# Modeling Digital Humanities Collections as Research Objects

Katrina Fenlon
kfenlon@umd.edu
College of Information Studies, University of Maryland, College Park
College Park, Maryland

## ABSTRACT

Advancing digital libraries to increase the sustainability and usefulness of digital scholarship depends on identifying and developing data models capable of representing increasingly complex scholarly products. This paper considers the potential for an emergent model of scientific communication, the *research objects* data model, to accommodate the complexities of digital humanities collections. Digital humanities collections aggregate and enrich diverse sources of evidence and context, serving simultaneously as "publications" and dynamic, interactive platforms for research. The research objects model is an alternative to traditional formats of publication, facilitating aggregation and description of all of the inputs and outputs of a research process, ranging from datasets to papers to executable code. This model increasingly underpins research infrastructures in some scientific domains, yet its efficacy for representing humanities scholarship, and for undergirding humanities cyberinfrastructure, remains largely untested. This study offers a qualitative content analysis of digital humanities collections relying on a content/context analytical framework for characterizing collection components and their interrelationships. This study then maps those components and relationships into a research objects model to identify the model's strengths and limitations for representing diverse digital humanities scholarship.

## CCS CONCEPTS

• **Information systems** → *Data structures.*

## KEYWORDS

data models, digital humanities, digital libraries, research objects

**ACM Reference Format:**
Katrina Fenlon. 2019. Modeling Digital Humanities Collections as Research Objects. In *Proceedings of ACM Conference (JCDL '19).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Across disciplines, the growth and evolution of digital scholarship has overwhelmed traditional systems for the representation and communication of research. Digital scholarship in the humanities produces resources that range widely beyond our traditional concept of publication, resources that incorporate not only narratives and rich media, but also datasets and linked data, interactive and functional components, and objects and processes that are physically and logically dispersed as well as dynamic and evolving over time. Despite the rise of digital scholarship, most existing research infrastructures lack support for the creation, management, sharing, maintenance, and preservation of complex, networked digital objects.

This paper considers the potential for emergent models of scientific communication and publication to accommodate the complexities of digital humanities scholarship, and therefore to underpin shared research infrastructure in the humanities. In particular, this study analyzes the suitability of the *research objects* model,[1] one among several emergent models for representing and describing complex digital objects that interweave data, workflows, and supplementary and contextual information, models for logically bundling the diverse inputs and outputs of research processes [3, 4]. Research objects comprise metadata frameworks with associated packaging standards. The model has gained uptake in some disciplines and witnessed concomitant growth in related tools, management systems, and supportive communities [2, 11, 26], which indicate its usefulness and contribute to its sustainability.

This study offers a starting point for answering the question: To what extent may existing (scientific) data models for representing research objects accommodate DH research products and processes? This paper focuses on a common form in DH scholarship: digital collections (often called *digital archives* and *thematic research collections*), which are scholar-built aggregations of digital sources of evidence about a topic [12, 15, 27]. This study provides selected results of a qualitative content analysis of DH collections, and offers a content/context analytical framework to characterize collection components and their interrelationships. This study then retrospectively maps those components and their relationships into the research objects model in order to identify the strengths and limitations of that model for representing DH scholarship.

### 1.1 Digital scholarship and sustainability

In the past few decades, research and scholarship have witnessed sweeping efforts to rethink existing formats for knowledge transfer and scholarly publication, and to develop technologies that support the publication and interlinking of data, software, workflows, and narratives, all as first-class research objects [8]. In the humanities, scholarship takes an increasing variety of forms, ranging from digital scholarly editions (e.g., the *Walt Whitman Archive*[2]) to curated collections of content (e.g., *Colored Conventions*[3]), from layered

[1] http://www.researchobject.org/
[2] https://whitmanarchive.org/
[3] http://coloredconventions.org/


*JCDL '19, June 2019, Urbana-Champaign, IL, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

visualizations (e.g., the *Torn Apart/Separados* project[4]) to models and simulations (e.g., the *MayaArch3D* Project[5]). The outputs of DH research are increasingly media-rich, data-centric, interactive, dynamic, interlinked, and subject to indefinite evolution.

As infrastructures for sustaining digital research struggle to keep pace with the advance of scholarly communication technologies, DH confronts sustainability challenges [19, 22, 23]. Digital libraries—including data repositories, aggregations of cultural records and artifacts, and certain publication platforms—are important components of research infrastructure in the humanities. While the capacity of digital libraries for representing complex digital objects and workflows continues to advance [6, 20, 29], there remains an urgent need for data models and standards to represent and describe increasingly complex scholarly products [13, 19].

Digital humanities (DH) collections, including those analyzed in this paper, often resemble cultural heritage digital libraries, broadly conceived. But DH collections are differentiated in several ways that make sustainability uniquely problematic. DH collections are often developed and maintained outside of the walls and purview of dedicated memory institutions. They tend to be centered in scholarly communities; scholars create them and maintain them for their own purposes, with fluctuating resources and support. Because they function simultaneously as scholarly "publications" and as platforms and hubs for ongoing research and communication among scholarly communities, and because they tend to be funded on short cycles, they often rely on bespoke infrastructures and take unique forms to serve specific research purposes. These factors combine to make DH collections uniquely difficult to sustain over time, and suggest the urgent need for shared infrastructure that does not limit the diversity of digital scholarship.

### 1.2 Research objects in the humanities

The basic concept of the *research object* is simple. Conceptually, research objects are composed of two main parts: aggregated resources (listed in a manifest with minimal metadata, and packaged into the research object using one of several packaging formats), and annotations (used to express metadata about, provenance of, and relationships among aggregated and external resources). The standard model specifies how relationships are declared, relying on extant linked data standards, primarily on OAI-ORE,[6] and W3C standards including the Annotation Data Model[7] and Prov[8]. The research object may be packaged and serialized in different ways, but always contains a manifest of metadata about the research object and its contents represented in JSON-LD. There are other models closely related to the research objects model, including for enhanced publications [2], executable papers, and scientific publication packages. Research objects have seen growing application in several domains, in various commercial and open-source implementations [5, 7, 10, 11, 18].

In the humanities, research objects and closely related models have been applied to repository and data-sharing architectures [1, 6], digital preservation and archive serialization [21, 30], semantic publishing [24], and digital libraries for musicology [25]. These applications are compelling, and suggest the need for and timeliness of a systematic investigation of whether or to what extent the model could serve to represent a range of DH collections as whole, cohesive objects, and therefore have potential to underpin a widely adoptable, sustainable DH infrastructure with cross-disciplinary investment and impact. Data modeling is a pervasive scholarly practice in DH [16]. Like research objects, DH collections may be conceptualized and modeled as assemblages of resources with semantic interconnections, designed to support research objectives [4, 13]. This study considers to what extent that resemblance bears out in the application of the *research objects* data model to complete representation of collections.

## 2 METHODS

The analysis presented in this paper builds upon an ongoing, multimodal study of digital collections [12, 13]. The study seeks to thoroughly characterize DH collections as a scholarly genre using three approaches: (1) a survey and typological analysis of DH collections (n=150 to date); (2) a qualitative content analysis of exemplary collections; and (3) interviews with researchers and practitioners who build digital collections, to identify challenges for libraries and other institutions in supporting and sustaining DH scholarship. The typological analysis identified three primary types, useful for describing DH collections in terms of their purposes and the completeness toward which they are developed; those types are briefly described in Table 1. Complete results of the first phase of the study and a detailed account of the interrelated methods are given in [13].

### 2.1 Qualitative content analysis

The current paper extends the qualitative content analysis to address the question: What *components* of these collections must be modeled in order to logically represent DH collections as research objects? In other words, what are the main products of the collection—its discrete, publishable outcomes—and how are they related to one another and to other resources? The initial phase of content analysis identified close to forty distinct aspects of the content, design, and contexts of digital collections. Table 2 gives an overview of the whole content analysis protocol and each aspect of the sample collections that has been subject to analysis and characterization.

The two most immediately relevant aspects of this protocol to the analysis at hand are *items* and *interrelatedness*. These aspects concern (1) what are the items in the collection, and (2) how are they interrelated with one another, with contextual information, with external resources, etc.? A closer analysis of *items* and *interrelatedness* in each of our sample collections identified discrete components of collections along with the relationships, both technical and abstract, that obtain between components. This study uses the terms "item" and "component" loosely, not only to indicate a collection's main conceptual units of gathering (such as books or artifacts), but also other parts of collections that substantially contribute to a collection's intended contribution to the scholarly and cultural records. The analysis focuses on discrete logical pieces that may be understood to have some kind of mereological, membership, or *isGatheredInto* relationship to the collection as a whole

[4] http://xpmethod.plaintext.in/torn-apart/volume/2/index
[5] http://www.mayaarch3d.org/language/en/sample-page/
[6] https://www.openarchives.org/ore/
[7] https://www.w3.org/TR/annotation-model/
[8] https://www.w3.org/TR/prov-overview/

**Table 1: Collection types**

| Type | Purpose |
|---|---|
| Definitive-source | Provide access to high-quality, authoritative, or otherwise definitive primary sources, (re-)assembling and shaping the affordances of the cultural record on the Web |
| Exemplar-context | Interrelate and (re-)contextualize diverse primary sources, building rich context and connection within and around exemplary sources |
| Evidential platform | Aggregate, deconstruct, and remodel sources for new uses, leveraging evidence into more flexible platforms for analysis and interpretation |

**Table 2: Content analysis protocol overview**

| Cluster | Categories of analysis |
|---|---|
| Context | Theme; Purposes; Impact; Creators; Audience; Documentation; Provenance; Related collections; Related projects and publications; Review; Funding; Developmental stage; Host; Rights; Sustainability and preservation plans; Method of collection |
| Content | **Items**; **Interrelatedness**; Diversity; Size; Narrativity; Quality; Language; Completeness; Density; Spatial coverage; Temporal coverage |
| Design | Data models; Navigation; Infrastructural components; Interface design; Interactivity; Interoperability; Openness; Identification and citation; Modes of access and acquisition; Accessibility; Flexibility |

[31], and which contribute to its scholarly purpose according to the collection's self-described objectives.

## 2.2 Content/context component framework

To refine the analysis of collections, this study developed and applied an analytical framework for characterizing components of collections more precisely. This characterization leverages a few different properties of components—including whether they are primary or secondary sources, and whether they are original to a collection—with the goal of identifying different ways in which components contribute to collections as wholes and, in turn, to the wider scholarly record. Figure 1 illustrates the "content/context" analytical framework used to focus the content analysis of collections in anticipation of applying the research objects model.

The framework is intended to refine analysis of how collections are constituted, and how their constitution determines the ways in which they contribute to scholarship. Using this framework, each component is first categorized as either *content* or *context*. "Content" includes components that are discrete, independent sources of evidence for scholarship. "Context" includes components that play a supportive, interpretive, representational, or functional role that is essential or utilitarian for the use and understanding of content. The reason for differentiating these categories conceptually, despite the difficulty of teasing them apart in practice, is to refine our understanding of collection contributions.

The next question put to components identified as *content* is: Are they primary or secondary sources, or would it be more accurate to say they fall somewhere in between? For both content and context components, a third question is: Is the component original, or has it been previously published or published externally to the collection? The final question is, how are both context and content components interrelated? These questions are intended to challenge our intuitions about aspects of collections that are commonly understood to be peripheral to collections.

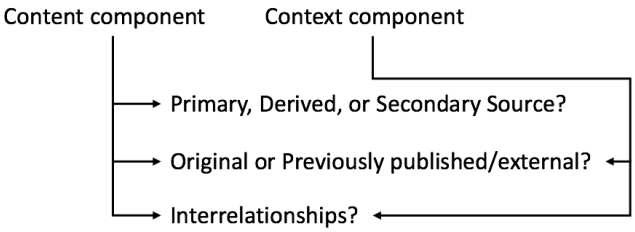


**Figure 1: Content/context analytical framework**

Content components in these collections include primary sources, secondary sources, and derived sources. Primary sources are well understood to be representations of original documents or first-hand evidence, while secondary sources offer substantial interpretation of primary sources. However, some resources seem to fall between these two categories, such as datasets extracted from primary sources. This study considers such sources to be *derived*. Derived sources are generated "directly" from primary sources through some interpretive intervention, where interpretation is manifested in the mode or method of derivation, such as an algorithm or encoding scheme designed to foreground or extract specific pieces of data from the sources. I posit that derived sources are more closely related to primary sources than other secondary sources because they are intended as alternative (usually computational) representations of primary sources.

Content components further divide into categories of original versus previously published/external. "Original" implies that a source is the first (digital) source of its kind, or has no available counterpart. "Previously published" implies that a source or comparable version has been published or digitized elsewhere, or is a reference component that exists externally to a collection.

Contextual components in these collections include elements that are essential or important to the interpretation, use, management, curation, and preservation of collections, but which do not constitute the main content. For example, contextual components include documentation and data models such as markup schemas or ontologies. Finally, many contextual components are functional, dynamic, and interactive features or affordances. Context components may also be original, previously published or external, or somewhere in between.

## 2.3 Collections

The following three collections were selected for close qualitative content analysis: the *Shelley-Godwin Archive*,[9] the *Vault at Pfaff's*,[10] and *O Say Can You See: Early Washington D.C. Law & Family*.[11] These collections were selected to represent three distinctive types of collection, summarized in Table 1 [13], which were identified in prior typological analysis.

The *Shelley-Godwin Archive* (*Shelley-Godwin*) represents a *definitive source* collection, a digital library focused on the representation of definitive primary sources, such as scholarly editions and authoritative archival sources intended for close study by scholars in a domain. *Shelley-Godwin* provides digitized, transcribed manuscripts from the Shelley-Godwin family of 18th- and 19th-century writers, including Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft. The collection aims to be a definitive digital source for close study of the Shelly-Godwin manuscripts—including major literary works such as *Frankenstein* (M. W. Shelley) and *Prometheus Unbound* (P. B. Shelley). Manuscripts are supplemented with biographical, bibliographical, and other secondary sources.

The *Vault at Pfaff's* (*Vault*) represents an *exemplar-context* collection, which aims to present exemplary (rather than definitive) sources on a subject, and to interrelate them with interpretive, contextual materials. *Vault* gathers primary and secondary sources about the historically significant bohemians of antebellum New York, U.S.A., particularly the social network revolving around the historical bar *Pfaff's*, which became an epicenter for a literary movement. The site provides a searchable annotated bibliography of more than 8,000 texts, linking to full-text internal and external sources. Critically, while some of the primary sources are hosted by *Vault*, many are instead references (with some linked to external sites), because the main content of this collection is the records of primary sources and the rich, interwoven contextual information with which records are augmented. The site also provides a map, timelines, biographies, and historical essays. Unlike *Shelley-Godwin*, *Vault* does not aim to provide an original or definitive set of primary sources for close study, but rather a massive set of interrelated sources, social entities, and contextual information to support the discovery of new connections.

*O Say Can You See: Early Washington, D.C., Law and Family* (*O Say*) represents an *evidential platform*, a digital library focused on gathering sources to provide evidence for a specific interpretive or analytical goal [13]. *O Say* gathered, digitized, and analyzed freedom suits filed in Washington, D.C., and surrounding areas between 1800 and 1862, in order to explore family, legal, and social networks. Like *Shelley-Godwin*, *O Say* provides carefully transcribed and encoded primary sources, but with a central goal of deconstructing and remodeling those sources for use as data (e.g., for computational social-network analysis).

# 3 COMPONENTS OF COLLECTIONS

In this section I consider what *components* of our sample collections must be modeled in order to logically represent them as research objects, to lay the groundwork for attempting a retrospective mapping to the research objects data model. For each collection, content analysis and the application of the content/context framework serve to identify the main products of the collection—its discrete, publishable outcomes—and how they are related to one another and to other resources. The remainder of this section characterizes the *items* and *interrelatedness* of the current instantiation, identified through content analysis of each collection.

## 3.1 *Shelley-Godwin Archive* components

*Shelley-Godwin* aims to provide a definitive collection of manuscripts, digitized as high-quality page images with corresponding TEI-encoded transcriptions. These manuscripts are augmented by innovative modes of access and participation for users, including features for multimodal and comparative reading, and features for facilitating future participation in the archive through user annotation and curation of manuscripts. What are the original contributions and important contextual components, and how are they related? Content analysis of the collection identified the following components:

- **Manuscripts**: Manuscripts are abstract objects, with multiple possible orderings, of sequential transcriptions and corresponding page images, currently instantiated through TEI-XML files that reference and order the separate TEI-XML files representing transcribed pages (see below).
  - **Page images**: Digitized manuscript page images. The image files are hosted remotely and appear on the site through a call to the Bodleian digital library's IIIF API; but images were digitized under the auspices of the Shelley-Godwin Archive project and thus constitute a contribution of the project.
  - **Encoded transcriptions**: Transcriptions of page images, encoded in a TEI-XML schema for representation of primary sources. Multiple representations of the page images and transcribed text stem from Shared Canvas manifests that are generated based on these TEI files; these transcriptions are the foundation of this project's contribution.
- **Narrative components**:
  - **Original texts**: The project offers manuscript descriptions, currently instantiated as HTML files.
  - **Excerpted texts**: The project includes excerpts of previously published texts, including manuscript descriptions and a chronology, currently instantiated as HTML files.
- **Browse and search functionalities**: Browse and search of *Shelley-Godwin* operate across manuscripts as wholes, and across components of manuscripts. These functionalities are customized to offer multiple reading orders, taking advantage of the highly rich encodings.
- **Reading viewer**: The custom implementation of the reading viewer takes advantage of Shared Canvas/IIIF representations of the manuscript images in addition to the encodings, to allow readers to compare the original handwritten text with its transcriptions, and to limit views by authorial hands.
- **Schemata and utilities**: *Shelley-Godwin* relies on multiple custom data models and utilities for constituting the manuscripts from numerous components.

[9] http://shelleygodwinarchive.org/
[10] https://pfaffs.web.lehigh.edu/
[11] http://earlywashingtondc.org/

**Table 3: Collection objectives**

| Shelley-Godwin | Vault | O Say |
|---|---|---|
| Provide access to a complete set of encoded manuscripts | Aggregate access to distributed, related sources | Digitize, transcribe, encode archival documents to extract data for analysis |
| Facilitate multi-modal, comparative reading and user participation | Illuminate a network of works (sources), people, places | Reconstruct and expose hidden relationships and personal histories |

The components of *Vault* and *O Say* are described in less detail, below, to facilitate comparison with *Shelley-Godwin*.

### 3.2 *Vault at Pfaff's* components

*Vault*, which aims to help users discover connections among a large set of related sources and people, decomposes into the following main components: *Annotated bibliographic metadata records*, which include annotated internal hyperlinks to related *people* entities (whether authors or mentions) and internal/external hyperlinks to electronic sources when available; *annotated biographical records* (people entities); a dedicated *relationships browser*, along with other browsing and searching facilities; *original narrative components* including historical essays and full biographies; an extended *timeline* and *interactive map*; and *transcriptions* and *page images* of the subset of primary sources hosted by *Vault* (most primary sources in this aggregation are externally linked).

### 3.3 *O Say Can You See* components

*O Say* provides encoded primary sources and extracted data. Its main contributions may be decomposed into the following components: *Page images* of archival documents; *encoded transcriptions* of archival documents in TEI-XML; extracted and augmented *person data* (represented as RDF data documenting relationship and personal information, derived from a central CSV file, all extracted from case documents); *family guides* (family trees that interrelate "people" entities, derived from the same central data source); *cases* (abstract entities, a mechanism for aggregating extracted data and documents, such as person entities and case documents references); *annotated cases* (which are the same as *cases*, but including long annotations with hyperlinks); a *relationships ontology* (OWL) and other customized data models; a special browse and search functionality, including relationship browse and search with multiple serialization options and simple relationship API; *stories* (original long-form narratives heavily linked both to internal entities/resources and external resources); and a *bibliography* with links to related projects, and primary and secondary sources.

### 3.4 Content and context components

Applying the context/context framework to the components identified through content analysis exposes a few important characteristics of DH collections, which any data model intended to represent and describe collections must take into account. As an example of how this analytical framework applies to collections, Figure 2 shows selected content and context components of all three collections mapped to a two-dimensional grid, to demonstrate how components fall along two spectra of (1) Primary/Derived/Secondary sources and (2) Previously published (or external) versus Original sources. The grid differentiates six boxes or categories for the sake of making the framework more legible, but in reality the category boundaries are fuzzy and each axis should be understood as a spectrum. Components of the three collections fall into almost every category. (The only category into which no components fall, in this analysis, is the category of components that are both derived from primary sources and previously/externally published; but it is easy to imagine components that would fall into such a category, such as datasets hosted in an external repository.)

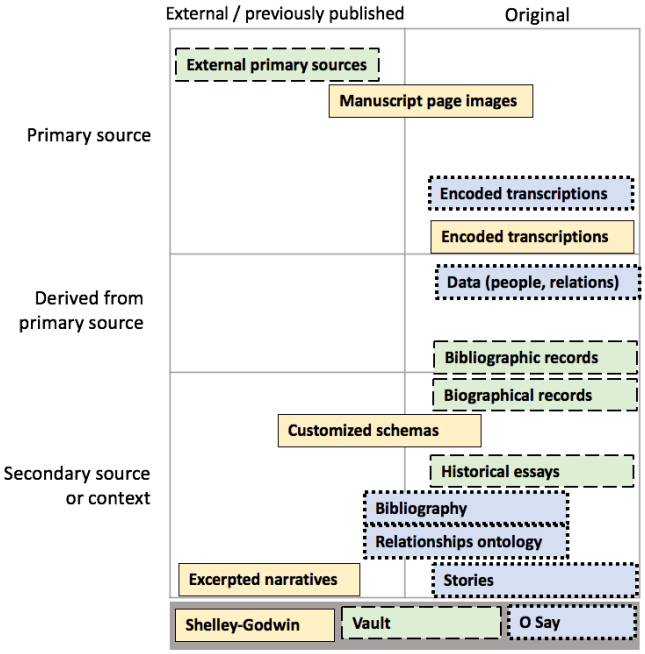


**Figure 2: "Content" components mapped to framework**

Mapping components identified above to this framework, as in Figure 2, exposes the following essential and interesting characteristics of DH collections:

**Components contribute to scholarship in diverse ways.** The mapping illustrates the great variety among the components of even just a few collections—variety not only in type and form, but also in less predictable dimensions, including their originality and how they participate in the scholarly record, whether as primary, secondary, or derived sources. The contributions of a collection are often framed in terms of concrete, novel *additions* to the scholarly and cultural records, but such additions are more various, and sometimes more abstract, than usually imagined. The multidimensional

diversity of the components that constitute our collections may complicate our judgments about which pieces are priorities for sustainability and preservation.

**Not all essential content is original or internal to the collection.** For example, many of the primary sources that make *Vault* a valuable resource for discovery were previously published and constitute external references. In a different case, the manuscript page images that constitute a major part of *Shelley-Godwin*'s contribution to scholarship are original but externally referenced, which will pose fundamentally different challenges to the sustainability and preservation of the collection as a whole than if they were co-located with the rest of the collection components.

**Content is not the only essential contribution of a digital collection.** The contribution may be partly or even centrally manifested in the *interrelationships* among components, or in the *context* surrounding the content. These relationships and context have been called the "connective tissue" of a collection [13]. For example, the customized schemas and utilities used to constitute the archive and its contents may represent a technical contribution to DH as a field of practice. The custom relationships browsers of *Vault* and *O Say* serve to enact scholarly interpretations; the ability to search and browse fine-grained relationships within and among components in bespoke ways is essential to the purposes of those collections. Flanders (2014) invites us to "consider what happens to our understanding of a 'collection' when its constituent items are no longer the primary unit of meaning" [15]; at the least, this idea suggests that standard repository models for representing "items + metadata" as constituting a collection are insufficient to represent and describe DH collections. The next section breaks some of the connective tissue down to have a closer look, prior to the application of the research objects model.

## 3.5 Relationships among components

Components of collections are interrelated both conceptually and technically, and these relationships are essential to representing and describing collections as complex and cohesive wholes. In the case of *Shelley-Godwin* relationships are implemented in various ways. The collection leverages identifiers, schemata, utilities (scripts or processes), and data files to construct the archive's representation of each manuscript.

Figure 3 offers a reductive illustration of components and relationships of *Shelley-Godwin* and relationships among them. In Figure 3, items included in the collection are enclosed in (blue) squares. Note that page images appear in a separate square; while they are logically part of *Shelley-Godwin*, they are maintained and hosted by a different institution in a separate digital library (Digital Bodleian[12]) and called via API. In Figure 3, arrows represent relationships. Solid arrows represent referential relationships that are formalized and actionable (if not semantically encoded), such as relationships performed by hyperlinked URIs. These include the following (broadly described):

**(a)** Custom data models refer to (and extend) standard, external data models, for purposes of validation and documentation. For example, the *Shelley-Godwin* TEI-ODD file references the TEI standard, in addition to the standard that defines the ODD.

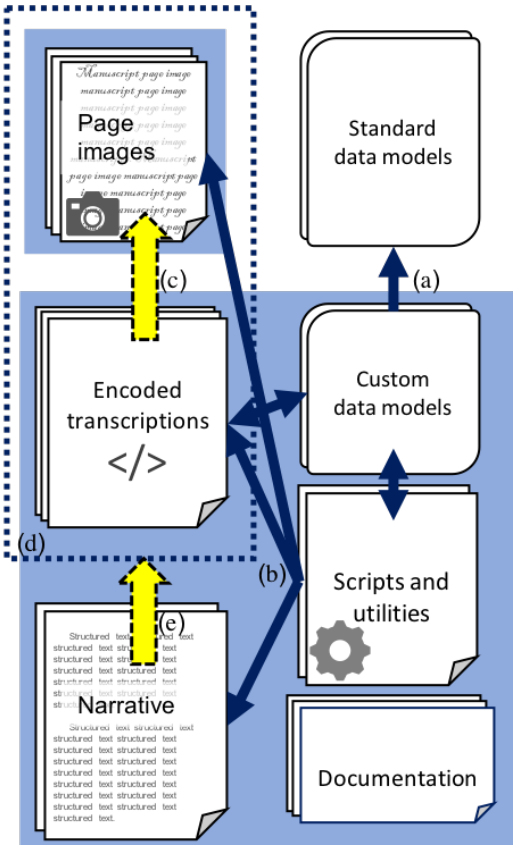


**Figure 3: Conceptual and technical relationships among components**

**(b)** Scripts and utilities refer to all components in order construct or enact the functional website. For example, the site relies on the Unbind utility, a Python utility to create Shared Canvas manifests (which underlie the interactive reading viewer) from *Shelley-Godwin* TEI files.

Dashed (yellow) arrows represent conceptual or abstract relationships, which are implemented indirectly through various means. These are conceptual relationships, made visible to users by the design of the site, but technically performed by completely separate components of the collection. These include the following:

**(c)** Relationships between page images and corresponding encoded transcriptions. For users this relationship is experienced via the juxtaposition of both in the reading viewer. Behind the scenes, this juxtaposition is created by the utilities described above.

**(d)** Relationships between each manuscript and its components. Each manuscript is an abstract entity with a proxy in the form of XML documents, one for each *volume*, which list the URIs for the individual pieces, or pages, that constitute the volume and manuscript. The identifiers for pieces of

[12] https://digital.bodleian.ox.ac.uk/

the manuscript serve to identify both page images and corresponding XML files, because scripts and utilities expand the identifiers into URIs. The dashed circle in Figure 3 encompasses the abstract object of the manuscript, an abstract entity that is evident and interactive to users through browsing mechanisms and the comparative reading viewer, but which is constructed behind the scenes through a complex, distributed process.

**(e)** Relationships between narrative components and manuscripts. References to manuscripts within the narrative components of the site are implemented as hyperlinks between textual references the landing pages for corresponding manuscripts.

Through this analysis of the final aspect of our "content/context" framework—the aspect of relationships among components—we find another crucial observation about DH collections: **not all relationships among components are equal.** Some are implemented directly using mechanisms such as URI addresses, which would readily translate to alternate representations, such as semantic relationships in a linked data or research objects model. Other relationships are implemented indirectly via processes that may prove more difficult to translate or migrate. Dwelling on relationships within *Vault* and *O Say* is out of scope for this paper, but those collections, even more than *Shelley-Godwin*, realize their purposes and contributions through their connective tissue, and demand a deeper analysis in future work.

## 4 RESEARCH OBJECTS AND COLLECTIONS

So far this analysis has broken collections down into sets of logical components and relationships, with the goal of applying the research objects model to describing and representing them. By way of reminder, research objects are comprised of two main kinds of things: *aggregated items* and *annotations*. In this model, a research object may be serialized as a *bundle*, which is a zip archive of a file structure and all constituent data files, along with a JSON-LD manifest of metadata about the aggregation contents.

How well can this model capture the logic and meaning of digital collections? This section suggests a basic mapping of components and relationships of one collection, *Shelley-Godwin*, to the research objects model, in order to begin to identify challenges and implications of this model for representing DH scholarship. The following examples assume the goal of trying to migrate the Shelley-Godwin—the complete collection, as data—into a research object bundle. The collection could then be migrated into a research objects management system, so that other digital humanists could access and use the data alongside (presumably) many other collections, or so that third-party applications could draw on the data to support custom interactions. The details of access and use are not imagined here, but some potential implications for varieties of access and use are considered in section 5.

First, adopting the model means capturing components that fall into the *content* category of the content/context framework articulated above. For *Shelley-Godwin*, these components are (at least): (1) page images, (2) encoded transcriptions corresponding to page images, and (3) narrative components that serve to describe manuscripts. Manuscripts, in turn, are abstract entities that are manifested by relationships among page images and encoded transcriptions. In a research object, each component would be referenced in the manifest as an aggregated item. The following example record shows a portion of a research object manifest, which lists aggregated items including (1) an XML file (ending in "volume_i.xml") representing Volume 1 of Mary Shelley's *Frankenstein* manuscript, and which references the individual pages in order; (2) a single digital page image (in JPEG2000 format); (3) an XML file (ending in "c56-0001.xml") representing a single page of the *Frankenstein* manuscript; (4) an HTML file representing a narrative introduction to the manuscript; and (5) the TEI-ODD schema that governs the *Shelley-Godwin* implementation of TEI-XML.

```
"aggregates": [
    {
        "uri": ".../tei/ox/ox-frankenstein-volume_i.xml",
        "mediatype": "application/xml",
        "conformsTo": ".../odd/shelley-godwin-FULL.odd"
    },
    {
        "uri": "https://digital.bodleian.ox.ac.uk/inquire/p/cb631387-c307-
                454a-9aad-53b7e1541d2f",
        "mediatype": "image/jp2"
    },
    {
        "uri": ".../tei/ox/ox-ms_abinger_c56/ox-ms_abinger_c56-0001.xml",
        "mediatype": "application/xml",
        "conformsTo": ".../odd/shelley-godwin-FULL.odd"
    },
    {
        "uri": ".../site/_pages/frankenstein/ the-
                frankenstein-notebooks-introduction.html",
        "mediatype": "text/html"
    },
    {
        "uri": ".../odd/shelley-godwin-FULL.odd",
        "mediatype": "application/xml",
        "conformsTo": "http://www.teic.org/
                release/xml/tei/custom/schema/relaxng/tei_odds.rng"
    }
    …
],
```

**Figure 4: Snippet of partial manifest for *Shelley-Godwin* research object aggregation**

Note that the *aggregates* field already captures several important relationships among the components of *Shelley-Godwin*, even prior to the addition of explicit relationship annotations. First, the research object manifest represents and make explicit the relationships between "tangible" or self-contained components (such as files or documents) and abstract components of the collection. In this example, the volume-level XML file stands as a proxy for a manuscript, which, as discussed above, is an abstract object in Shelley-Godwin's architecture. It would also be possible to represent the manuscript as an abstract entity more explicitly in this model, perhaps relying on the OAI-ORE *proxy* mechanism.

In addition, URIs for aggregated objects may reference both local files contained within a research object and remote resources. In Figure 4, relationships to external resources are highlighted. The *conformsTo* field allows a research object creator to indicate schemas or standards to which a given aggregated resource conforms; in this case *conformsTo* references schemas both internal and external to the collection. Relationships between the encoded transcriptions and relevant schemas and standards, embedded in the TEI-XMl file headers, can also be described in the research object manifest, where they can be exposed to consumption by independent applications. Figure 4 gives an example of how a *Shelley-Godwin* research object might reference page images hosted externally to the collection, in Digital Bodleian. Digital Bodleian is in fact the

source of page images displayed in the *Shelley-Godwin* website. But in the current *Shelley-Godwin* site, this referential relationship is only made explicit within the code used to generate pages. The research objects model makes this relationship explicit, semantic, and discoverable in the outward-facing manifest.

*Annotations*, constituting the second major piece of a research object manifest, are used to express descriptive metadata about aggregated resources, including relationships among resources (internal or external) and detailed provenance information. Annotations rely on domain-specific ontologies and vocabularies. Figure 5 exemplifies annotations that make explicit several relationships among aggregated components of *Shelley-Godwin*, including *Shelley-Godwin* relationships (c), (d), and (e), identified in section 3.5, above:

- **(c)** Relationships between page images and corresponding encoded transcriptions: In this example research object, these relationships are made explicit in annotations that link each XML file representing a single transcribed page to its corresponding page image, via *prov:wasDerivedFrom*. There are, of course, other ways to express this relationship.
- **(d)** Relationships between the various components that constitute a manuscript: In this example research object, the relationships are made explicit in annotations that link each XML file representing a single transcribed page to its corresponding TEI-XML file representing a single volume, via *dct:hasPart*. There are other ways this relationship could be represented.
- **(e)** Relationships between narrative components and *manuscripts*. Hyperlinks forge relationships between textual references and manuscripts; therefore these relationships are best modeled not at the document level but at a lower level within the text. These relationships could simply remain as embedded hyperlinks, relying on unique identifiers for manuscripts (assuming the URLs continue to function in the new context of a research object). Alternatively, the fact that a narrative component refers to a manuscript could be made explicit in the manifest, via an annotation such as *crm:refersTo*.[13] But it is not immediately clear how a document-level annotation indicating references would be useful.

Figure 6 offers an alternative view of these relationships, expressed as an RDF snippet derived from an ROHub research object[14] and visualized.[15]

The research objects model supports the use of domain ontologies (such as CIDOC-CRM and bibliographic ontologies) for rich descriptions of the interrelationships among collection components and external sources. There are numerous alternative ontological approaches to modeling the relationships given in the examples above. Current research object management systems (such as ROHub) offer a limited set of terms for adding annotations to objects, mainly oriented toward description of computational and scientific research workflows. For example, ROHub's "RO Basic Requirements" require research objects to include hypotheses or research questions, along with conclusions. For expressing relationships among the aggregated *research object* resources, ROHub relies on terms from the Prov and Wf4Ever Research Object ontologies, which are both focused on scientific workflows. Such ontologies will prove inadequate to fully describe the processes or workflows of digital scholarship in the humanities.

```
"annotations": [
        {
            "uri": "...",
            "about": ".../tei/ox/ox-frankenstein-volume_i.xml",
            "dcterms:hasPart": ".../tei/ox/ox-ms_abinger_c56/
                                    ox-ms_abinger_c56-0001.xml"
        },
        {
            "uri": "...",
            "about": ".../tei/ox/ox-frankenstein-volume_i.xml",
            "dcterms:hasPart": ".../tei/ox/ox-ms_abinger_c56
                                    /ox-ms_abinger_c56-0002.xml"
        },
        {
            "uri": "...",
            "about": ".../tei/ox/ox-ms_abinger_c56/ox-
                                    ms_abinger_c56-0002.xml",
            "prov:wasDerivedFrom": "https://digital.bodleian.ox.ac.uk/
                inquire/p/cb631387-c307-454a-9aad-53b7e1541d2f"
        },
        {   "uri": "...",
            "about": ".../site/_pages/frankenstein/the-
                            frankenstein-notebooks-introduction.html",
            "crm:refersTo":".../tei/ox/ox-frankenstein-
                                                    volume_i.xml"}

            ...
    ],
```

**Figure 5: Snippet of partial manifest for *Shelley-Godwin* research object annotations**

This example application of the research objects model has not accounted for the components of collections that are interactive, dynamic, and functional, such as *Shelley-Godwin*'s custom search and browse options, and its comparative reading viewer. These are essential aspects of the project's contributions to scholarship. Not only do they represent technological contributions to the DH landscape, but they were built for symbiosis with *Shelley-Godwin* data, which was modeled to support the use of these advanced tools. As flat code, of course, these pieces readily fit into the research objects model, which has been shown to be useful for aggregating data and code for migration and preservation purposes. But as performative, interactive components that function to enable new kinds of exploration and encounter with collection contents, these components challenge the research objects model. While the model has been applied to software preservation [5], and while workflow-oriented research objects usefully represent certain kinds of dynamic and executable content, the functional and interactive components of DH collections are really about enabling specific, purposeful kinds of real-time, end-user interaction. The duties of the functional, contextual components of collections—to enable exploration, discovery, connection-making, learning, etc.—would be assumed not by a data model but by the interactive components of a research objects management system or other applications built on top of a research objects management system. The potential for such systems and applications to enact the diverse methodological and functional goals of DH scholarship is a topic for future investigation.

## 5 DISCUSSION AND FUTURE WORK

This study has analyzed three DH collections using qualitative content analysis, employing a novel content/context analytical

[13] http://www.cidoc-crm.org/cidoc-crm/
[14] http://www.rohub.org/
[15] http://ontodia.org/

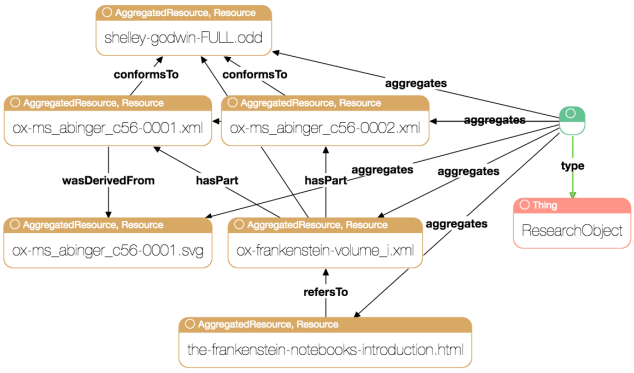


**Figure 6: Partial visualization of RO**

framework in order to characterize collection components and their interrelationships. Applying the framework highlighted a few important characteristics of DH collections that complicate our understanding of how collections are constituted, and which therefore carry implications for the data models that represent collections along with approaches to sustaining and preserving them. These characteristics are: (1) Components of collections contribute to scholarship in diverse ways. (2) Not all of the essential content of a collection is necessarily original or internal to the collection. (3) Contextual components and interrelationships among components may be equally as essential as the main content of a collection.

Research objects have the potential to represent and describe a wide range of scholarly products—more fully and more sustainably than models that currently dominate content management and publication systems. In this paper, the components and interrelationships of the sample DH collections were retrospectively mapped into a research objects model in order to identify strengths and limitations of that model for representing DH scholarship. The following three central strengths emerged.

(1) Research objects readily perform the most essential function of a collection: to aggregate related resources in order to support scholarly objectives. (For this reason, research objects have been leveraged to support digital preservation and big-data transfer [9]).

(2) Research objects have the capacity to accommodate rich semantic descriptions of interrelationships among components, using domain ontologies. These interrelationships may obtain between components with identifiable and addressable representations, such as documents or files, and components that are more abstract. In DH collections, such interrelationships are often inexplicit or "hidden", enacted by or encoded in the layers of scripts and processes that operate to assemble collections for presentation on the Web. When these relationships are hidden, they may be more vulnerable to dissolution in the course of data manipulation, preservation, and migration processes. Formalizing these relationships not only makes them more sustainable; it also opens them to linked data representation and computational uses.

(3) The research objects model accommodates aggregations of linked data, offering researchers the opportunity to create and annotate virtual, fully referential collections in any context and at scale. In addition, structured descriptions of aggregations in research objects are amenable to third-party annotation, and can be leveraged by external applications. These advantages of the research objects model for representing DH collections suggest new possibilities for collaboration, communication, and data reuse within scholarly communities.

The most immediate limitation of the model for DH collections is that functional components designed for end-user interaction are not usefully captured in a basic research objects model. Instead, these components raise questions about the capacities of research objects management systems to serve the distributed development of a diversity of applications. How can management systems serve to underpin experimental, interactive, and dynamic platforms? Different kinds of DH scholarship aim to facilitate different kinds of interactions between users, evidence, and context; the diversity of DH scholarship and the compulsion toward experimentation and innovation have hindered large, sustainable, cross-disciplinary infrastructures.

Realizing the advantages of research objects and related efforts for DH will depend on implementations that establish dynamic platforms for experimentation, participation, and co-creation. This study has treated collections in terms of their logical components and relationships, setting aside for now several other important characteristics and properties, such as collections' look and feel, their digital materiality, and the detailed contours of their implementations. These aspects are essential to the experience and preservation of some collections; it is hard to see how the research objects model could benefit such projects after their development, in retrospective sustainability or preservation efforts, but it is clear that the model could underpin systems going forward that support a wide variety of implementations.

DH research objects would necessarily represent extensions of the basic research objects model, based on the representational and user requirements in different domains and scholarly communities. The work of ontologizing the humanities is well underway. A research objects profile[16] specific to representing collections such as *Shelley-Godwin*, *Vault*, and *O Say* will depend on cobbling together ontologies and vocabularies to express a diversity of relationships among primary, derived, and secondary sources, in addition to workflows, people, and contextual entities. Prior research has emphasized the necessity of highly granular systems of identification, addressability, and reference for supporting DH research and collection practices within digital libraries [14]. Indeed, implementing the research objects model at scale within a linked data paradigm would demand more pervasive use of persistent identifiers for DH objects at varying levels of granularity, including ideally addressable identifiers for each component of a collection, the pieces that make up a component, and so on.

In terms of architecture, DH collections bear significant resemblance to other kinds of digital libraries. The benefits, constraints, and practical challenges of applying the research objects model for DH collections seem, for the most part, likely to hold for cultural heritage digital libraries generally. Emerging linked data collections of cultural heritage institutions stand to support the rise of research objects and similar publication models across disciplines. Future work will investigate the potential intersections between research

[16] http://www.researchobject.org/scopes/

objects and linked data representations of cultural collections in libraries, archives, and museums.

There are numerous emergent models for representing digital publications and digital objects, including models for publishing media-rich and interactive digital monographs along with supplementary materials, and experiments with alternative scientific publishing models such as nanopublications [17]. Future work will investigate the intersections between the research objects model and various alternatives for representing the breadth of DH scholarship, collections, and data, including forerunning applications of research objects to humanities collections [1, 6, 21, 24, 25, 30], and ongoing studies of other approaches to containerization in DH.[17] The research objects data model evaluated in this paper is "data-centric"; workflow-oriented research objects, as a closely related alternative, extend the basic model to capture holistic, executable research workflows. While workflows have received growing attention in the humanities from both technical and strategic perspectives [20, 28], the implications of workflow-oriented data models for capturing the idiosyncracies of humanities research processes need further investigation. Future work will extend this analysis to a more complete study of DH scholarship, scholars, and workflows, in order to advance data models that may help us realize the benefits of standard infrastructure while minimally attenuating the irrepressible diversity of digital humanities scholarship.

## REFERENCES


[1] Bridget Almas. 2017. Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities. *Data Science Journal* 16, 0 (2017).

[2] Alessia Bardi and Paolo Manghi. 2014. Enhanced Publications: Data Models and Information Systems. *LIBER Quarterly* 23, 4 (2014).

[3] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (2013).

[4] Sean Bechhofer, David De Roure, Matthew Gable, Carole Goble, and Iain Buchan. 2010. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Proceedings* (2010).

[5] Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gable, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, and Carole Goble. 2015. Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics* 32 (2015), 16–42.

[6] Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems* 29, 2 (2013).

[7] Joshua Borycz and Bonnie Carroll. 2017. Managing Digital Research Objects in an Expanding Science Ecosystem: 2017 Conference Summary. *Data Science Journal* (2017).

[8] Phil E. Bourne, Tim Clark, Robert Dale, Anita de Waard, Ivan Herman, Eduard Hovy, and David Shotton. 2011. *FORCE11 Manifesto: Improving Future Research Communication and e-Scholarship.* Technical Report. FORCE11.

[9] Kyle Chard, Mike D'Arcy, Ben Heavner, Ian Foster, Carl Kesselman, Ravi Madduri, Alexis Rodriguez, Stian Soiland-Reyes, Carole Goble, Kristi Clark, Eric W. Deutsch, Ivo Dinov, Nathan Price, and Arthur Toga. 2016. I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. *2016 IEEE Conference on Big Data* (2016).

[10] Tracey Clarke and Andy Bussey. 2018. Research Information Systems – fit for the future? A report on the situation and plans of the University of Sheffield Library. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* (2018).

[11] Anita de Waard. 2018. FAIR4CURES: A research object authoring tool for the data commons. *CNI Fall Membership Meeting* (2018).

[12] Katrina Fenlon. 2017. Thematic research collections: Libraries and the evolution of alternative digital publishing in the humanities. *Library Trends* 65, 4 (2017), 523–539.

[13] Katrina Fenlon. 2017. *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign, http://hdl.handle.net/2142/99380.

[14] Katrina Fenlon, Megan Senseney, Harriett Green, Sayan Bhattacharyya, Craig Willis, and J. Stephen Downie. 2014. Scholar-built collections: A study of user requirements for research in large-scale digital libraries. In *Proceedings of the American Society for Information Science and Technology.*

[15] Julia Flanders. 2014. *Advancing Digital Humanities.* Palgrave Macmillan UK, Chapter Rethinking Collections.

[16] Julia Flanders and Fotis Jannidis. 2012. *Knowledge Organization and Data Modeling in the Humanities.* Technical Report. Brown University.

[17] Richard Freedman, Raffaele Viglianti, and Adam Crandell. 2017. The collaborative musical text. *Music Reference Services Quarterly* (2017).

[18] Andres Garcia-Silva, José Manuel Gómez-Pérez, Raul Palma, and Ikay ...Altintas. 2018. Enabling FAIR Research in Earth Science through Research Objects. *arXiv:1809.10617 [cs]* (2018).

[19] David Hansen, Liz Milewicz, Paolo Mangiafico, Will Shaw, Mattia Begali, and Veronica McGurrin. 2018. *A framework for library support of expansive digital publishing.* Technical Report. Duke University.

[20] Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Küster, and Malcolm Illingworth. 2013. TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship. *Journal of the Text Encoding Initiative* 5 (2013).

[21] Inna Kouper, Beth Plale, Dharma Akmon, and Margaret Hedstrom. 2014. Practical and conceptual considerations of research object preservation. *Digital Preservation 2014* (2014).

[22] Christine Madsen and Megan Hurst. 2018. Are digital humanities projects sustainable? A proposed service model for a DH infrastructure. In *CNI Membership Meeting.* https://www.slideshare.net/mccarthymadsen/are-digital-humanities-projects-sustainable-a-proposed-service-model-for-a-dh-infrastructure.

[23] Nancy L. Maron and Sarah Pickle. 2014. *Sustaining the digital humanities: Host institutional support beyond the startup phase.* Technical Report. ITHAKA S+R, https://sr.ithaka.org/publications/sustaining-the-digital-humanities/.

[24] Dominic Oldman and Diana Tanase. 2018. Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. *The Semantic Web - IWSC* (2018).

[25] Kevin Page, David Lewis, and David Weigl. 2017. Contextual interpretation of music notation. *Digital Humanities* (2017).

[26] Raúl Palma, Piotr Hołubowicz, Oscar Corcho, José Manuel Gómez-Pérez, and Cezary Mazurek. 2014. *ROHub — A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science.* Springer International Publishing, 77–82.

[27] Carole L. Palmer. 2004. *A Companion to Digital Humanities.* Blackwell Publishing, Chapter Thematic research collections.

[28] Roger C. Schonfeld and Donald J. Waters. 2018. The Turn to Research Workflow and the Strategic Implications for the Academy. *CNI Spring Membership Meeting* (2018).

[29] Sarah J. Sweeney, Julia Flanders, and Abbie Levesque. 2017. Community-Enhanced Repository for Engaged Scholarship: A case study on supporting digital humanities research. *College and Undergraduate Libraries* 24, 2-4 (2017), 322–336.

[30] David Tarrant, Ben O'Steen, Tim Brody, Steve Hitchcock, Neil Jefferies, and Leslie Carr. 2009. Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications. *The Code4Lib Journal* (2009).

[31] Karen M. Wickett, Allen H. Renear, and Jonathan Furner. 2011. Are collections sets?. In *Proceedings of the American Society for Information Science and Technology*, Vol. 48.


[17] http://digits.pub/