# Building Descriptive Metadata Framework and Taxonomy to Organize Topic-Specific Collection: Text Mining for No Gun Ri Archives

**DONGHEE SINN**

**Abstract:** The web environment and newer technology have provided users greater opportunities to access information, regardless of type and format. Recent studies identify that users consider it useful to have one place to search different types and levels of information. Having multiple types of datasets and various formats of information in one system creates challenges in organizing and describing. Current metadata schemes, such as MARC, EAD, Dublin Core, etc., have been developed to generally organize datasets that are loosely homogeneous in nature. Therefore, utilizing a particular metadata standard may not be effective for an aggregated collection with various types and formats of materials.

This study intends to find a way to organize heterogeneous collections under one system. This study uses an aggregated collection on a historical event, the No Gun Ri killing during the Korean War. The materials about No Gun Ri include various types of textual information (archival documents, oral histories and transcripts, government reports, academic publications, legal documents, etc) as well as other formats (photos, multimedia, web resources, etc.). This study aims (1) to identify data categories for descriptive metadata framework and (2) to create taxonomy of semantic contents of the collection. To do so, this study uses the actual contents of the collection as a core dataset to extract necessary categories and a knowledge base on the topic. Text mining is used to identify and explore representative descriptors from the unstructured text in the collection. A simple text mining tool, TAPor, is used to generate top descriptors from the collection.

Archivists cannot overlook the challenges associated with organizing diversified collections effectively in the digital realm. This study provides an example to handle a collection within a specific topic that has various natures in types and formats.

**About the author:**

***Donghee Sinn*** is an assistant professor of the Department of Information Studies, University at Albany (State University of New York). Her research interests are based on her experiences in archives and libraries. The major research interests lie with the archival research in relation to public memory, archival use/user studies, archival memory in the digital realm, personal archiving in the web environment, topic-focused approach in information organization, and archival education with practical components. One of the approaches in her research centers on the use of archival holdings and the actual influence of archival materials in the entity of knowledge in a certain area and even more specifically for a certain historical event. She has observed the impacts of archival use on the final products of historical research on a focused topic in her research studies. She would like to

further this understanding to see what the actual contents in historical research tell about necessary elements to organize such resources.

She holds a Ph. D. from the University of Pittsburgh. She previously worked as an archivist at the National Archives of Korea as well as a librarian in academic libraries. In her current position, she teaches Archival Representation, Information and Knowledge Organization, and Collection Development and Management.