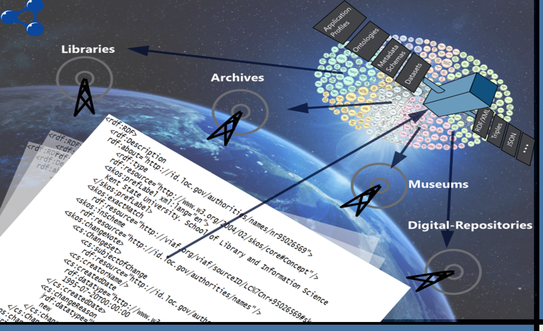

Helping Users Find the 'Good Stuff': Using the Semantic Analysis Method (SAM) Tool to Identify & Extract Potential Access Points from Archival Finding Aids

Karen F. Gracy, PhD; Sammy Davidson, MLIS, MS
School of Library and Information Science, Kent State University

ABSTRACT

The Semantic Analysis Method (SAM) Project aims to develop an open source tool for identifying and analyzing unstructured descriptions of archives and special collections materials to generate potential access points suitable for linked data applications. This presentation reports on the development of the SAM tool, which is a software application that utilizes the semantic analysis engine Open Calais to linguistically process archival finding aids and generate potential metadata (access points) through entity extraction. The entities derived from the analysis are parsed and saved in the comma-separated value (CSV) database format, and can then easily be imported into a data cleanup application such as OpenRefine. **This tool provides an important bridging application for assisting in converting valuable, unstructured information found in archival descriptions into usable, semantically-defined access points.**

ACKNOWLEDGEMENTS

The development and testing of the SAM Tool was completed as part of the Metadata Vocabulary Junction (MV-Junction) Project.

Principal Investigator: Marcia Lei Zeng
Co-Investigator: Karen F. Gracy

Funding for the MV-Junction Project was provided by the IMLS National Leadership Grant program.

CONTACT

Karen F. Gracy
School of Library and Information Science
Kent State University
Email: kgracy@kent.edu
Website: http://lod-lam.slis.kent.edu

BACKGROUND

Finding aids are rich descriptive tools that contain many significant details about the provenance and content of historical records, and may include dozens, hundreds, or even thousands of potential access points into the contents of a collection, including personal and family names, organizational and corporate names, events, geographic names, topical terms, and genre terms.

The current encoding standard used for markup of finding aids, Encoded Archival Description (EAD), allows archivists to assign semantic tags to these names for the purposes of indicating the creator(s) of the materials as well as significant topics documented in the records. Yet only a select few access points are labeled in this way due to the significant time and cost involved in manually marking up an entire document. **Archivists need a quick, easy, and inexpensive way to analyze archival records and tag new access points, so many more entry points can be created for indexing the records.**

HARNESSING THE POWER OF SEMANTIC ANALYSIS

As an aid in determining the number of entities in a text block, a natural language processing (NLP) tool was employed, called OpenCalais. This tool can analyze unstructured documents in text, HTML, or XML formats to identify named entities, generate known facts about those entities, and identify events associated with those entities. The analysis relies on the underlying knowledge base, i.e., embedded business rules**. This analysis results in a list of entities, including personal names, corporate body and organization names, geographic locations and features, and events, and also supplies semantic metadata that can be used for other applications**, such as news aggregators and blogs.

A number of libraries, archives, museums, and digital humanities projects have experimented with OpenCalais for the purposes of generating named entities and tags that might be used to enhance description of cultural heritage collections.

SAM Tool source code located at: https://github.com/sammysemantics/SAM

OVERVIEW OF THE SAM TOOL

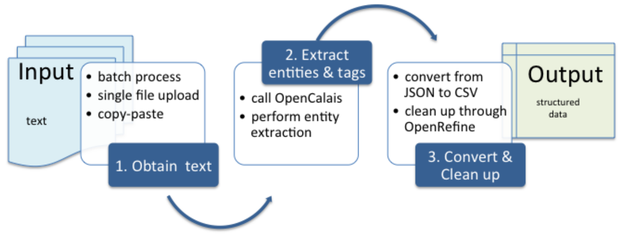

Figure 1. Overview of SAM Tool Functionality.

The Semantic Analysis Method (SAM) Tool automates identification and extraction of potential access points and parses the resulting data into a database for further cleanup and editing. **The SAM program integrates j-calais, a third-party library that provides a Java interface to the Open Calais semantic analysis API service, with additional scripts in Java to streamline the tasks of: (1) obtaining text files from a finding aid data repository; (2) calling the OpenCalais web service API; (3) performing the tasks of access point extraction and social tagging through the Open Calais service; (4) converting the resulting data to the CSV database format.**

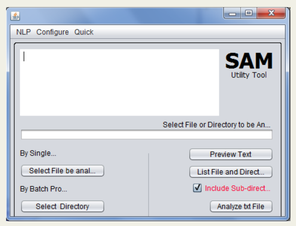

Figure 2. SAM Tool user interface.

Figure 3. Sample results of OpenCalais semantic analysis.

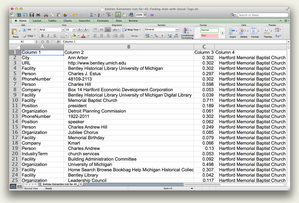

Figure 4. Example of extracted entities from finding aids.

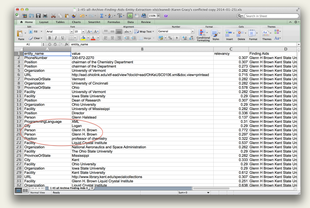

Figure 5. Example of cleanup activity in resulting database.

The database can then be imported into the Open Refine data cleanup tool to (1) improve the quality of the resulting datasets (merging synonyms into single data points and deleting incorrect extractions); and (2) establish linking between extracted entities and terms, and outside authority data sources.

TESTING THE SAM TOOL

Using a test set of 45 archival finding aids drawn from 16 repositories, this research measured the success of the tool in performing the initial entity extraction using OpenCalais API, and also identified the implementation challenges discovered during the import and cleanup of the resulting entities data in the OpenRefine environment. **Raw analysis of the 45 finding aids using the OpenCalais tool extracted 8,096 entities and 336 suggested social tags.** These figures were somewhat reduced by data cleanup to deduplicate, collapse synonyms into a single data point, and remove incorrect extractions.

Table 1 shows types of entity categories that OpenCalais includes in its ontology. Table 2 shows common errors found in results of semantic analysis that needed to be resolved through data cleanup using OpenRefine.

Table 1. OpenCalais Entity Types

Entity Groupings	OpenCalais Entity Types
Personal names	Person
Corporate body names	Company, Facility, Organization, Product (see also Topics)
Geographic names	City, Continent, Country, NaturalFeature, ProvinceOrState, Region
Publications (Titles)	MusicAlbum, Movie, PublishedMedium, RadioProgram, TVShow
Topics	IndustryTerm, Position, Product (see also corporate body names), Technology

Table 2. Errors Generated by the Semantic Analysis Process

Error Type	Examples
Entity duplication	New York (extracted and listed five times from same finding aid)
Entity variants	Margaret Muir, Margaret Muir Read (variants of personal name)
Entity miscategorization	Two Gentleman of Verona (title miscategorized as Movie, should be PublishedMedium) Sandy Hook, Virginia Key (geographic names miscategorized as Persons)
Inclusion of unrelated text as part of entity name	Box 9 Traveling Pictures Animation Company ("Box 9" not part of corporate body name)

CONCLUSIONS AND FUTURE DIRECTIONS

The SAM Project successfully achieved its goal to develop an open source tool to aid in extraction and identification of entities from unstructured textual descriptions of cultural heritage material. While the test data set of archival material descriptions indicated that the tool is in need of further assessment and refinement, the SAM tool provides a proof-of-concept for the automation of certain tasks related to entity extraction and refinement. This tool holds promise to assist catalogers with the enhancement of current surrogates beyond manually assigned controlled vocabulary terms.