**GRACY — Creating**

<0 images>

**Creating Linked Data within Archival Description: Tools for Extracting, Validating, and Encoding Access Points for Finding Aids**

Gracy, K. F. and Zeng, M. L.

In current archival information systems, useful linkable information is embedded in narrative descriptions known as finding aids. While the finding aid excels at providing contextual information to understand the nature and scope of archival collections, the format of the genre is characterized by large text blocks in which many types of information are intermingled and unlabeled. In the blocks of histories of records and the scope notes for collections, dozens or even hundreds of names of persons, organizations, places, and events, as well as topical terms can be found. While these text blocks in finding aids are full-text-searchable and amenable to natural language processing techniques, the lack of semantic distinction among the different entities and topics hinders efficient and effective information retrieval and restricts the ability of information systems to create the links that would gather widely dispersed information about the same person, organization, or thing into one place.

Because of these challenges of converting archival descriptions to archival linked data, the Linked Open Data—Libraries, Archives and Museums (LOD-LAM) research group at the School of Library and Information Science, Kent State University (http://lod-lam.slis.kent.edu/index.html) has been exploring various automated and semiautomated ways to enrich archival description with semantic tagging. This process involves (1) the identification of name entities and topical terms from finding aids, (2) extraction of these entities and terms and processing them using semantic analysis services, (3) validating the names of each extracted identity, and then (4) encoding the entities and topics that can be validated within the finding aids with Uniform Resource Identifiers (URIs).

In a pilot study, the LOD-LAM research group developed a software program that facilitated the first and second steps mentioned above of this multistep process, as reported in the NKOS Workshop in 2014 DL Conference (Gracy et al., 2014).The Semantic Analysis Method (SAM) Tool first obtains the archival descriptions by one of three methods: copying and pasting text from a finding aid document, upload of an individual PDF file, or batch upload of multiple PDF files. Then, the SAM Tool sends the file or files to a semantic analysis service, such as OpenCalais (http://viewer.opencalais.com/). These services generate semantically tagged output in the JSON format, which the SAM Tool then converts to a CSV file. The resulting CSV database can then be viewed as a Microsoft Excel spreadsheet. The CSV files can be used in the OpenRefine tool (openrefine.org) to validate the names and topical terms against various controlled vocabularies, such as the Library of Congress Name Authority File, the Library of Congress Subject Headings, and the Getty vocabularies. As a continuing study, additional functionality that the research group will be exploring in the next year will be automated methods to embed those validated entities and topics in finding aids, connect those entities to the same entities in other data sources, and enhance finding aids with information found in those other sources.

Similar to OpenCalais, other semantic analysis tools are available, with free or demo versions that allow the performance of the first few tasks (the identification of name entities and topical terms from finding aids, extraction of these entities and terms, processing them using semantic analysis services, validating the names of each extracted identity). The APIs can be tested (after obtaining the keys) as well. Some of the tools have more functions, such as semantic reasoning, categorization, and fact mining, in addition to text mining and entity name extractions. Visualization of the relationships between entities and on geographic maps are other functions of these tools. In this poster, research team members Zeng and Gracy will present the results of experiments to test the performance of several semantic analysis engines in identifying and extracting entities and topical terms, including OpenCalais, MachineLinking (http://www.machinelinking.com/wp/), Cogito Intelligence (http://www.intelligenceapi.com/), and Zemanta (http://www.zemanta.com/api/). The research findings include the comparison of functions, the core functions identified through the study, and the estimated values added to extracting, validating, and encoding the access points for finding aids.

**Reference**

**Gracy, K. F., Davidson, S. and Zeng, M. L.** (2014).Semantic Analysis Method (SAM): A Tool for Identifying Potential Access Points in Unstructured Text. Paper presented at the *13th European Networked Knowledge Organization Systems (NKOS) Workshop, Digital Libraries 2014 Conference*, London, 11–12 September 2014.