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

SAMMY DAVIDSON AND KAREN F. GRACY

Abstract: 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, thus giving users many more entry points into records.

The Semantic Analysis Method (SAM) Tool automates identification and extraction of potential access points and parses the resulting data into a database for further clean-up and editing. The SAM program integrates j-calais, a third-party library that provides a Java interface, to the Open Calais semantic analysis API web 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 API; (3) performing the tasks of access point extraction and social tagging through the Open Calais service; and, (4) converting the resulting data to the CSV database format. This database can then be imported into the OpenRefine data clean-up tool to (1) improve the quality of the resulting datasets (included merging synonyms into single data points and deleting incorrect extractions); and (2) establish linking between the extracted entities and terms and outside authority data sources. The SAM tool serves as the first tool in a larger toolkit (currently in development) to enrich and enhance finding aids with semantic markup.

About the authors:

Sammy Davidson has his B.A. in Sociology from Baldwin-Wallace College. He graduated with his MLIS in August of 2013 and also expecting to graduate with his MS in Information Architecture and Knowledge Management from Kent State University at the end of this summer.

Karen F. Gracy is an associate professor at the School of Library and Information Science of Kent State University. She possesses an MLIS and PhD in Library and Information Science from the University of California, Los Angeles and an MA in critical studies of Film and Television from UCLA. Recent publications have appeared in *JASIST*, *Archival Science*, *The American Archivist*, *Journal of Library Metadata*, and*Information and Culture*. Dr. Gracy's scholarly interests are found within the domain of cultural heritage stewardship, which encompasses a broad range of activities such as preservation and conservation processes and practices, digital curation activities that consider the roles of heritage professionals and users in the lifecycle of objects and records, as well as knowledge representation activities such as definitions of knowledge domains, development of standards for description, and application of new technologies to improve access to cultural heritage objects.