Automatic Metadata Extraction for Archival Description and Access

WILLIAM UNDERWOOD

Abstract: The objective of the research reported here is to develop techniques for automatically extracting metadata from electronic records that is necessary for automatically describing items, file units and records series and for supporting access to these records. Archival metadata and elements of descriptions include document type, date, author, addressee, and topic. The elements of documentary form are those elements of documents of the same type that do not change, or vary just slightly, from document to document. These include not just keywords or captions such as "MEMORANDUM FOR", "FROM: ", and "SUBJECT:", but semantic categories such as dates and person names, for example, "MEMORANDUM FOR <person name>". The methods developed are described via example. These include methods for: (1) annotating dates, persons names, location names, organization names, postal addresses, and other semantically relevant categories that appear in e-records, (2) recognizing the intellectual and physical elements of documentary form, (3) recognizing the documentary form of a record and extracting metadata by using a parser with grammars for documentary forms, (4) automatically generating item titles and scope and content notes from the metadata, and (5) automatically populating access point attributes such as personal name, geographic name and topics. It is illustrated how these results potentially support improved access to electronic records.

About the author:

William Underwood is the Principal Research Scientist at Georgia Tech Research Institute. Dr. Underwood is the Principal Investigator of the PERPOS project sponsored by the U.S. National Archives and Records Administration. The objective of this project is to aid archivists in Presidential Libraries in preserving, reviewing and describing digital records created on personal computers. He is also a member of the U.S. InterPARES II research project sponsored by the National Historical Publications and Records Commission. The objective of that project is to identify technologies and methodologies for long-term preservation of authentic electronic objects created during the course of scientific, artistic and government business activities. He is a member of the Consultative Committee for Space Data Systems, Panel 2 that is developing standards for archival information interchange. His current research interests are in developing formal, theoretical foundations for records management and archival science, experimental investigations of alternative preservation strategies, and the application of natural language processing technologies to the support of archival description and Freedom of Information Act (FOIA) review.