Tracing the Evolution of Imaging in Libraries and Archives: A Look at Professional Preservation Discourse via *Preservation, Digital Technology & Culture*

Presented by Karen F. Gracy, Ph.D.
School of Information
Inspiration for This Research (Work-in-Progress)

• 50th Anniversary of PDTC is this year (1972-2022).
• Journal’s scope covers a significant period in the evolution of the imaging and preservation fields.
• Curiosity about how the journal’s scope and editorial focus has changed over the years.
• Wanted to do a special issue that looked at this question using a variety of approaches, both quantitative and qualitative.
  • Study of editorials and interviews with past editors where possible to determine editorial intent (solicitation and selection of content).
  • Creation of network of critical actors (based on who published articles and reviews in the journal, who was on the editorial board, etc.).
  • Identification and analysis of patterns of discourse and topical coverage in journal using NLP approaches.
Microform Review

1972-1998 (ind. publ.)

1998-2005 (K.G. Saur)

Microform & Imaging Review

2005-2012 (De Gruyter)

Microform & Digitization Review

2012-present (De Gruyter) Peer-reviewed

Preservation, Digital Technology & Culture
Journal Titles and Editors-in-Chief, 1972-2022

**Microform Review, 1972-1998**
- Allen B. Veaner (founding editor), 1972-1985
- Thomas A. Bourke, 1985-1989
- Susan Marie Szasz, 1990-1993
- Wendy Thomas, 1994-1998

**Microform & Imaging Review, 1998-2005**
- Julie Arnott, 1998-1999
- Steven D. Smith, 1998-2004
- Suzy Szasz Palmer, 2004

**Microform & Digitization Review, 2005-2012**
- Ken Middleton, 2005-2012

**Preservation, Digital Technology & Culture**
- Michèle V. Cloonan, 2013-2018
- Leisa Gibbons, 2018-2019
- Karen F. Gracy, 2018-2022
A "Digital Humanities" Approach: Using NLP Techniques to Uncover Patterns in Journal Discourse

• Research Goals:
  1. Identify key topics covered by the journal over the course of 50 years
  2. Compare differences in scope as journal title and editorial responsibility changed
  3. Relate scope shifts to changes in uses of imaging/reformatting technologies in cultural heritage stewardship
     • Milestones in preservation field gathered via secondary literature

• NLP Offers Quantitative Approaches for Achieving First and Second Goals
  • Computer-aided data analysis allows for much quicker examination of texts and identification of patterns than can be accomplished via manual processes
  • Can help see larger picture and test hypotheses
  • Might also have predictive power in certain circumstances
The Journal Corpus (Data Sources and Associated Metadata)

• Authors for all feature articles
• Titles for all feature articles
• Abstracts for all feature articles
• Keywords (of limited use, as only included from 2015-onward)
• Not included at this time:

  • Editorials
    • Not always included in each issue; editorials often focus on summaries of issue contents rather than statements about the field
  • Reviews of microform collections, books, events, etc.
  • Text and metadata of reviews (of microform collections, books, other resources)
    • Less likely to be relevant as discourse for this research
  • Full text for feature articles
    • Data collection and clean up too labor-intensive at this time.
Data Collection and Preparation/Clean-Up

• Complete run of journal was available in print form, with access to metadata via electronic indexes.

• Authors, titles, years of publication, volume/issue numbers, abstracts, and keywords were collected via a combination of manual and semi-automated techniques.
  • Spreadsheet was used to put data in tabular form (each row corresponding to one document).
  • Plain text documents were created as needed for analysis, depending on need (e.g., all abstracts for articles appearing in a particular date span)
  • Clean-up was light, mostly involving fixing structural errors and inconsistencies (like typos and variations in abbreviations)
  • Keywords were limited (2015 onward); will investigate use of volume indexes found in issues to add more keywords, but this will involve a lot of manual labor

• Full-text of feature articles were not collected at this time (to be done in the future)
Analysis of Data

• Types of Analyses
  • Term frequency and distribution (over time)
  • Collocations
    • Shows terms whose frequencies rise and fall together
  • Concordances
    • Shows words in context, with a bit of surrounding text
  • Topic modeling
    • Generates term clusters from a document or corpus and then shows how each term cluster is distributed across the document or corpus

• Toolset Used: Voyant Tools (www.voyant-tools.org)
  • This is a basic set of tools for newcomers to text analysis.
  • It doesn’t have all the features one would like, such as being able to chart text against date (but there are work-arounds for this corpus, as we will see).
Visualization of Data

• **Techniques Used:**
  • Word clouds (shows top frequencies of words in a document or corpus)
  • Line graphs
  • Tabular displays
  • Topics (generates term clusters from a document or corpus and then shows how each term cluster is distributed across the document or corpus)
NLP Example #1: Comparing Term Frequencies (Top Ten Over Journal Run)

NLP Example #1: Comparing Term Frequencies (Top Ten Over Journal Run)

**Microform & Imaging Review, 1998-2005**

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>library</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>preservation</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>information</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>collection</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>libraries</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>university</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>digitization</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>collections</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>project</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>
NLP Example #1: Comparing Term Frequencies (Top Ten Over Journal Run)

Microform & Digitization Review, 2005-2012

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>preservation</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>library</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>article</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>heritage</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>archives</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>project</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>university</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>cultural</td>
<td>68</td>
<td></td>
</tr>
</tbody>
</table>

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (2022) Privacy v. 2.6.0 [2022/07/27]
NLP Example #1: Comparing Term Frequencies (Top Ten Over Journal Run)

Preservation, Digital Technology & Culture, 2013-2021
NLP Example #2A: Use of Collocations to Identify Multi-Word Concepts

[Image of Voyant Tools interface with collocates for terms such as digital, preservation, heritage, etc.]
NLP Example #2B: Use of Concordances to Examine Words in Context

Phrases in which the word “digital” appears (PDTC, 2013-2021)
NLP Example 3: Comparison of Frequency of Words Over Journal Run

Use of the words microform and digital in titles and abstracts over the run of the journal

(could be even more precise by using word stemming, i.e., digit* and micro*)
Next Steps

• Add to corpus by collecting and cleaning full-texts of articles.
• Generate additional N-grams for corpus.
  • N-gram = “Contiguous sequence of $n$ items from a given sample of text or speech”
  • This will allow me to track multiple-word concepts across the corpus and see how concepts become associated with particular discourses relating to the use of various technologies in imaging for preservation.

• Triangulate these findings with other types of inquiry, e.g.:
  • Conduct qualitative analysis of editorials and interviews with past editors.
  • Generate list of concepts via robust secondary literature review in order to build ontology (domain knowledge) for the preservation field
    • Preferably with historical dimension to track change over time.

• Assess this project as a model for conducting similar projects with other journal/scholarly literature corpora in the LIS/archives fields.
• Compare use of terms in this specialized corpus to much larger corpora with broader scope (e.g., HathiTrust, Google Books).
Thank You for Listening!

Questions?

Want to Hear About Progress on This Project?

Contact me at kgracy@kent.edu.

Thank You for Listening!

Many thanks to my graduate assistant, Ms. Elaine Yuen, Ph.D. student (College of Communication and Information) for her data collection and analysis support.