Mining Oral History for Enhanced Access

Jody Perkins, Metadata Librarian and Becky Yoose, Bibliographic Systems 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 shareollections of their Miami years
- 100 oral history interviews
  -> Average length of interview: 2 hours
  -> Half have been fully transcribed

OpenCalais

- Refined 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
- Popular Open Source content management system (CMS) built with PHP
- Used widely for web sites and blogs
- Omeka
- Application of Subject Terminology
- CONTENTdm
- Starting with the Oral History Collection
- CONTENTdm includes a common data model
- CONTENTdm faceted searching
- CONTENTdm provides metadata that is easy to use and easy to enter
- CONTENTdm integrates well with other platforms

Pilot project goal:
- Experiment with applications that automatically generate index terms from full text as keywords
- Name indexing
- Subject indexing
- The Libraries did not have the staff to manually extract metadata
- This pilot project is a step toward automating the metadata creation workflow and eliminating as much variability as possible between human indexers

OpenCalais

- 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
- OC can provide substantial efficiencies when working with large volumes of full text especially for collections where terms representing people, organizations, facilities and locations are deemed critical access points

Possible Next Steps

- Data quality study - measure data quality using established criteria for:
  - Aboutness / substantive coverage
  - Accuracy
  - Completeness
  - Context
  - Consistency
  - Interoperability
  - Usability

Integration, display, and sharing:
- Currently the Oral History Project is hosted on CONTENTdm; however, the Libraries are in the process of migrating several collections to DSpace. In light of this move, the metadata generated from this project, along with the videos, transcriptions, and descriptive metadata, might be calling one of the following platforms “home” in the near future:
  - Drupal
  - Omeka
  - OpenWMS

Outcomes

- Outcome:
  - OC can provide substantial efficiencies when working with large volumes of full text especially for collections where terms representing people, organizations, facilities and locations are deemed critical access points.

Migrating transcriptions into Drupal & batch processing using the Calais Drupal module in six [oversimplified] easy steps!

1. Export XML from CONTENTdm to MySQL
2. Import MySQL table using Table Wizard
3. Migrate content into Drupal using modules
4. Edit Calais node settings
5. Batch process transcripts using Calais
6. Profit!

Sample of name entries issues

<table>
<thead>
<tr>
<th>OpenCalais Name Entries</th>
<th>CONTENTdm Name Entries</th>
<th>OpenCalais Data Quality Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charles Wilson</td>
<td>Wilson, Charles</td>
<td>qualified</td>
</tr>
<tr>
<td>Curtis Ellison</td>
<td>Ellison, Curtis [interviewer]</td>
<td>unqualified</td>
</tr>
<tr>
<td>Ed Branch</td>
<td>Branch, Edgar Marquess, 1913-</td>
<td>OC indexed other form</td>
</tr>
<tr>
<td>Ethridge</td>
<td>Ethridge, Robert</td>
<td>OC indexed first name</td>
</tr>
<tr>
<td>History’s “Men of Fame”</td>
<td>N/A</td>
<td>OC false hit</td>
</tr>
<tr>
<td>Roland Delator</td>
<td>Delator, Roland</td>
<td>N/A</td>
</tr>
<tr>
<td>Roland Delato</td>
<td>N/A</td>
<td>OC indexed spelling error</td>
</tr>
</tbody>
</table>

Sample of assigned subjects

<table>
<thead>
<tr>
<th>CONTENTdm topics</th>
<th>OpenCalais tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Vietnam War protests</td>
<td>War</td>
</tr>
<tr>
<td>Black Student Action Committee</td>
<td>The Organ (false hit)</td>
</tr>
<tr>
<td>Black Cis. Shariff</td>
<td>Music</td>
</tr>
<tr>
<td>Faculty leaving the University</td>
<td>Education</td>
</tr>
<tr>
<td>Faculty Senate</td>
<td>Jim Zwerg (misc. category)</td>
</tr>
<tr>
<td>Gentle Revolution</td>
<td>Politics</td>
</tr>
<tr>
<td>Long hair and beard on men</td>
<td>Religion</td>
</tr>
<tr>
<td>Rowan Hall occupation</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 http://bit.ly/1YHEID for more information!

Miami Stories Oral History Project
http://dl.lib.muohio.edu/dlib/musonomies/OpenCalais
http://www.opencalais.com/
Drupal
http://drupal.org/

For further information
Jody Perkins, Metadata Librarian
perkinsj@muohio.edu
Becky Yoose, Bibliographic Systems Librarian
yoosebj@muohio.edu
Miami University Libraries
http://miamilib.muohio.edu

Download the full text at: http://miamilib.muohio.edu/content/services/downloads/mharc.pdf