UNDERSTANDING USER DISCOVERY OF ETD: METADATA OR FULL-TEXT, HOW DID THEY GET THERE?

DANIEL GELAW ALEMNEH, MARK EDWARD PHILLIPS, LAURA WAUGH, AND HANNAH TARVER
UNIVERSITY OF NORTH TEXAS
OUTLINE

• BACKGROUND
• RESEARCH QUESTION
• ETDs CASE STUDY
• USAGE STATISTICS
• ANALYSIS
• SUMMARY
BACKGROUND

• The digital environment has now introduced new resource types and new user expectations into the current information landscape.
• As forms of digital scholarship evolve, so do users’ and creators’ roles and expectations.
• Advancing knowledge requires not only enhancing our capacity to generate more knowledge, but also cultivating our ability to provide seamless discovery process to the vast quantities of knowledge we continue to generate.
INTRODUCTION

• In an ideal world, digital content would also be given the same consideration as other library materials
  • When conducting collection development, organization and cataloging of works, etc.

• Discussions of library cataloging and metadata groups:
  • But are users are in fact making use of the metadata records we create for them?
  • Are they able to find the resources they want with just the full text of the item they are looking for?
RESEARCH QUESTION

• To get a better sense of users discovery of digital resources, we decided to assess and see:
  • Whether users were arriving at our digital resources from searches that were answered by an item's descriptive metadata or by parts of the full-text of the item.
  • If or when a resource was found with metadata, what fields were being used to retrieve that item.
ETDs CASE STUDY

• At UNT, we have a combination of both born-digital ETDs and digitized analog theses and dissertations.
  • The UNT ETD collection consists of close to 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 created for them as part of the inclusion of the item in our Digital Library.

• UNT tracks the use of its ETD, as well as the use of other digital collections hosted by UNT Libraries.
  • ETD Collection is heavily used, with daily use level ranging between 2,000 and over 5,000 uses from 200+ countries.
FIGURE 1. THE USAGE STATISTICS FOR UNT ETDS, (AUGUST - SEPTEMBER, 2015)

Statistics for UNT Theses and Dissertations

Item Usage  Added Items  More Data

5,036,997 Total Uses / 16,855 Total Items (2,044,056 files) / 12,207 Visible / 4,648 Hidden
### FIGURE 2. TOP 10 COUNTRIES USERS FROM WHICH USED THE UNT ETDS COLLECTION
(AUGUST - SEPTEMBER, 2015)

<table>
<thead>
<tr>
<th>Country</th>
<th>Sessions</th>
<th>% New Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>390,314</td>
<td>86.60%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>33,357</td>
<td>90.66%</td>
</tr>
<tr>
<td>India</td>
<td>31,098</td>
<td>92.89%</td>
</tr>
<tr>
<td>Philippines</td>
<td>24,878</td>
<td>93.16%</td>
</tr>
<tr>
<td>Canada</td>
<td>18,398</td>
<td>91.50%</td>
</tr>
<tr>
<td>Australia</td>
<td>10,440</td>
<td>90.13%</td>
</tr>
<tr>
<td>Germany</td>
<td>8,988</td>
<td>89.92%</td>
</tr>
<tr>
<td>Malaysia</td>
<td>8,951</td>
<td>88.63%</td>
</tr>
<tr>
<td>(not set)</td>
<td>8,603</td>
<td>91.63%</td>
</tr>
<tr>
<td>China</td>
<td>6,095</td>
<td>89.01%</td>
</tr>
</tbody>
</table>
METHODOLOGY

• This study analyzed access to UNT’s ETD Collection from two sides:
  • Searches that were answered by an item’s descriptive metadata
  • Users request met by parts of the full-text of the item.

• For use analysis, we used 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 available with metadata
  • Occurred between May 4, 2014 and January 24, 2015.
METHODOLOGY …

• 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.
METHODOLOGY …

• The UNT Digital Library uses the Solr full-text indexer, which can provide:
  • Explanatory information noting why a query yielded certain results, if a specific document would be returned by a query, and in which specified fields the terms appear.
  • To utilize this, the item-query pairs were fed to the Solr search system.

• The final (43,420) dataset used in the remainder of this paper lists the percentage of each search query that was found in the:
  • Full text,
  • Metadata (full descriptive record),
  • The four specified fields -- title, subject, agent (both creator and contributor values) and description further analyzed.
### TABLE 1. EXAMPLE DATASET ENTRIES FOR THREE SEARCH QUERIES

<table>
<thead>
<tr>
<th>Dataset Field</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>metadc129697</td>
<td>metadc146510</td>
<td>metadc155618</td>
</tr>
<tr>
<td>Query</td>
<td>susan cheal</td>
<td>human trafficking</td>
<td>article writing</td>
</tr>
<tr>
<td>Query Tokens</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PageText</td>
<td>0%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Metadata</td>
<td>100%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Title</td>
<td>0%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Subject</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Agent</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Description</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>
FINDINGS

• Table 2. Statistics for the number of queries per unique item.

<table>
<thead>
<tr>
<th>N</th>
<th>min</th>
<th>median</th>
<th>max</th>
<th>sum</th>
<th>mean</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>9794</td>
<td>1</td>
<td>2</td>
<td>24</td>
<td>43420</td>
<td>4.43</td>
<td>2.92</td>
</tr>
</tbody>
</table>

• Table 3. Statistics for the number of tokens per query.

<table>
<thead>
<tr>
<th>N</th>
<th>min</th>
<th>median</th>
<th>max</th>
<th>sum</th>
<th>mean</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>43420</td>
<td>1</td>
<td>2.5</td>
<td>31</td>
<td>104102</td>
<td>2.40</td>
<td>1.59</td>
</tr>
</tbody>
</table>

• Although queries varied in length, they were analyzed as individual words (or tokens) rather than phrases. This allowed for partial matches in a given field, resulting in percentages less than 100. The distribution of tokens across queries ranged from 1 to 31 tokens.
Table 4 gives a breakdown of the total number of queries found:

- 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**
Matches found in: Percentage of Queries Found:

Queries ONLY in metadata: 4%
Queries ONLY in full text: 22%
Both any metadata and full text: 74%
Table 5 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 (806) that could be found only through the combination of metadata and full-text versus either index alone.

<table>
<thead>
<tr>
<th>Overlapping matches found in:</th>
<th>Number of queries:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata / full text equally (m=100/p=100)</td>
<td>23,935</td>
</tr>
<tr>
<td>ONLY with metadata / full text combined (m&lt;100/p&lt;100)</td>
<td>1,376</td>
</tr>
<tr>
<td>Metadata / partial full text (m=100/p&lt;100)</td>
<td>5,726</td>
</tr>
<tr>
<td>Partial metadata / full text (m&lt;100/p=100)</td>
<td>12,383</td>
</tr>
</tbody>
</table>
At a more granular level, Table 6 displays average query matches by field: title, subject/keyword (Subj.), agent, or content description (Descr.). Most terms were found in the agent and subject fields.
Table 7 shows record discoveries broken down by match percentages of each field, for the entire dataset. This shows the extent of the matches (partially for longer query strings) and the overlap across multiple fields.

<table>
<thead>
<tr>
<th></th>
<th>0%</th>
<th>1-49%</th>
<th>50-74%</th>
<th>75-99%</th>
<th>100%</th>
<th>%&gt;=1% found in field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>33,597</td>
<td>2,086</td>
<td>2,297</td>
<td>120</td>
<td>5,320</td>
<td>22.62%</td>
</tr>
<tr>
<td>Subj.</td>
<td>28,661</td>
<td>1,591</td>
<td>1,736</td>
<td>61</td>
<td>11,371</td>
<td>33.99%</td>
</tr>
<tr>
<td>Agent</td>
<td>29,276</td>
<td>193</td>
<td>293</td>
<td>4</td>
<td>13,654</td>
<td>32.57%</td>
</tr>
<tr>
<td>Descr.</td>
<td>29,274</td>
<td>3,048</td>
<td>3,454</td>
<td>350</td>
<td>7,294</td>
<td>32.57%</td>
</tr>
</tbody>
</table>

Table 7. Record Discoveries Per Field Based on Percentage of Query Present in Field. (N=43420)
Table 8 shows the match percentage by field for the subset of 1723 items discovered through metadata-only retrieval. Most query matches were in the agent fields, followed by subject with few matches coming from description or title.
Distribution of Queries Found within Metadata Fields

- Agent: 53%
- Others: 32%
- Subj.: 13%
- Descr.: 2%
- Title: 0%
CONCLUDING REMARKS

• Arguably, there has been a shift in the way users search, access, and use information resources.

• Although the starting point for new researches are increasingly digital, the challenge yet to be overcome in the provision of access to digital contents or achieving a level of description sufficient to ensure success in information retrieval.

• Effective metadata and taxonomies add value and amplify the mostly interdisciplinary ETDs—allowing users to explore and delve deeper in multi-dimensional ways.
Percentage of Matches Queries Found:

- Queries ONLY in full text (22%)
- Queries ONLY in metadata (4%)
- Both any metadata and full text (74%)
CONCLUDING REMARKS

• Considering the diverse global ETDs users’ communities, the authors further argue that effective retrieval depends not only on the subject terms assigned to describe an item, but on the search query terms entered by users as well.
TOP 50 TERMS REPEATEDLY USED BY UNT ETD SEARCHERS (MAY 2011 TO MAY 2015)
THANKS!

AND

QUESTIONS?

DANIEL.ALEMNEH@UNT.EDU