Testing EAD Encoding in the *Texas Archival Resources Online* (*TARO*) System with Textual Analysis Techniques

Vidya Narayan *(narayanv@ischool.utexas.edu)*

School of Information, University of Texas at Austin

Patricia Galloway *(galloway@ischool.utexas.edu)*

School of Information, University of Texas at Austin

Electronic archival finding aids encoded in Encoded Archival Description (EAD) are transported across networks and rendered into HTML for display on the browser. Considering the time, effort and money involved in marking up the finding aids, has the markup been used for retrieval purposes? Has the multilevel hierarchical nature of finding aids been used for searching? A few online EAD tag based retrieval systems that process queries look for occurrences of the search term in the corresponding EAD tag, but do not seem to address subject- or topic-based queries. This study explores the possibility of using the content of specific EAD tags for subject retrieval purposes. We studied the consistencies, commonalities and discrepancies in usages of various critical tags across repositories participating in the *Texas Archival Resources Online (TARO)* project. These usages were compared to EAD tagging guidelines as well as TARO guidelines. We identified the `<abstract>`, `<scopecontent>` and `<bioghist>` tags as good representatives of the finding aid from standard archival descriptive practice and examined their content for a sample of repositories within TARO. The content of these tags was processed using text processing techniques to further study and arrive at possible similarity metrics to identify similar finding aids. We feel this would help evaluate EAD as an information retrieval tool within *TARO* and if our experiments help conclude that EAD can be effective as such a tool (or can be made effective by better descriptive practice), then the prospect of creating a highly interconnected web of finding aids exploiting the hierarchical nature of EAD is possible.

This study was conducted on 1226 EAD encoded finding aids from nine archiving institutions which are part of *TARO*. Our study was conducted in three phases. First, we verified the usage of EAD tags across repositories within *TARO* with an aim of determining if there exists a core set of tags within these finding aids. This part of the study was motivated by the underutilization of the Dublin Core tags as reported by Shreves, Kirkham, Kaczmarek and Cole and Ward. From this part of our study we arrived at a core set of 27 EAD tags from the entire EAD tag library comprising 146 tags. These 27 EAD tags form a superset of the tags deemed mandatory by ISAD(G) as well as the EAD tagging guidelines of other archiving institutions. Additionally, we observed the varied usage of the hierarchy of these tags within these finding aids and very limited usage of tags to achieve electronic linking between documents.

In the second part, we studied if these finding aids have been encoded according to standard archival descriptive practice (i.e. if the text within these EAD tags was appropriate). This was achieved through text processing involving extraction of the text from the specific tags and processing these to arrive at a vocabulary. We conducted this study on the part of the finding aids corresponding to the *University of Texas Alexander Architectural Archive (UTAAA)* and *University of Texas Benson Latin American Collection (UTLAC)* repositories. Comparisons, of the vocabularies of the `<abstract>` tag between two different repositories indicate that the vocabularies for the said repositories are quite distinct. We found that the content of different tags has different word counts and correspondingly different vocabulary sizes. Additionally, we observed that up to 65% of the total word count in each of the three tags studied (`<abstract>`, `<bioghist>` and `<scopecontent>`) represents the vocabulary, thus indicating that significant information is embedded in the textual content of each of these tags.

In the third step, using the vocabularies obtained, we represented these finding aids as vectors in the vocabulary space. In such a vector representation of finding aids, we compared finding aids using a cosine similarity in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) weighting. The TF-IDF scheme weights rarely used words higher than commonly used words, and also accounts for the size of the document. We then clustered these finding aids with an online clustering tool (wCLUTO) using the agglomerative clustering algorithm. The agglomerative clustering groups finding aids based on the similarity of content, resulting in a tree of documents. The lowest levels of the tree correspond to individual finding aids and the highest levels of the tree correspond to the entire sets of finding aids. Our study focused on low-level clusters, which are of particular interest to archivists, as these clusters address the descriptive material embedded in the various EAD tags. To determine the similarities between finding aids, we extracted vocabularies for individual tags like `<abstract>`, `<scopecontent>` and `<bioghist>` and clustered the finding aids based on the similarity of textual content with respect to these individual tags. Further, we combined the similarity relations between finding aids, based on these individual tags, to build a space that encompasses the content similarity for a combination of tags. Our clustering results on individual and combination of tags are in agreement with the classification provided by the curators of the *UTAAA* repository.

We conclude from our study that if finding aids are marked up according to standard archival descriptive practice then they yield meaningful content-based clusters of similar finding aids. Further, we were able to demonstrate; i) the ability of forming 'neighborhoods' of similar finding aids using either individual tags or a combination of tags, and, ii) that the 'neighborhoods' were different for different tags or combination of tags. From this idea of a 'neighborhood' of finding aids, we propose a searchable interface for a repository of finding aids by means of the EAD tags. This search facility, we think, enhances the prospect of creating a web of similar and, thus, interconnected finding aids, which, in turn would facilitate research in the field of archives and help researchers form cliques by common research interests and goals.

Our study demonstrates the ability to apply the text processing techniques from the field of information retrieval to the field of archives with a goal of enabling EAD encoded finding aids transition to the digital world and be visible in the realm of online documents and be accessible to researchers.

Bibliography

Shreves, S.L., C. Kirkham, J. Kaczmarek, and T.W. Cole. "Utility of an OAI Service Provider Search Portal." *Proceedings 2003 Joint Conference on Digital Libraries.* Los Alimotos, CA: IEEE Computer Society, 2003. 306-308.

Ward, J. "A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative." *Proceedings 2003 Joint Conference on Digital Libraries.* Los Alimotos, CA: IEEE Computer Society, 2003. 315-317.