Indexing Quality and Effectiveness: An Exploratory Analysis of Electronic Theses and Dissertations Representation

Daniel Alemneh  
University of North Texas  
Denton, TX 76203, USA  
Daniel.Alemneh@unt.edu

Mark Phillips  
University of North Texas  
Denton, TX 76203, USA  
Mark.Phillips@unt.edu

ABSTRACT  
Theses and dissertations (ETDs) represent a wealth of scholarly and artistic content created by graduate students in masters and doctoral programs in the degree-seeking process. Considering the multi-disciplinarity and interdisciplinarity characteristics of ETDs, often several subjects and indexing terms need to be supplied to adequately represent ETDs for efficient access.

This study analyzes the quality and effectiveness of indexing terms -- both authorized terms from controlled vocabularies and free-text keywords -- used to succinctly describe the content of the electronic theses and dissertations (ETDs). Based on the comparisons of the search terms entered by users to discover and access the ETDs, the authors also discuss the subjectivity and objectivity of the process and the need to distinguish functional representation from mere descriptions of a topic.

Keywords  
Guides, instructions, author’s kit, conference publications.

INTRODUCTION  
Since the late 1990s, more and more academic institutions have mandated the electronic submission of theses and dissertations. A move to an all-digital means of providing electronic theses and dissertations (ETDs) is accelerating their discovery and facilitating their use, value, and impact in research.

The University of North Texas (UNT), one of the first three North American universities to require ETDs for graduation, began accepting theses and dissertations in electronic format in 1999. The UNT ETD collection consists of more than 17,000 items and more than 2 million pages or files. All of these documents share the characteristic that they are full-text searchable and have had metadata records.

Figure 1. The Usage Statistics for UNT ETDs, as of August 10, 2016:  
http://digital.library.unt.edu/explore/collections/UNTETD/stats/.

CHARACTERISTICS OF ETD  
Successful retrieval of materials that are useful to a user relies on the quality of the information representation. As stated by various information retrieval researchers (Rolling (1980), Alemneh et.al (2014), among others), indexing quality determines whether the information content of an indexed document is accurately represented. ETDs usually constitute original research; each is unique to the bibliographic world. As a result, catalogers need to provide original cataloging (i.e., creation of metadata record from scratch) as opposed to copy cataloging (i.e., use of pre-existing metadata record with or without augmentation) to describe each ETD.

Figure 2. The Classic Information Retrieval (IR) Model (Modified from Bates, M. J., 1989).
INDEXING AND INFORMATION RETRIEVAL
The successful management of information resources requires effort across the entire life cycle to ensure that contents are identified, organized, managed, preserved, and made accessible in a manner that today's users expect. Given the pressure of reading more in less time, today's users demand access to various formats regardless of temporal and spatial restrictions and the types of devices used. An index is simply a systematic representation of an information-bearing object (text, images, or other resource) that helps users find needles in the information haystack. It points users to specific items on topics of interest. It shows users related topics and indicates information trails through the vast information stores we have out there.

Identification of search terms and comparison of search results with the existing indexing terms used in describing ETDs provides a basis for assessing the relative usefulness of the indexing terms supplied by professional indexers and free-text keywords supplied by authors of ETDs in facilitating access to ETDs. At the end, indexing effectiveness measures whether an indexed item is correctly retrieved every time it is relevant to a query (Rolling, 1980).

To fully understand what an index is, it is necessary to be both micro- and macro-minded. On the micro level, we concern ourselves with the specific mechanics of creating an index. On the macro level we put an index into the larger context of an information retrieval system. Figure-3 shows the inverse relationship of precision and recall.

A good index helps users find what they need, even when they are not sure themselves what they need. Traditionally, inter-indexer consistency is a measure of indexing quality or effectiveness. However, as depicted in Figure-4, digital contents exhibit many aspects that make representation difficult. Recognizing the fact that aboutness is often in the eye of the beholder, the authors also discuss the subjectivity and objectivity of the process and the need to distinguish functional representation from mere descriptions of a topic.

REPRESENTATION AND ACCESS
The changes and expansion of types of indexes have been necessary because of the changes in information delivery mechanisms. Five hundred years ago books and book-like manuscripts made up the major part of a library. Two hundred years later the scholarly journal emerged and with it the invention of the abstracting journal. By the middle of the 19th century journals had proliferated to the point where indexes across journals were needed. As we moved into the digital age, new forms of indexes were needed to get access to the electronic information. Libraries now include media other than print. There are graphic materials, photographs, slides, drawings, art, and so on.

A review of the current landscape in digital libraries and emerging trends shows that there is no shortage of opinions on the role of indexing or appropriate representation in digital resources lifecycle management. Just to highlight some trends that impact resource organizations and related best practices:

- Heterogeneous, multimedia contents instead of text documents
- Complex retrieval systems instead of matching queries and document representations
- Visualization of the information space instead of a ranked list of search results
- Human information behavior instead of information needs
- Users as both creators and consumers of information instead of one or the other
- Deployment of new cataloging standards in libraries, including RDA (Resource Description and Access) and the exposure of RDA-based data in the linked data cloud instead of traditional cataloging tools such as MARC
- Various emerging Web applications driven by semantic web technologies such as the Web ontology language (OWL), the resource description framework (RDF), semantic Web rules language (SWRL), and other members of the World Wide Web Consortium (W3C) family of specifications offer powerful data organization, combination, and query capabilities.

**Figure 5. Representation, Access, and Delivery**

**METHODOLOGY**

This study reanalyzed the index terms in UNT’s ETD Collection from two sides -- the document side (supplied by creators and librarians) and the user side. It specifically tried to answer the effectiveness of the terms in matching users’ queries. To get a better sense of users’ discovery of digital resources, we also looked at the following two questions:

1. Were users arriving at our digital resources from searches that were answered by an item’s subject metadata? Users’ requests met by parts of the full text of the item were shown only for comparison purpose, but not fully analyzed. For use analysis, we employed Web server logs from the application server that provides access to the UNT Digital Library. The log files were limited to discoveries of items in the UNT ETDs collection. At the time this research was conducted, 11,873 unique items were available with metadata.

   The original raw dataset contained 172,115,682 lines during that timeframe, in the standard extended log file format.

   - Further limitations removed:
     - Requests made by known robots,
     - Requests without known search queries,

   The resulting (two-column) intermediary dataset contained 84,837 item-query pairs. A local identifier for each discovered item and the request used.

   Following normalization, the dataset contained 46,366 unique item-query pairs. However, the final dataset has 43,420 unique query results.

   2,946 samples were combined during processing because after further normalization they were no longer unique samples. As indicated in Tables 1 and 2, the final (43,420) dataset used in the remainder of this paper lists the percentage of each search query that was found in the:

   Metadata (full descriptive record). The four specified fields -- title, subject, agent (both creator and contributor values) and description further analyzed.

   Sample data from UNT’s ETD collection were collected. For the users’ side, the transaction logs of UNT Libraries’ ETDs collection use, collected by the Google Analytics application, were analyzed. For the indexing terms, data occurring in subject fields of metadata records describing electronic theses and dissertations were analyzed.

   This study analyzed access to UNT’s ETD Collection from searches that were answered by items’ subject metadata. Users’ requests met by parts of the full text of the item were shown only for comparison purpose, but not fully analyzed. For use analysis, we employed Web server logs from the application server that provides access to the UNT Digital Library. The log files were limited to discoveries of items in the UNT ETDs collection. At the time this research was conducted, 11,873 unique items were available with metadata.

   The original raw dataset contained 172,115,682 lines during that timeframe, in the standard extended log file format.

   - Further limitations removed:
     - Requests made by known robots,
     - Requests without known search queries,

   The resulting (two-column) intermediary dataset contained 84,837 item-query pairs. A local identifier for each discovered item and the request used.

   Following normalization, the dataset contained 46,366 unique item-query pairs. However, the final dataset has 43,420 unique query results.

   2,946 samples were combined during processing because after further normalization they were no longer unique samples. As indicated in Tables 1 and 2, the final (43,420) dataset used in the remainder of this paper lists the percentage of each search query that was found in the:

   Metadata (full descriptive record). The four specified fields -- title, subject, agent (both creator and contributor values) and description further analyzed.

   Sample data from UNT’s ETD collection were collected. For the users’ side, the transaction logs of UNT Libraries’ ETDs collection use, collected by the Google Analytics application, were analyzed. For the indexing terms, data occurring in subject fields of metadata records describing electronic theses and dissertations were analyzed.

   This study analyzed access to UNT’s ETD Collection from searches that were answered by items’ subject metadata. Users’ requests met by parts of the full text of the item were shown only for comparison purpose, but not fully analyzed. For use analysis, we employed Web server logs from the application server that provides access to the UNT Digital Library. The log files were limited to discoveries of items in the UNT ETDs collection. At the time this research was conducted, 11,873 unique items were available with metadata.

   The original raw dataset contained 172,115,682 lines during that timeframe, in the standard extended log file format.

   - Further limitations removed:
     - Requests made by known robots,
     - Requests without known search queries,

   The resulting (two-column) intermediary dataset contained 84,837 item-query pairs. A local identifier for each discovered item and the request used.

   Following normalization, the dataset contained 46,366 unique item-query pairs. However, the final dataset has 43,420 unique query results.

   2,946 samples were combined during processing because after further normalization they were no longer unique samples. As indicated in Tables 1 and 2, the final (43,420) dataset used in the remainder of this paper lists the percentage of each search query that was found in the:

   Metadata (full descriptive record). The four specified fields -- title, subject, agent (both creator and contributor values) and description further analyzed.

   Sample data from UNT’s ETD collection were collected. For the users’ side, the transaction logs of UNT Libraries’ ETDs collection use, collected by the Google Analytics application, were analyzed. For the indexing terms, data occurring in subject fields of metadata records describing electronic theses and dissertations were analyzed.
• The numbers overlap in cases where tokens appeared in both indexes: 32,056.
• A number of record discoveries were dependent entirely on the full text: 9463.
• A number of record discoveries were dependent entirely on the metadata: 1723.

<table>
<thead>
<tr>
<th>Matches found in:</th>
<th>Total No. of Queries Found:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any part of query in full text</td>
<td>41,519</td>
</tr>
<tr>
<td>Any part of query in metadata</td>
<td>33,779</td>
</tr>
<tr>
<td>Both any metadata and full text</td>
<td>32,056</td>
</tr>
<tr>
<td>100% of query in full text</td>
<td>36,318</td>
</tr>
<tr>
<td>Queries ONLY in full text (but not in metadata)</td>
<td>9463</td>
</tr>
<tr>
<td>100% of query in metadata</td>
<td>29661</td>
</tr>
<tr>
<td>Queries ONLY in metadata (but not in full text)</td>
<td>1723</td>
</tr>
</tbody>
</table>

Table 3. Record discoveries based on matches in metadata and full text. (n=43420)

Figure 6 shows how many items could be found using either index equally, how many had a partial match in one index, with a full match in the other index, and the number of queries that could be found only through the combination of metadata and full text versus either index alone.

CONCLUSION

Considering the multiple stakeholders in the digital ecosystems and their evolving needs and requirements, successful management of digital materials requires constant maintenance, which is to say individual items are curated and actively managed during their entire lifecycle. The findings of this study will help to revisit the two traditional approaches that demonstrate the process of representation: the document-oriented approach, which claims that indexing summarizes or represents the content of a document, and the user-oriented approach, which requires that indexing reflect the requests for which a document might be relevant. Considering the interdisciplinary nature of ETDs and the diverse global user communities, the authors further argue that effective retrieval depends not only on the indexing terms assigned to describe ETDs, but on the search query terms entered by users as well. Regardless, effective metadata and taxonomies add value and amplify the mostly interdisciplinary ETDs—allowing users to explore and delve deeper in multidimensional ways.

REFERENCES


