Significant Features of Thematic Research Collections

Katrina Fenlon
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Champaign, IL, 61820
kfenlon2@illinois.edu

ABSTRACT
In order to support innovative scholarship, digital library development must be informed by an understanding of new modes of scholarly production. This paper considers an emergent genre of scholarly publication in the humanities: the digital thematic research collection. While thematic research collections have grown in number and academic significance over the past decade, and this is reflected in the literature on scholarly practice, we do not have a definitive grasp on this exciting phenomenon. What significant features distinguish thematic research collections as a mode of scholarly production? This paper describes a pilot study conducted as part of a project investigating the significant features of thematic research collections. A stronger understanding of thematic research collections will contribute to evolving scholarly evaluation processes for tenure and promotion and the advancement of digital libraries to support new modes of research.

Keywords
Digital libraries, collections, scholarly communication.

INTRODUCTION
In response to technological evolutions and concomitant disruptions in systems of publication, scholars are embracing new forms of communication and dissemination. These forms, which range from shared datasets to blogs to experimental digital monographs, are widely unsupported by current publishing models and evaluation systems. Libraries and digital libraries must understand more about new modes of scholarly production in order to continue to realize their institutional missions to support the creation, maintenance, preservation, and dissemination of scholarship. Increased understanding of new scholarly products will also help us reconcile systems of evaluation (for tenure and promotion processes) with ongoing innovations in scholarly production.

This paper considers one emergent genre of scholarly product in the humanities: the digital thematic research collection (Palmer, 2004; Unsworth, 2000). Thematic research collections (TRCs) gather and present digital primary and secondary source materials on a theme, in order to support research. Beyond this baseline, TRCs exhibit great diversity in purpose, presentation, functionality, and scope. For examples of TRCs that range in theme and presentation, see “Crossroads of War: Maryland and the Border in the Civil War”1, “Gothic Past”2, or “Nineteenth Century Disability: Cultures and Contexts”3. Scholars increasingly demand recognition and support for digital collection development (MacKay, 2014; Fenlon et al., 2014). Digital libraries, with vast and growing stores of digitized content, stand in prime position to support researchers making collections, or to leverage the intellectual contributions of resulting collections. While much has been written on TRCs from the perspective of scholarly practice (e.g., Brockman et al., 2001; Palmer, 2004; Palmer et al., 2009), our understanding remains vague and inadequate to fully support and evaluate this rising form of scholarship.

The overarching goal of this project is to answer the research question: What significant features distinguish the TRC as a mode of scholarly production?

This paper reports on our pilot investigation into this question, exploring the potential of multiple qualitative and computational methods for shedding light on features that may or may not generalize across this emerging, highly eclectic genre.

CONTEXTUAL MASS
Unsworth (2000) and Palmer (2004) identify five basic features of TRCs: they are digital, thematically coherent, heterogeneous, structured, and open-ended. TRCs are

---

2 Highlighted in Omeka Showcase: http://gothicpast.com/
3 Highlighted in NINES: http://www.nineteenthcenturydisability.org/
hypothesized to exhibit what Palmer (2004) and Palmer et al. (2009) term “contextual mass”. Contextual mass is a posited development principle for digital collections, libraries, and aggregations. A collection with contextual mass is one in which items have been purposefully selected, organized, and bestowed with sufficient context to support deep, multifaceted inquiry on a theme (Palmer et al., 2009). At least two of these hypothesized features evade precise definition. In particular, we aim to determine whether it is possible to generally define and measure (1) thematic coherence and (2) contextualization of items in a collection.

**METHODS**

The full investigation into our stated research question will rely on content analysis: qualitative content analysis, bolstered with multiple threads of computational text analysis, applied across collection and item descriptions from select thematic research collections. For this pilot study, we tested a few of our proposed methods on one exemplary test collection: “Nineteenth Century Disability: Cultures and Contexts”, a peer-reviewed TRC drawn from the NINES federation. At the time of data collection, “Nineteenth-Century Disability” contained 53 primary source items of varying types, all pertaining to the theme of disability in the long nineteenth century, each augmented with scholarly annotations. Treated as a text corpus, the collection comprises 54 documents (including the collection record) or 70,292 words (10,807 unique words). Item and collection descriptions were gathered from the collection website using a Chrome extension, Web Scraper. We broached thematic coherence and context-provision from multiple directions. First, we conducted a manual qualitative assessment of context-provision in item records. We framed this assessment using Lee’s (2011) framework of contextual information in digital collections. This framework defines nine classes of contextual entity: object, agent, occurrence, purpose, time, place, form of expression, concept or abstraction, and relationship. Of these, object, agent, time, and place commonly appear in descriptive metadata. In general, digital collections are less reliable about occurrences (such as relevant events or processes), purposes (of items or related contextual entities), relevant concepts or abstractions, and relationships (between items and contextual entities) are less reliable in regular descriptive practice. See Lee (2011) for detailed definitions of contextual entities. For a contrastive approach to context provision, we conducted a part-of-speech analysis of the collection (“Evaluating Context Provision” below) using the Natural Language Toolkit 3.0 (Bird et al., 2009). We compared the collection vocabulary to the vocabulary of a baseline corpus: The Brown News Corpus, which is built into NLTK 3.0, and serves as a crude baseline for quick comparison of natural language. Finally, to begin to assess thematic coherence, we conducted a term-frequency analysis (discussed in “Evaluating thematic coherence”) using Voyant Tools (http://voyant-tools.org/) and a minimally customized stoplist.

**EVALUATING CONTEXT-PROVISION**

Inherent in the concept of contextual mass is the collocation of items with contextual information imperative to scholarly use. We hypothesize that TRCs are distinguished from other kinds of collection, and indeed attain status as a scholarly genre, in part by providing layers of “intellectual context” (Beaudoin, 2012) on top of items, beyond basic descriptive information. Intellectual context may include:

- Information about interrelationships both among items and between items and other entities;
- Expert interpretation, which asserts or reveals such contextual aspects as meaning, function, and significance of items.

Is it possible to identify, understand, and assess the nature and extent of contextual information in TRCs?

We suggest that TRCs obtain contextual mass in part by manifesting (where possible), through description or presentation, all of the nine classes of contextual entity in Lee’s (2011) framework of contextual information in digital collections. To pilot qualitative content analysis, we analyzed the text surrounding items in our test collection for adherence to Lee’s framework. Here we offer one example from the test collection. We aim to catalog contextual information, in the item description, by which contextual entity it represents. The results are given in Table 1. Each of the nine contextual entities was represented in the item description, multiple times and in multiple ways. All nine were also explicitly, carefully interrelated in the item description. While this pilot analysis cannot confirm our hypothesis, our example does suggest that the framework may valuably structure our planned qualitative content analysis of TRCs.

In addition, we conducted an experimental part-of-speech analysis of language in the test collection for distinguishing features. We offer one tentative finding relevant to the provision of context in TRCs. In his review of the literature about humanities scholars’ information needs, Lee (2011) shows that proper names are among the most imperative contextual elements for scholarly use of digital items and collections, along with events and places. Our pilot analysis suggests that our test collection provides substantially more than one would expect from normal use of language.

Figure 1 shows the ratios of occurrence of the 10 most common parts of speech to the whole vocabularies of two corpuses: the test collection, and the Brown News Corpus.
While ratios of most parts of speech are approximately equal for the two corpuses, the test collection contains approximately double the occurrences of singular proper nouns (first column), which include proper names, events, and places. This suggests that TRCs are distinguished by language they contain, and consequently that part-of-speech analysis and named entity recognition may offer new kinds of insight into what distinguishes TRCs as a genre.

**EVALUATING THEMATIC COHERENCE**

Thematic coherence is perhaps the most difficult posited property of TRCs to evaluate. Intuitively, a TRC obtains thematic coherence first by self-definition and then by inclusion of relevant items. In the case of our test collection, for example, each item is selected in part for its pertinence to the theme stated by the collection description, namely: “physical and cognitive disability in the long nineteenth century... emphasizing the technologies, institutions, and representations in literature and popular science.”

---

**Table 1. Cursory qualitative analysis of test item again Lee’s Framework (2011) of contextual entities**

<table>
<thead>
<tr>
<th>Contextual entity</th>
<th>Descriptive manifestations from selected item exemplar</th>
</tr>
</thead>
</table>
| **Object**        | (1) Primary source image: “Treatment of consumption... by mesmerism” (Davey, 1842)  
                    (2) Primary source text: “Letters on Mesmerism,” (Martin, 1844)  
                    (3) Linked primary source: “Medical Report on the Case of Miss H—M—,” (Greenhow, 1845)  
                    (4) Reference for further reading: Dickens and Mesmerism; The Hidden Springs of Fiction (Kaplan, 1975)  
                    (5) Reference for further reading: Medical Body and Lived Experience: The Case of Harriet Martineau (Rowan, 2000)  
                    (6) Reference for further reading: Spellbound: Studies on Mesmerism and Literature (Tatar, 1979)  
| **Agent**         | (1) Author of primary source image: William Davey  
                    (2) Author of primary source text: Harriet Martineau  
                    (3) Related person: Franz Anton Mesmer, originator of theory  
                    (4) Related person: Thomas Greenhow, Martineau’s physician and brother-in-law  
                    (5) Related person: Spencer Hall, mesmerist  
                    (6) Contributor to item: Nadine Laceur  
                    (7) Authors of all references for further reading: Kaplan, P., Rowan, A., Tatar, M., and Winter, A. |
| **Occurrence**    | (1) Narrated events and processes: Martineau’s illness, mesmerism, and recovery  
                    (2) Narrated events: Publication of and ensuing public dialog about “Letters on Mesmerism” |
| **Purpose**       | (1) Of this item in this collection: To contribute to the mission of the collection by providing “striking illustrations of the debate between patients and physicians for authority over the disabled or ill body during the Victorian period and of Martineau’s role in this debate”  
                    (2) Of primary source “Letters on Mesmerism”: to confirm the scientific validity of mesmerism and that this would strengthen the movement of patients taking control over their own medical cases and assuming some of the power and authority of the medical profession |
| **Time**          | (1) Victorian period  
                    (2) Creation and publication dates of objects referenced above  
                    (3) Lifespans of agents referenced above |
| **Place**         | (1) (Implied) London |
| **Form of expression** | (1) (Implied) Letter  
                    (2) (Implied) Text  
                    (3) (Implied) Image |
| **Concept or abstraction** | (1) Mesmerism (explicit tag)  
                    (2) Invalid (explicit tag) |
| **Relationship**  | (1) Narrative description of relationships between agents referenced above  
                    (2) Narrative description of relationships between agents referenced above and the provided and linked primary sources  
                    (3) Implied relationships between primary sources themselves  
                    (4) Implied relationships between this item and references for further reading  
                    (5) Sometimes implied, sometimes explicit relationships between all of these contextual entities |

---

**Figure 1. Parts of speech in test collection**
culture that shaped ideas about disability which are still current today." For now, we offer an exploration of one potential method: applying simple term frequency measures across the collection. Using Voyant Tools, we examined (1) most frequent words across the collection and (2) distinctive words for each document, as determined by a tf-idf measure. Figure 2 gives most frequent words across all items in the collection. This visualization suggests that language used in item descriptions exhibits striking coherence with the collection’s stated theme and scope. Distinctive terms for each document corroborate this suggestion of thematic coherence. Figure 3 gives a snippet of results from the Voyant Tools summary of distinctive terms for the first ten items in the collection. Even among those terms that set items apart, linguistically, from the rest of the corpus, we find a high proportion of terms that are intimately related in meaning or referent to the collection’s stated theme.

CONCLUSION
This paper has examined possible approaches to identifying and understanding two significant, yet vaguely defined features of thematic research collections: contextualization and thematic coherence. Qualitative analysis of item records suggests that TRCs surround primary source evidence with multiple layers of every kind of contextual information identified in Lee’s (2011) framework, and that narrative item records serve to interrelate different facets of item context. A part-of-speech analysis tentatively supports this outcome, showing that the vocabulary of this test collection is distinguished from normal language by an abundance of proper nouns, which Lee (2011) recognizes as imperative contextual entities for humanities scholars. In addition, term-frequency-based measures hold potential to shed light on the nature and extent of thematic coherence in TRCs. The methods and tentative outcomes described here serve as a foundation for an ongoing study of what has emerged as an increasingly important, yet understudied and underserved genre of scholarly publication. A fuller understanding of the highly eclectic genre of TRCs, along with the proposed property of contextual mass, will help us adapt systems of scholarly evaluation for new modes of production. A clearer understanding of how collocations of evidence obtain value – or, perhaps, how they are coherent, heterogeneous, and richly contextualized – may help us leverage the enormous potential value in our burgeoning digital libraries of primary sources.

REFERENCES


http://www.nineteenthcenturydisability.org/