Using Open Source Tools to Improve Access to Oral History Collections

Becky Yoose, Bibliographic Systems Librarian and Jody Perkins, Metadata Librarian
Miami University Libraries, Miami University, Oxford, OH

Miami Stories Oral History Project

- Begin in 2005, coordinated by University Archives, CONTENTdm collection maintained by Digital Initiatives
- Current and former students, faculty, and staff, as well as friends of the University share recollections of their Miami years
- 100 oral history interviews
  - Average length of interview: 2 hours
  - 100 oral history interviews
  - Half have been fully transcribed

OpenCalais

- Released in 2008, used by various companies, news agencies, and publishers
- Uses natural language processing and machine learning to extract categorized metadata (in RDF format) from full text documents
- API modules, applications available for different platforms

Drupal

- Popular Open Source content management system (CMS) built with PHP
- Used widely for web sites and blogs
- Flexible and customizable, over 8,000 modules

Miami Stories OpenCalais Pilot

Human metadata creation workflow: Each interviewer had a cover sheet to list key terms and topics relevant to the interview. These terms were entered as keywords into item records and supplemented with FAST (Faceted Application of Subject Terminology) headings - a controlled vocabulary based on Library of Congress Subject Headings and L.C. Authority files.

Human metadata creation issues: Data collected on cover sheets varied with the amount of time and number of staff available for a given interview – the sheets varied from having no data to over 50 topics for a single interview. Names entries even for interviewees were inconsistent. The Libraries did not have the staff to manually go through 60+ interview transcripts to manually extract metadata.

Pilot project goal: Experiment with applications that automatically generated index terms from full text as more efficient way to provide access points for this collection at the item level.

The OpenCalais API offered a number of advantages that made it ideal for this purpose. The faceted framework it employs makes it easy for categories that are essential to these kinds of historical collections – names, places and locations in particular.

Observations

- OC generates a much larger number of access points, but OC results also included a larger number of false hits/inaccuracies
- OC categories provide a less granular browsing structure
- Terms representing contextual and relational information are lacking in OC results
- Certain aspects of the OC schema don’t suit the content (many irrelevant categories) and there are numerous gaps when compared to the catalogue created metadata
- Meaning of many OC categories is ambiguous making index terms difficult to interpret
- Preservation and genre metadata not captured (since OC only processes text)
- Subject indexing seems to be a weakness of OC – it only generated a few very broad terms, though it did so with a great deal of accuracy
- Name indexing (people, organizations, facilities and locations) seems to be a real strength of OC

Sample of name entries issues

<table>
<thead>
<tr>
<th>OpenCalais Name Entries</th>
<th>CONTENTdm Name Entries</th>
<th>Report-calais false Quality Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charles Wilson</td>
<td>Wilson, Charles</td>
<td>updated</td>
</tr>
<tr>
<td>Curtis Ellison</td>
<td>Ellison, Curtis [interviewee]</td>
<td></td>
</tr>
<tr>
<td>Ed Branch</td>
<td>Branch, Edgar Marquis, 1913</td>
<td></td>
</tr>
<tr>
<td>Ehlers</td>
<td>Ehlers, Robert</td>
<td>OC created index terms</td>
</tr>
<tr>
<td>Histori, Moe Zemper</td>
<td>N/A</td>
<td>OC false hit</td>
</tr>
<tr>
<td>Roland Delaney</td>
<td>Delaney, Roland</td>
<td></td>
</tr>
<tr>
<td>Roland Delaney</td>
<td>N/A</td>
<td>OC indexed without error</td>
</tr>
</tbody>
</table>

Even though the following table shows some of the name entries issues from OC, it should be noted that OC on average created more name entries than catalogers.

Sample of assigned subjects

<table>
<thead>
<tr>
<th>CONTENTdm topics</th>
<th>OC topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Vietnam War protests</td>
<td>War</td>
</tr>
<tr>
<td>Black Student Action</td>
<td>Music</td>
</tr>
<tr>
<td>Vietnam War veterans</td>
<td>Education</td>
</tr>
<tr>
<td>Butler Co. Sheriff</td>
<td>Jim Zwick (monocatagorized)</td>
</tr>
<tr>
<td>Faculty leaving the University</td>
<td>Politics</td>
</tr>
<tr>
<td>Faculty Senate</td>
<td>Religion</td>
</tr>
<tr>
<td>Great Revolution</td>
<td>Technology (false hit)</td>
</tr>
</tbody>
</table>

Want to learn more about the technical details of this project? Scan the QR code or visit bit.ly/1HEID for more information!

For further information

Miami Stories Oral History Project
http://library.lib.miamiu.edu/clir/moa/stories/
OpenCalais
http://www.opencalais.com/
Drupal
http://drupal.org/
Becky Yoose, Bibliographic Systems Librarian
perkins@muohio.edu
Jody Perkins, Metadata Librarian
perkins@muohio.edu