TEMPLE
UNIVERSITY LIBRARIES

Organization: Temple University Libraries

Project title: Automating keyword assignments for the Nineteenth-Century Knowledge Project

Primary mentor	Peter Logan, Academic Director, Digital Scholarship Center, Temple University Libraries and Professor of English
Supporting mentor	Holly Tomren, Head of Metadata Strategy and Digitization Services, Temple University Libraries
Description	The Nineteenth-Century Knowledge Project is an NEH-funded effort to digitize and index individual entries from historical editions of the Encyclopedia Britannica. All material is generated as TEI-XML files. We want to add LOD subject terms to each entry using the HIVE automated indexing program. We will compare the results of using current and nineteenth-century comprehensive vocabularies, to see whether a historical ontology produces significantly different outcomes when indexing historical material.
Problems	(1) Historical vocabulary will need to be cleaned up from initial OCR output. (2) Getting an historical ontology into a SKOS format involves complex cross-walking between historical and current terms. (3) Comparing differences in keyword outputs for two vocabularies over thousands of entries will require advanced analytical techniques.
Techniques	Creating a SKOS-enabled vocabulary using a thesauri program, like MultiTes; work closely with TEI-XML; work with HIVE online; interact with eXist database online.
Data	TEI-XML source files; OCR output for historical vocabulary; plain-text source files; csv files.
Outcome	The larger project seeks to advance our understanding of how knowledge changes over time by developing a comprehensive dataset of official knowledge from the French Revolution to WWI. The section implemented by the LEADS fellow will be instrumental in understanding how changes in controlled vocabularies over time relate to the analysis of knowledge change generally as well as provide the first test case measuring the value of using historical instead of contemporary controlled vocabularies when indexing older textual materials.