Archival Description and Linked Data: A Preliminary Study of Opportunities and
Implementation Challenges

KAREN F. GRACY

Abstract: This poster presents selected results of a study to investigate how archives can connect their collections to related data sources through the use of Semantic Web technologies, specifically linked data. Questions explored included: a) What types of data currently available in archival surrogates such as Encoded Archival Description (EAD) finding aids and Machine Readable Cataloging (MARC) records may be useful if converted to linked data? b) For those potentially useful data points identified in archival surrogates, how might one align data structures found in those surrogates to the data structures of other relevant internal or external information sources? c) What features of current standards and data structures present impediments or challenges that must be overcome in order to achieve interoperability among disparate data sources?

To answer these questions, the researcher identified metadata elements of potential use as linked data in archival surrogates, as well as metadata element sets and vocabularies of data sets that could serve as pathways to relevant external data sources. Data sets chosen for the study included DBpedia and schema.org; metadata element sets examined included Friend of a Friend (FOAF), GeoNames, and Linking Open Description of Events (LODE). The researcher then aligned tags found in the EAD encoding standard to related classes and properties found in these linked data sources and metadata element sets. To investigate the third question about impediments to incorporating linked data in archival descriptions, the researcher analyzed the locations and frequencies at which controlled and uncontrolled access points (personal and family name, corporate name, geographic name, and genre/form entities) appeared in a sample of MARC and EAD archival descriptive records by using a combination of hand counts and the natural language processing (NLP) tool, OpenCalais. The results of the location and frequency analysis, combined with the results of the alignment process, helped the researcher identify several critical challenges currently impeding interoperability among archival information systems and relevant linked data sources, including differences in granularity between archival and other data source vocabularies, and inadequacies of current encoding standards to support semantic tagging of potential access points embedded in free text areas of archival surrogates.

About the author:

Karen F. Gracy is Assistant Professor at the School of Library and Information Science, Kent State University. She teaches in the areas of digital preservation, digital curation, and archival studies, with a particular interest in media archives and knowledge representation for cultural heritage resources. Her research interests include preservation and curation of digital media, audiovisual archiving, linked data applications for cultural heritage material, and the social contexts of information creation and use, focusing on ethics and values. Her most recent publications include "Ambition and Ambivalence: A Study of Professional Attitudes Toward Digital Distribution of Archival Moving Images" for *The American Archivist* and "Exploring Methods To Improve Access to Music Resources by Aligning Library Data With Linked Data: A Report of Methodologies and Preliminary Findings," which will appear later this year in the *Journal of the American Society of Information Science & Technology*.