Semantic Analysis Method (SAM): A Tool for Identifying Potential Access Points in Unstructured Text

Karen F. Gracy, Sammy Davidson, Marcia Lei Zeng
School of Library and Information Science
Kent State University
kgracy@kent.edu, sdavids6@kent.edu, mzeng@kent.edu

Widespread adoption of linked data technologies will facilitate the interconnection of library, archive, and museum (LAM) information systems, enrich systems with data drawn from external trusted information sources, and allow users to search for and use materials and information across a variety of datasets. In order to achieve this goal for next-generation cultural heritage information systems, data must be structured in a way that such interlinking is possible. Convergence among heterogeneous datasets is difficult, particularly in the cultural heritage sector, due to problems of converting legacy data into linked data and aligning the vocabularies used by different datasets.

Legacy descriptions that contain large amounts of unstructured data remain one of the most vexing stumbling blocks for achieving the goal of launching semantically-enabled information systems for cultural heritage materials, such as archives collections. Whereas linked data applications rely on the use of Uniform Resource Identifiers [URIs] to define each significant entity and topic in a dataset, legacy descriptions such as archival finding aids are characterized by: (1) large blocks of unstructured narrative prose and (2) detailed descriptions of content at different hierarchical levels of arrangement (such as record group, series, subseries, file, and item), which often reflect the intellectual and/or physical organization of the materials).

While controlled vocabulary terms are often included as part of finding aids and collection-level catalog records, those terms represent a small percentage of the larger number of potential entities and topics that could conceivably be identified and converted to structured data suitable for interlinking. The number and variety of controlled vocabulary terms assigned by catalogers is not consistent, and term assignment tends to be limited by resources and institutional cataloging and description. Given these practical restrictions that keep archivists from providing additional entry points into archival collections, the research team at Kent State University suggest that a machine-based solution may help catalogers increase the number of entities and topics available for interlinking to other databases. More specifically, we suggest harnessing the power of semantic analysis technology to aid in the identification and extraction of entities and topics for further processing and conversion to linked data.

This presentation describes a tool developed to help solve the challenge of converting unstructured textual descriptions of cultural heritage material, specifically archival descriptive data, into linked data. The Semantic Analysis Method (SAM) tool serves as a bridge between unstructured narrative description and semantically-defined access points. SAM identifies name entities and topics through the use of a semantic analysis engine, OpenCalais, with a JSON data file as an initial output, and also parses and saves the results of the OpenCalais analysis as a comma-separated value (CSV) database file. Once in a database form, this list of potential access points may be imported into a data cleanup application such as OpenRefine for further editing and removal of any misidentified entities.

Figure 1 below provides an overview of this process:

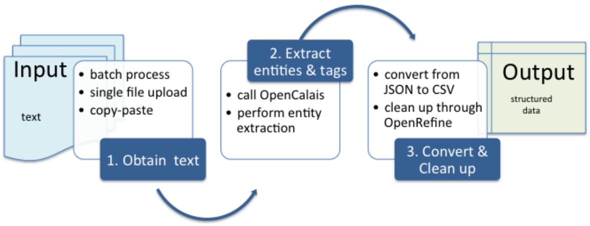

Figure 1. Overview of SAM Tool Functionality

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 clean-up of the resulting entities data in the Open Refine environment.

Raw analysis of the 45 finding aids using the OpenCalais tool extracted 8,096 entities and 336 suggested social tags. These figures would be somewhat reduced by data clean-up to deduplicate (when the same entity is mentioned more than once in a document), collapse synonyms into a single data points, and remove incorrect extractions. While the average number of data points extracted for each entity type varied significantly, depending on the content and extent of the collection being described, it is worth mentioning that in most cases the number of entities extracted often exceeded the number of controlled access points assigned to the archival collection by the cataloger who created the finding aid.

The research team then explored data generated from the OpenCalais analysis in the OpenRefine tool in order to assess the success of the initial entity extraction, and to identify the types of problems that would need to be remedied to make this data suitable for interlinking. The OpenCalais tool had successfully extracted recognizable entities from the raw text drawn from finding aids on most occasions; however, those entities could sometimes be miscategorized. The categories for which OpenCalais most successfully identified entities were personal, corporate body, and geographic names, although synonym control was still needed to deduplicate and control for name variants.

The OpenRefine tool can assist in resolving duplication and variant errors, but does not help in resolving problems of entity miscategorization or incorrect extractions. Additionally, OpenRefine may also introduce new errors into the extracted data due to the algorithms used to suggest refinements; thus, great care must be taken in the use of this tool to improve the quality of data.

While the SAM Tool successfully provides the functionalities of text retrieval, semantic analysis for entity extraction, and conversion of results to a more manageable format for later data clean-up, much work remains to streamline these processes and improve accuracy in data extraction and characterization. There are certain challenges to be overcome in the areas of entity extraction and name resolution for historical names and places (where those names may not be available in authority files), and misidentification of certain phrases as entities. Due to this issue of not finding many names in well-established national and international data sources, it is clear that establishing a local name authority needs to be another task to be incorporated into the SAM tool in order to improve accuracy in identifying and correctly categorizing entities.

In future iterations of the SAM tool, the research team will be exploring the addition of new functionality such as:

- Generating RDF instead of JSON during the OpenCalais analysis and extraction.
- Incorporating the option to choose other semantic analysis engines.
- Adding the functionality to directly query linked data sets such as LCNAF, DBpedia, and VIAF to find proposed matches to established access points.
- Further modularization of the various processes and procedures to make future updates easier.

The research team also hopes to identify and test additional cultural heritage data sources, such as other types of descriptions and transcriptions of interviews and oral histories, to determine the efficacy of the SAM tool in increasing potential access points and facilitating interlinking of individual documents to relevant information found in other data sources.