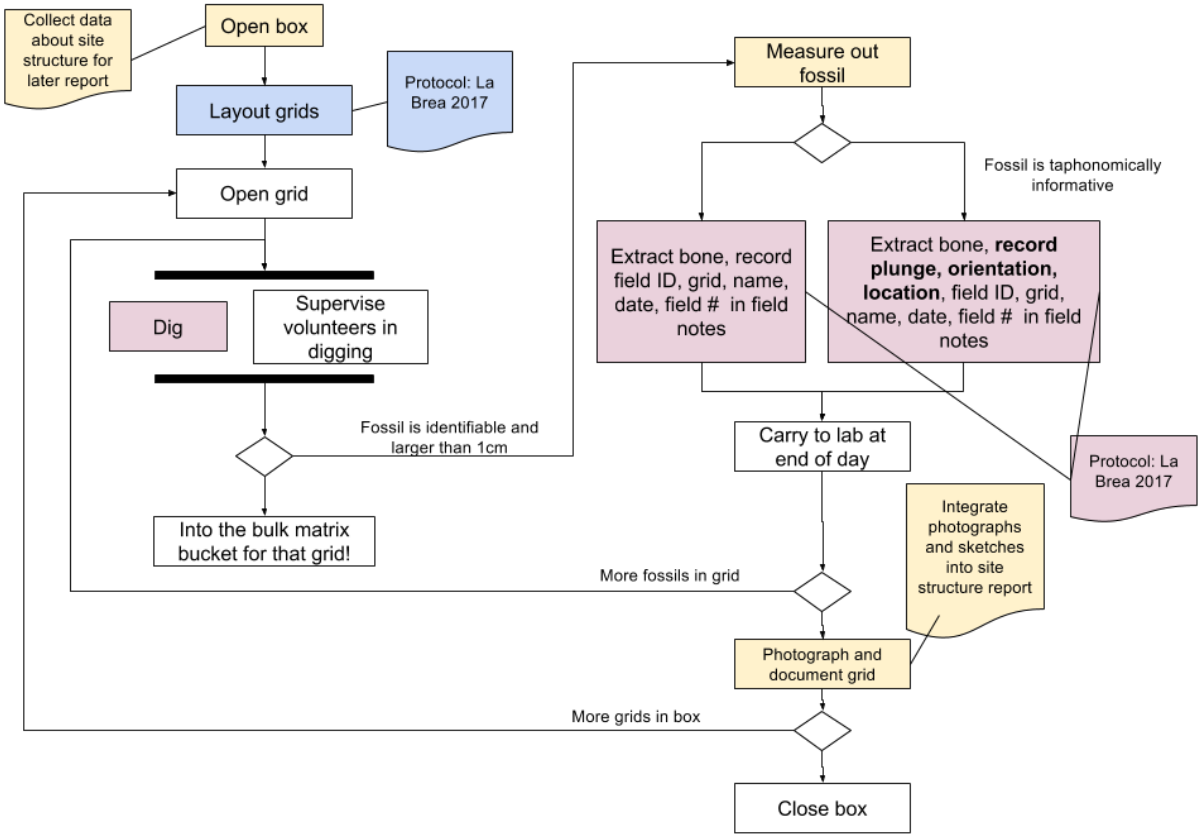
STREAMLINE THE FOSSIL MEASUREMENT METHOD TO FOCUS ON CREATING AN INVENTORY, COLLECTING USABLE TAPHONOMIC DATA

There are known problems with the present measurement system. Excavating fossils individually runs the risk of damaging them; and the measurement data creates a huge downstream backlog for curatorial/lab personnel. Though a fossil may "only" take a few minutes or so to measure, enter into a field book and bag, it creates hours of preparation, curation and databasing work for the lab, and likely slows down the excavation process prior to the moment it's measured out.

From my interviews, the measurements seem to have two primary functions:

1) They would support the reconstruction of the fossils' original position in the earth, which is important for *in situ* deposits with taphonomic structure. Many of the researchers felt it was good to collect just in case it became useful in the future. That said, some of the La Brea staff felt that this information wasn't necessary, or wasn't collected in a manner that was usable
2) They include field identifications, which are then entered into a spreadsheet by the excavators, and aggregated by the collections managers and used as a tentative inventory. This inventory is important as a rough estimate of unprepared fossil holdings, and can help the La Brea staff plan for curatorial work, storage space, and so on.

I've consequently tried to revise the data collection workflow to focus on updating the unprepared fossil inventory, and streamlining the collection of taphonomically informative data. I propose the following workflow:



This workflow divides fossils into 3 categories:

1) Indeterminate fossils & fossils less that 1cm in size: by indeterminate fossils, I mean fragments that could not be easily determined down to element and genus or species. **Indeterminate fossils and fossils under 1cm in size would go into the matrix bucket for bulk processing.**
2) Determinate fossils that are not taphonomically informative. These are fossils that are complete enough to merit a field identification but that would not be considered *taphonomically informativ*e, and therefore don't need to have positional data collected. These fossils could still be excavated individually, entered into field notebooks in the same manner as they have been (with carbon paper and an envelope or 3x5 card) along with their grid number, the field identification, and the date, but *without* any measurements. The field identification would allow the collections staff to continue maintaining a running inventory of excavated fossils, and the grid number would provide rough positional data.

3) Determinate fossils that are taphonomically informative. These are fossils that have been assessed as potentially informing the future interpretation of a deposit's taphonomy and depositional structure. Excavators should record the field identification, the grid number and some further positional data. **In lieu of the current measurement method, I propose that future excavations begin collecting:**
    a. **the slope of the fossil,**
    b. **the orientation of the fossil, and**
    c. **the fossil's rough location within the grid. This could be as simple as a 1-point measurement of the center of the fossil.**

A few caveats here:

- The goal of these recommendations is not to define what fossils should be considered taphonomically informative, but rather, to define the information that ought to be collected if fossils are deemed informative. The definition of a "taphonomically informative fossil" will need to be determined by the collections staff, will likely change from project to project, and even may change over time. For instance, in early Pit 91 excavations, all fossils and fossil fragments over one quarter inch in size were considered taphonomically informative; whereas in the beginning stages of the P23-1 excavation, only identifiable "3-point" fossils were. **Collections staff will have to determine whether a deposit is *in situ* first, and then should decide what fossils may be taphonomically informative.**
- Further changes to this workflow will be needed if the excavators move to pulling large blocks or jackets out of the deposits to more fully excavate in the prep lab. The biggest change might be to the databasing/inventory workflow. In the past, bulk "bone bags" have created problems in the KE Emu databasing process. The same issue could arise with blocks or jackets. I imagine that the excavators could either skip entering these records into their field books; skip entering them into spreadsheets; or the collections managers could skip migrating the records into EMu. In short: **it would be important to determine how jackets and blocks would impact the database and cataloging workflow** (unless, of course, the collections managers may already have a work-around for this, which I'm just not aware of!).
- As noted above, some of the researchers felt the measurements were important to collect. However, in general they also seemed happy to defer to the judgement of the La Brea collection staff. **If the La Brea staff change how they collect the measurements, it may be worthwhile to make the reasons and justifications for this change extremely clear, perhaps through a publication or at least a press release on the website. This could help ease researchers' used to the old style of excavation through the transition.**
- Another way to collect positional data may be to transition to a photography-based system of documenting the deposits' structure. These may remove the need for measurements. However, without a specific set up to critique, it's difficult to advise as to data collection protocols. I would caution that if photography was being used to document the taphonomy of a deposit, there would likely be a need to develop a workflow tying a specific fossil to a specific point on an image. This may be challenging,

and could require an image processing workflow that gives excavators "edit" access to the pictures in the field (in other words, the excavation staff would possibly need a tablet where they could view and annotate pictures on the fly).

RECOMMENDATIONS FOR FUTURE WORK WITH PIT 91 AND PIT 91 DATA

EXCAVATION

Researchers that had experience either working in Pit 91, or working with Pit 91 data, generally agreed that it would be best to continue collecting Pit 91 data in the same way (e.g. according to the 1970 protocol). I agree that this would likely be best for later data curation and cleaning work; it will be easier to store, clean, reformat and use the entire Pit 91 dataset if it's all collected in the same manner. However, there are likely changes that could be made the curation and cataloging processes used with the Pit 91 material – particularly the microfossil associations.

SPECIMEN CURATION

The Pit 91 collection is the least used fossil collection at La Brea, primarily because it is physically difficult for collections staff and scholars to find what they're looking for. Where the Hancock Collection is physically organized by taxon, element and then location, the Pit 91 collection is physically organized according to catalog number and microfossil association. This has made the Pit 91 collection difficult to browse through physically; it is essentially designed to be accessible by computers and robots – not humans. Thus, the Pit 91 fossils may need to be cataloged, curated and stored in a different way. Otherwise, it will remain difficult to access, and therefore underutilized.

Additionally - both researchers and La Brea staff seem to agree that the associations between microfossils and the individually excavated bones (the "primary fossils") are largely arbitrary. I did not get the impression that anyone was actually using this association data in their research; at most, researchers working with microfossils would want to know what grid they came from. Additionally, the collections staff repeatedly noted that managing these associations, both physically and digitally, is extremely time-consuming and burdensome. Consequently, La Brea staff may wish to stop cataloging microfossils as associated with "primary fossils" as they have been in the past. However, this would require a lot of work both physically rearranging the collection, and reformatting the Pit 91 records in KE-EMu. This re-curation work may be to0

Near-term: Database the existing Pit 91 collections to make them more accessible
Fundamentally, the existing Pit 91 collections could be made more accessible by:

- Digitizing all available Pit 91 specimen records (including field notes) and including them in the database;
- making this database publically accessible, so researchers could browse the collection and flag specimens for research use;
- and, if possible, photographing at least the "primary" Pit 91 specimens and making those photographs publically available in the database.

Making digitized specimen records and field notes available online would make it much easier for visiting researchers to identify the specimens they need for their work.

Longer term: Rearranging the Pit 91 collections, and refactoring the Pit 91 database
Even with the improvement of the Pit 91 database, there will still be challenges in physically managing and curating the Pit 91 collection. Specimens will still take longer to pull and put away for researchers. Though there are certainly robotic shelving technologies available (sometimes called "automatic retrieval systems"), these are quite expensive to build and implement, and are typically designed for use with books (which are homogenously shaped and fairly sturdy), not fossils (which are small, delicate and differently shaped). Implementing this technology would require the new storage facilities, additional curatorial staff, and likely a huge amount of funding to tailor the technology to La Brea.

More feasible would be a re-shelving project, ideally along with a databasing effort. The Pit 91 collection could be integrated into the taxa- and element-based Hancock collection, hopefully without needing to re-catalog the materials.

Changing microfossil storage: eliminating "associations" between microfossils and primary fossils
"Microfossils" could also be sorted into taxa- and element-based storage as well. In the long run, this could make them easier to access and manage. However, this would likely mean effectively severing the association between microfossils and primary fossils. There are also some potential issues curation and storage issues to address before making this shift:

- *Storing cataloged fragments:* Some (perhaps many) of the cataloged microfossils are fragments. Curatorial staff have mentioned that these are likely not useful to many researchers, and have also noted that these likely would perhaps not merit a catalog number in the future. Would these fragments be worth integrating into new taxa- and element-based systems? If not, where would they go? Simply binning them into a can

could present problems in the event that someone did want to find a specific cataloged fragment in the future. However, storing them individually will take much more space, particularly if they need to be paired with human-readable catalog/locality cards.

- *Labeling microfossils.* Even if only storing complete microfossils, storing them along with human-readable catalog/locality cards will increase the amount of space they need. Small machine-readable codes could reduce the amount of space needed, but would take time to set up and new software and hardware to manage (e.g. QR code readers, etc).

Database implications of eliminating associations: preserving or splitting parent-child relationships?

Changing how microfossils and fragments are stored might mean eliminating associations between the fossils, or at the very least, changing how that information is stored.

The collections managers reported that the associations between the primary and associated fossils have motivated some complicated workarounds in the KE-EMu database. Maintaining these relationships in the database seems like it's causing a considerable amount of extra work for the collections managers. Given that the associations are not scientifically meaningful, it's worth exploring whether it would be possible to "flatten" the parent-child relationships between the bones. However, this would require careful collaboration between the collections staff, the database administrator and possibly an information scientist to understand how "flattening" these relationships might impact the database, the collection, and the cataloging system overall. **If undertaken, this transformation would require an iterative data processing workflow that includes frequent data quality and integrity checks.**

IN-HOUSE DATA CURATION PROJECTS

Thinking of data processing and curation as a necessary part of curation in general may help relieve some of the frustration at other data workflow issues. Data processing can be frustrating, especially with a collection as large as La Brea's – but it will be inevitable no matter what kind of measurement or data collection method is used. Data collected by a number of different people over long periods of time will always need some processing before it's usable. Rather than trying to eliminate this processing, it may be helpful to instead try to plan for it.

With that in mind, there are several initial data curation projects that the museum may wish to take on. Two are outlined below.

IN-HOUSE RE-FORMATTING OF MEASUREMENT DATA

Earlier, I noted that much of the existing measurement data is difficult to map because of its unique format and mapping structure. For instance, researchers have consistently had to re-plot the data from the Southwest corner (rather than the Southeast); they have also had to process the data considerably to plot the grids into one large map, and to accommodate "extensions" – fossil measurements that stretched from one grid to another.

Given how consistent these data curation needs have been, the museum could consider collaborating with information or computer scientists to clean and reformat this data in-house, and making the cleaned data available for research use. This would entail

- Replotting the measurements from the southwest corner
- Integrating the gridded data into one large graph
- And potentially normalizing taxonomic names and elements, if needed.

Some of this work could likely be automated through the use of scripting languages or programs such as Python or R. The automated cleaning could be done periodically, and then added to the shareable dataset. For what it's worth, I would be interested in collaborating on a grant towards this.

RECOVERING FOSSIL PREPARATION HISTORY

Several researchers and La Brea staff members said they felt that better preparation notes would be helpful in their work. It sounds like the collections and lab staff are already working on changing this for current projects. However, they may also need to reconstruct past prep notes before the institutional memory fades. It might be worthwhile to work with the database administrator to see if already curated specimens could be binned and bulk edited with the prior prep methods.