Mining Oral History for Enhanced Access

Jody Perkins, Miami University Libraries, Oxford, OH
Becky Yoose, Miami University/Grinnell College Libraries, Grinnell, IA*

***Becky was working at Miami University during the project. She has since started working at Grinnell College.**

A version of this paper was presented as a poster at the Society of American Archivists Annual Conference, August 25, 2011.

About the Miami Stories Oral History Project

Coordinated by the University Archives, Miami Stories invites groups of people with common experiences--current and former students, faculty, and staff, as well as friends of the University--to offer recollections of their Miami years and to hear reminiscences of those who shared them. A University Bicentennial legacy project that began in 2005, Miami Stories are recorded and stored in digital video formats so that future generations may enjoy and study them, and scholars may have ready access to a variety of perspectives on Miami's past.

The Oral History Project coordinators, Valerie Elliott and Jenny Presnell (University Libraries), work with an advisory committee of Library personnel and Miami faculty, along with the assistance of a program associate. The coordinators oversee the project as well as plan, schedule, and conduct interviews.

Throughout the year interview teams conduct story circles with diverse groups. These groups have included the Campus Owls dance band, fraternity members, the Hughes Society of 50-year alumni, Miami Student newspaper staff, the Alumni Band, Black Alumni, Educational Leadership alumni, McGuffey School faculty and alumni, and others. It is possible to conduct more interviews at certain times of the year than others; for example, Reunion Weekend is a campus event would be the only time of the year to reach certain alumni and former Miami faculty and staff.

The interviews are conducted in a story circle format. In a story circle, a moderator, typically a current Miami faculty or project staff member facilitates a conversation in which participants discuss their Miami experiences. The number of interviewees range from one to eight per interview. Interviews are videotaped by another staff member. While the interview is in progress or immediately afterward (staff and time permitting) keywords are noted on a cover sheet.

Currently the collection is comprised of approximately 100 videotaped interviews averaging two hours in length about half of which have been fully transcribed. The interviews – video, basic metadata and transcript – are then uploaded into a ContentDM installation hosted by the University Libraries. During the cataloging process terms are taken from cover sheets and entered as keywords into item records. Where possible keywords were supplemented with authorized headings from FAST (Faceted Application of Subject Terminology) - a controlled vocabulary based on Library of Congress Subject Headings and other related LC Authority files.

About the Miami Stories OpenCalais Pilot

Because of the workflow used to upload the interviews into ContentDM, access to the interviews depends heavily on human decision making and data entry. However, this process has proven to be inconsistent, making it difficult to provide a consistent level of access to the interviews through the public interface. Data collected on cover sheets varies with the amount of time and available staff. The number of keywords added to coversheets varied from none to over 50 for a single interview. Name entries, even for interviewees were inconsistent. Cataloging staff did not have the time that would have been required to listen to interviews or review transcripts for all of the items lacking sufficient coversheet metadata.

The Metadata Librarian had been experimenting with applications that automatically generated index terms from full text as a more efficient way to provide access points for this collection at the item level.

She began discussions with the Bibliographic Systems Librarian to explore ways to do the process in batch, which then lead to the possibility of creating an alternative facetted interface based on the index terms. After looking into available options two applications were chosen for the project – Drupal and OpenCalais.

Technical Process (See Appendix A for technical specifications)

OpenCalais is an application that uses natural language processing and machine learning to extract categorized metadata from full text documents. OpenCalais, during the time of the project, was used by major news websites to generate the metadata providing access to their content. An example closer to the Library environment was the Powerhouse Museum, where they used OpenCalais to create faceted metadata for their items. The developers released the API for use in different platforms as well as locally-developed applications. The OpenCalais API offered a number of advantages that made it ideal for this project. One of the main advantages was the faceting framework it employs makes use of categories that are essential to these kinds of historical collections – names, places and locations in particular. The ability to create a relevant list of terms from a document more quickly than a human was also particularly attractive, especially given the lack of staff to provide an in-depth metadata creation of each transcript.

Drupal became the platform of choice fairly early in the project. The University Libraries had worked extensively with Drupal before the project. The Libraries' website and discovery layer were both Drupal-based. Keeping with the same platform would allow the project to take advantage of an existing testing Drupal environment as well as staff who were experienced with Drupal. Drupal itself is well established and has an extensive community for support and development. The Drupal framework allowed for flexibility needed for an environment to explore ways for processing transcripts and creation of terms from the transcripts. The Calais module for Drupal provided several options in processing and term creation as well.

Exporting/Importing transcripts

The project started with the exporting of transcripts from one platform (ContentDM) to another (Drupal). The item level metadata and transcripts were exported in XML format from ContentDM. From there, the XML file was imported into Microsoft Access for preparation to import the transcripts into the MySQL database used for the test Drupal installation. Preparation of the data and transcript included creating column names and properties, such as setting the primary key. After the preparation was complete, the transcripts and metadata were imported into a new table in the Drupal MySQL using the OBDC database export in Access.

In order for the newly created table to be recognized in Drupal, the data had to go through two steps using two modules commonly used to migrate content from a different platform to Drupal. The first module was Table Wizard, a module that, among other functions, listed the tables that exist in the MySQL database. Table Wizard was also the first step in importing the transcripts into the test Drupal installation. From the "add existing table" option, the transcript table could be selected and added to the recognized list of tables for the Drupal installation to pull content from. The "analyze link" option of the module gave the opportunity to see if the table's properties, such as fields and primary key, had been imported correctly.

The Table Wizard module prepped the transcripts and metadata for the second step of migration. The second module, Migrate, was where the content is configured before migration into Drupal. The "add

content set" function set up the migration process by giving the options to select the MySQL table (prefixed with a "tw" indicating that it was imported using Table Wizard) to be migrated and to set the content type that the transcripts should be published in (page, story, etc.). For the project, each transcript was published as a node:page. The edit screen of the Migrate module was where one crosswalked the fields in the MySQL table to corresponding Drupal node destination fields. For example, the information in the "title" field in the MySQL table could be used to populate the node:Title field in the Drupal page. After filling out which table fields should migrate to node fields, the Migrate module could then migrate the content in the MySQL table to Drupal. After verifying that the content migrated successfully, the transcript pages were published.

Calais processing

The transcripts were now ready to be processed by the Calais module. This step in the project was a short one since the Calais module allowed for bulk processing of content types. The module required an API key, which one could apply for and receive fairly easily. Once the API key was added to the module settings, the settings were then adjusted in the Node settings menu. The Node Settings menu allowed for a range of controls for Calais processing. There was a global setting that applied for all content types, and a setting for each content type for those who wanted a more granular control for processing. For the project the Page settings were adjusted to apply all suggested terms on every update as well as to use the Calais global entity defaults. The Relevancy Threshold setting for Pages were set to zero, indicating to Calais to bring back the greatest amount of terms without regard to relevancy to the page content. While setting this option to zero would have been discouraged in other cases, the experimentation aspect of this project allowed for a greater tolerance of terms that may not be very relevant to the content displayed in the page.

The transcripts were ready to be batch processed. The Calais module contained a bulk processing feature which allowed for a relatively quick and automated process of a large number of pages. The bulk processing page settings repeated some of the settings in the node settings menu, such as term relevance. Again, we set this to zero for the project to return the greatest amount of terms per page. After setting the rest of the options in the bulk processing menu, the pages were processed. In a matter of minutes, the 60+ transcripts now contained a set of categories which were populated with terms found within those transcripts.

Display of Categories/Terms

The display of terms in each transcript was at the bottom of the page, lumped into a group of highlighted terms. This display was very difficult for one to browse as well as difficult for one to analyze the terms created by Calais. Therefore, we explored different ways to format the terms and their respective categories in a human readable format. The project ended up utilizing the Views module in Drupal. The Views module allowed one to create a custom display of information extracted from one's Drupal site. Through the Views module, we were able to create a block to list the terms that Calais extracted, grouped by the Calais vocabulary term groups. This block was then placed on the right side menu of the local Drupal installation for all the transcript pages.

While Views gave us a basic faceted list of terms, we experimented with other ways of displaying the metadata created by Calais. A site index was created using the Taxonomy VTN module. The first level of the site index consisted of the Calais vocabulary term groups in which the terms were grouped into. Another modification to the metadata display used the Taxonomy Hide module. Even though the Views block provided an easier to understand display of categories and terms, the list of terms at the bottom

of the page remained. Taxonomy Hide made it possible to hide most of the group of terms without affecting the display in the Views block.

Metadata Analysis

Challenges and limitations related to the nature of the content

A number of challenges were encountered around providing intellectual access to the resources in this collection. Some were related to the nature of oral history projects in general, while others may be specific to this project. Since OpenCalais is not designed to capture administrative or technical metadata, the focus of this analysis will be limited to the information content – the subject matter contained in the transcripts.

Both human and machine generated metadata were derived from transcripts once removed from the original audio visual format. Regardless of skill level or the methods employed human transcription will inevitably produce errors - some typographical and others related to interpretation of spoken content. However, errors which may have occurred during the transcription process are beyond the scope of this case.

The length of the interviews (1- 2 hours), the number of simultaneous interviewees (1-8 at a time), and the relatively unstructured, unedited, subjective conversational narrative that is characteristic of oral histories made it difficult for catalogers to quickly capture the aboutness of any given interview or interview segment. The only paths through the content were widely spaced open ended questions asked throughout each interview. Labor intensive analysis of either the videotape or the transcribed versions by project staff was not cost effective and did not always provide the desired results. Some interviews came with metadata recorded on handwritten cover sheets and some did not – the quality and quantity of which varied over time dependent largely on available staff. While this was helpful – it was not a reliable long term solution.

Metadata Structure

When comparing the original metadata produced by human catalogers using Dublin Core to that generated by OpenCalais there are significant gaps. As noted previously OpenCalais is only able to provide access points that represent subject content and cannot capture administrative (e.g. date and location of interview) or technical (e.g. format, extent, etc.) metadata. OpenCalais produces metadata that fall into the following high level categories: Entities, Events/Facts, Social Tags, and Document Categories. Though each of these includes more narrow facets (e.g. Political Events, Cities, etc.) they are not presented as a formal hierarchy.

Two of the most common vehicles used to represent information content are surrogate records or indexes (in various forms including faceted taxonomies). Records will typically present many characteristics of a given information object in addition to the subject matter. It does so within a specific metadata structure most often including subject headings (i.e. descriptors) based on human interpretation and analysis. Indexes are typically comprised of terms (descriptors) extracted from the text and left unaltered. These can be combined with additional representative terms selected from external sources. It's possible to generate indexes or surrogate records using human or machine based methods. This analysis compares descriptors generated by a machine based indexing method using semantic extraction to the descriptors (controlled subjects, name headings and keywords) selected by catalogers creating surrogate records.

Method of Analysis

This project was originally designed to explore practical solutions to problems commonly encountered when attempting to provide enhanced subject access to primary resource content. The following is an informal, unscientific analysis based on some initial observations.

The unit of analysis in this case is a single transcript representing two hours of videotaped interviews. Comparisons were made between information representations (metadata representing subject content) created by humans versus those generated by machine based methods. It should be noted that the referential sources varied somewhat between the machine generated and human created methods. OpenCalais had access to the transcripts and a proprietary authority file while human catalogers had access to transcripts, LC authority files, cover sheets, original video recordings and if need be the interviewer(s).

Evaluation Criteria

This analysis was based on selected characteristics of effective information representation not on the effectiveness of the resulting access points in information discovery. Results will be compared on the basis of the following criteria, defined very simply for the purpose of this project:

- Accuracy – the number of errors and the extent to which the terms generated correctly reflect the information content.

- Exhaustivity – the total number of unique terms generated.

Number of Unique Topical Descriptors by Category and Accuracy

OpenCalais Category	Number of unique topic descriptors generated	Correct	Incorrect	Accuracy Rate
Social Tags	9	6	3	67 %
Event Categories	6	6	0	100 %
Occupations	16	10	6	63 %
Document Categories	4	3	1	75 %
Total	35	25	10	71 %

Number of Unique Named Entities by Category and Accuracy

OpenCalais Category	Number of unique named entities generated	Correct	Incorrect	Accuracy Rate
Persons	37	34	3	92 %
Organizations	11	11	0	100 %
Cities	5	4	1	80 %
States or Provinces	2	2	0	100 %
Companies	3	1	2	33 %
Facilities	20	18	2	90 %
Holidays	1	0	1	0 %
Total	79	70	9	89 %

Human vs. Machine Generated Descriptors by Type

	Accuracy Rate		Exhaustivity*	
	Topical Terms	Named Entities	Topical Terms	Named Entities
Human	100%	100%	21	24
OpenCalais	71%	89%	25	70

* Only correct entries are counted

Topical Terms

While not surprising at this level of analysis – the accuracy rate for catalogers was 100% compared to rates starting at 63% for OpenCalais. Types of errors included errors parsing compound terms, errors in categorization (homographs), errors in disambiguation and errors related to relevance. Errors of omission were not counted for topics. The overall accuracy rate was 71%. For topical terms the difference in exhaustivity (measured as total number of correct terms generated) between OpenCalais and human generated results is not significant (25 v. 21). Topics generated by OpenCalais tended to very general compared to those chosen by catalogers (e.g. War, Politics, Education vs. Vietnam War Protests, Kent State Shootings).

Named Entities

As above the accuracy rate for catalogers was 100% compared to 89% for OpenCalais – much higher than the overall rate for topical terms. For OpenCalais there were no errors of omission for personal names - in fact, it correctly identified every person catalogers did and then went on to correctly identify 46 more. This is especially impressive given the local nature of most of the entities. In addition, OpenCalais did an outstanding job of correctly identifying facilities on campus (named buildings) using adjacency and contextual markers (terms paired with “hall”, “center”, etc.). Not so surprisingly it also did well with established locations (cities, states, etc.) since it could run these against an authority file. For named entities OpenCalais’s results were impressive in terms of both accuracy and exhaustivity.

However, exhaustivity also needs to be examined in terms of significance – how well does a particular term indicate the aboutness of the content? This was an issue which may be more critical for narrative than other kinds of content. Is there an inverse relationship between exhaustivity and significance? Does a given term/named entity merely represent something mentioned in passing or is it an indicator of more substantive content? Machine based methods can make estimates of overall aboutness at the document level as OpenCalais does based on term frequencies. How do we measure aboutness at more granular levels? Can such potentially subjective questions be answered by methods other than human judgment based on close review of the content? Is that something, on principle at least, that should be left to the user? These questions and others (e.g. preferences for browsing, most useful types of access points, etc.) which could lead to solutions that make collections of primary resources more accessible merit further investigation.

Conclusion

As with other technology-based projects, this project saw a fair amount of tweaking and troubleshooting. Since there was no other known Oral History project using OpenCalais in the manner that the project was using it, the project staff investigated the other uses of OpenCalais and then tried to adapt certain aspects into the project. The project did run into technical issues, like encoding and special character issues from the ContentDM export file that appeared when importing the file into Drupal. Most of the technical issues were not unique to the project, though.

Another aspect that we found in this project was that while Drupal provides various modules and tools for creating a locally customized alternative transcript viewing platform, building that platform could take a considerable amount of time and resources if one wanted to do so. There were no dedicated staff for this project, but because of Drupal's popularity and flexibility producing a variety of modules, a staff on a limited staff resource basis could produce a simple interface displaying the transcripts and metadata and organizing them on a basic level. Given more time and resources, the interface and functionality of the platform could be improved upon to provide more services and better organization and discovery. Nonetheless, resources in terms of staff, skills, and time should be considered when investigating potential Oral History platforms and automated metadata creation workflows. Drupal provides a more than capable framework to build a highly customized interface; however other more out-of -the-box platforms may be more appropriate depending on the needs of the intended users and the institution.

Using OpenCalais in a Drupal environment allowed for various customization and discoverability, which provides room for future expansion and functionality. The Calais Collection included different modules focusing on different aspects of discoverability. The Calais Marmoset module integrated the existing Calais Marmoset tool, creating specialized metadata for search engines. The More Like This and TopicHubs modules focused more on creating access by grouping related content or tags together in various ways. Other modules not under the Calais Collection could also make use of the data created by Calais, such as the GMap module taking geographic terms and plotting them on a map. More recently, the Calais Collection was included in OpenPublish, a Drupal distribution that was developed for publishing of a range of online content and has been used for various online news and other organizations. [1]

Another feature that the project did not delve into deeply but was one focus for future development was Linked Data. Calais Linked Data is part of the Linking Open Data Cloud, allowing for connecting entities identified by Calais to others in the Cloud, including Freebase and Wikipedia. [2] The Drupal versions 6 Calais module required the RDF module. While that provided some opportunities in possible

LD development, development in the Semantic Web/LD fields since the project began have shown great strides. One of the bigger strides was Drupal 7 incorporating more Semantic Web technologies and standards, including RDF, SKOS, and Dublin Core, into its core. Migrating the project to the newer Drupal version could provide more opportunities for Linked Data application in the future.

OpenCalais is ideal for handling large volumes of unstructured text. It's very effective at generating the kinds of access points (Named Entities: persons, places, events) assumed to be most useful for typical users of historical content and can do so with relatively high levels of accuracy. Although the results may still require some human review, it has the potential to save many hours of catalogers' time. OpenCalais can also be used as a stand-alone solution for generating index terms outside of a particular platform such as Drupal. The resulting terms can be entered as metadata in any system of your choosing. Among the disadvantages of this approach: it still requires transcripts (a very laborious undertaking); the topical terms generated tend to be very broad and the issue of significance as detailed previously remains.

[1]https://drupal.org/project/opencalais
[2]http://www.opencalais.com/documentation/linked-data-entities

Appendix A: Drupal specifications and modules list

- Drupal version 6.16
- Table Wizard 6.x-1.2
- Schema 6.x-1.7
- Views 6.x-2.10
- Migrate 6.x-1.0
    - Optional: Migrate Extras and Advanced help
- CCK 6.x.26
- Calais 6.x-34
- RDF 6.x-1.0-alpha7
    - ARC2 library needed as well, see RDF module page for more information
- If you want a site index
    - Taxonomy VTN 6.x-1.9
- If you want to hide the list of vocab terms at the bottom of the page
    - Taxonomy hide 6.x-1.02
        - Does not hide all of the terms at the bottom of the page – only applies to one vocabulary group (Social Tags) which can be turned off at the Calais global node setting and updating the page.

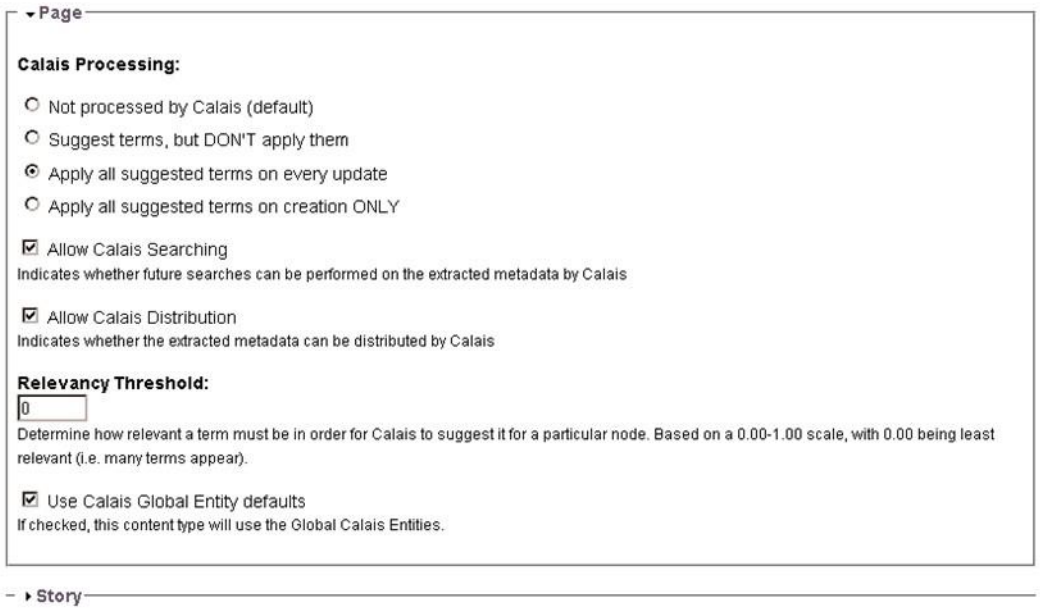

Figure 1. The page node settings of the Calais module in Drupal.

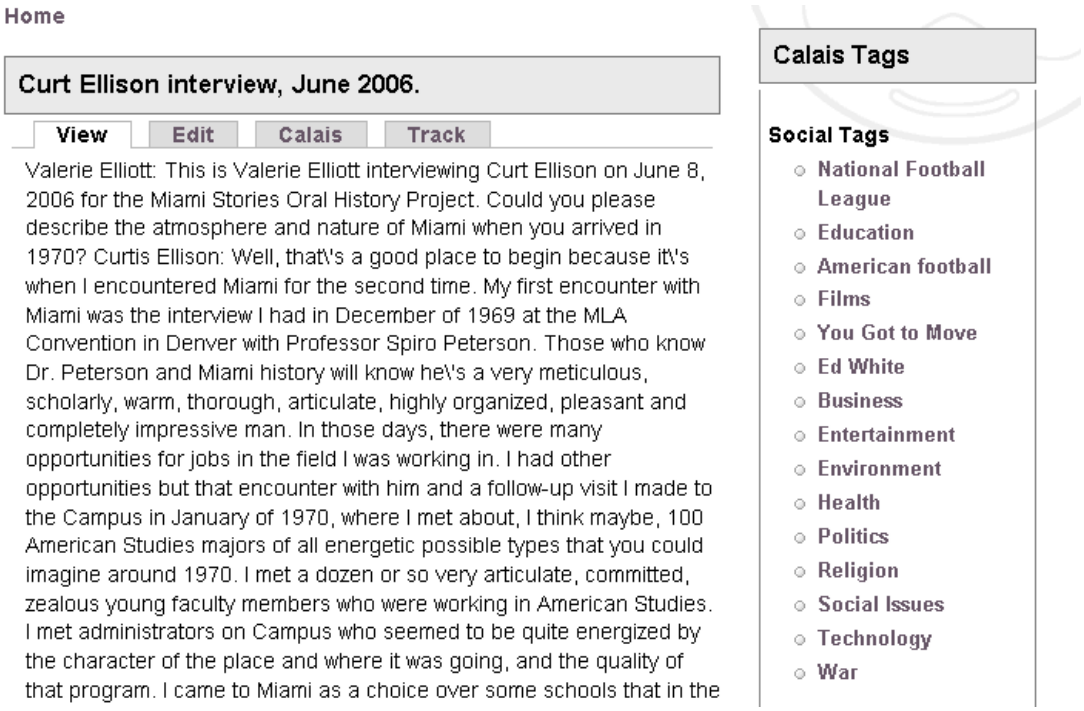

Figure 2. An example transcript page that was processed by the Calais module in Drupal.