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 share recollections of their Miami years
- 100 videotaped interviews
  - →Average length of interview: 2 hours
  - →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

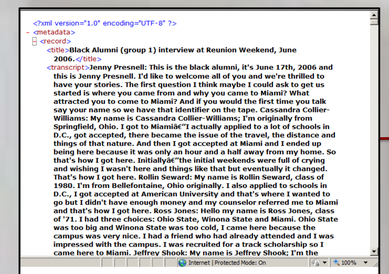
Human metadata creation workflow: Each interviewer had a cover sheet to list key terms and topics relevant to the interview. These terms were entered as keywords into item records and supplemented with FAST (Faceted Application of Subject Terminology) headings - a controlled vocabulary based on Library of Congress Subject Headings and related LC 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. Name entries even for interviewees were inconsistent. The Libraries did not have the staff to manually go through 60+ interview transcripts to manually extract metadata.

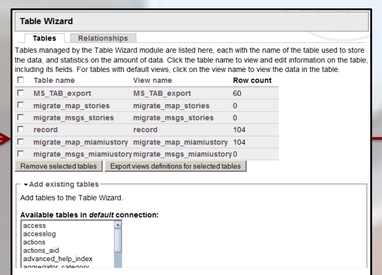
Pilot project goal: Experiment with applications that automatically generated index terms from full text as more efficient way to provide access points for this collection at the item level.

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

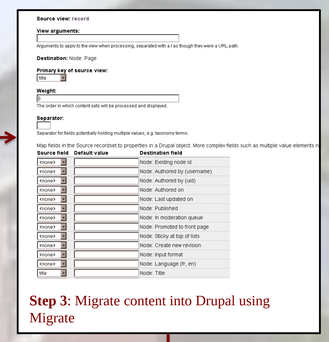
Migrating transcripts into Drupal & batch processing using the Calais Drupal module in six [oversimplified] easy steps!



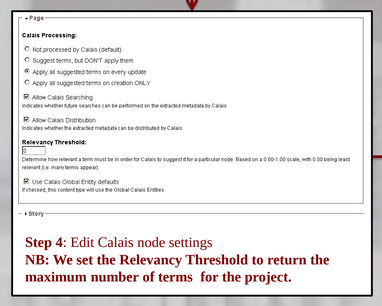
Step 1: Export XML from CONTENTdm to MySQL



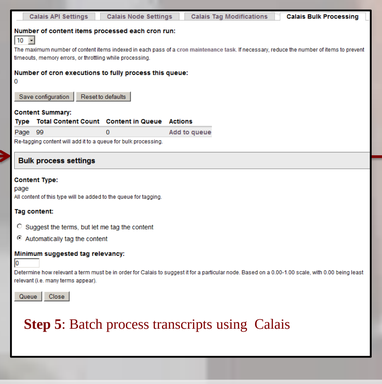
Step 2: Import MySQL table using Table Wizard



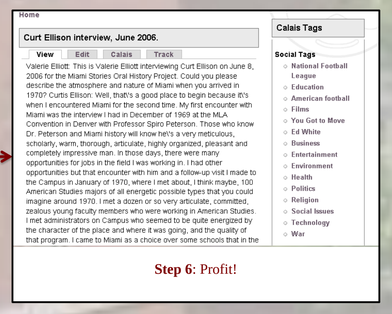
Step 3: Migrate content into Drupal using Migrate



Step 4: Edit Calais node settings
NB: We set the Relevancy Threshold to return the maximum number of terms for the project.



Step 5: Batch process transcripts using Calais



Step 6: Profit!

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 cataloger 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 significant **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 – particularly where context is not an issue.

Sample of name entries issues

OpenCalais Name Entries	CONTENTdm Name Entries	OpenCalais Data Quality Issues
Charles Wilson	Wilson, Charles	
Curtis Ellison	Ellison, Curtis [interviewer]	unqualified
Ed Branch	Branch, Edgar Marquess, 1913-	OC indexed other form
Etheridge	Etheridge, Robert	OC missed first name
Hitler's "Mein Kampf"	N/A	OC false hit
Roland Delattre	Delattre, Roland	
Roland DeLattro	N/A	OC indexed spelling error

It should be noted that OC on average created more name entries than catalogers. But some of those entries could not be associated with any substantive content.

Sample of assigned subjects

CONTENTdm topics
- Anti-Vietnam War protests
- Black Student Action Committee
- Butler Co. Sheriff
- Faculty leaving the University
- Faculty Senate
- Gentle Revolution
- Long hair and beards on men
- Rowan Hall occupation
- Voices of Reason

OpenCalais tags
- War
- The Organ (false hit)
- Music
- Education
- Jim Zwerg (mis-categorized)
- Politics
- Religion
- Technology (false hit)

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 that 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, transcripts, and descriptive metadata, might be calling one of the following platforms "home" in the near future:
- Drupal
- Omeka
  http://omeka.org/
- OpenWMS
  http://rucore.libraries.rutgers.edu/open/projects/openwms/

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

For further information

Miami Stories Oral History Project
http://doyle.lib.muohio.edu/cdm4/mustories/

OpenCalais
http://www.opencalais.com/

Drupal
http://drupal.org/

Becky Yoose, Bibliographic Systems Librarian
yoosebj@muohio.edu
Jody Perkins, Metadata Librarian
perkintj@muohio.edu