Extracting Metadata from Digital Records Using Computational Methods

KATE BARBERA and ANN MARIE MESCO

Abstract: Metadata is essential for access, management, utilization, and discoverability of digital collections, but many repositories do not have the necessary resources to describe objects effectively at the item level. At Carnegie Mellon University, we are examining various methods for extracting item-level metadata from digital records automatically using combined computational methods. This is an interdisciplinary and interdepartmental research project that is focused on providing automated metadata solutions for archivists, digital humanists, and other similar professionals.

In 2016, we began our research in this area while evaluating and assessing our digital object metadata as part of a larger repository migration. We have over 3 million pages of items in our digital collections that have inaccurate and inconsistent metadata. Our goal is to use computational methods and tools such as optical character recognition, regular expressions, topic modeling, and natural language processing to build an efficient workflow, which will allow us to automate portions of the revision and enrichment processes for our digital object metadata. The workflow we are developing focuses on creating "templates" for loose categories of objects in our digital collections, based on the genre, form, and other characteristics of the items.

We are currently testing our methodology on one of our smaller digitized collections—the William W. Cooper Collection—consisting of 2,885 objects. Ultimately, we hope to share our workflow, code, and any other resulting tools and technologies with the wider community via our GitHub repository. Our aim is to produce an approachable, practical, and adaptable method for automating metadata creation.

About the Authors:

Kate Barbera is an archivist and public historian. Her work focuses on blending these two fields to enhance archival access and outreach, always aiming to help researchers find the stories that mean the most to them. Kate is currently Assistant Archivist at Carnegie Mellon University where she supports the activities of the University Archives in a broad range of areas, from oral history to metadata. Over the years, Kate has helped organizations large and small to share and preserve collections on topics as varied as avant-garde film and accounting. She holds a master's degree in Public History from Duquesne University and she is a certified Digital Archives Specialist. Kate is also a proud member of Three Rivers Archivists, a Pittsburgh-based professional advocacy group.

Ann Marie Mesco is the Digitization Projects Manager at the Carnegie Mellon University Libraries. For nearly 20 years, she has maintained digitized collections of over 3 million images and managed the archival repository. Over the past several years, her position has transformed from simply overseeing digitization projects to managing the curation of the University Libraries' digitized collections, which vary from the correspondence of Andrew Carnegie to the congressional records of Senator H. John Heinz III. She received her MLIS from the University

of Pittsburgh, a BA in History and Policy from Carnegie Mellon University, an AS in Business Management from the Community College of Allegheny County, and a post-master's certificate in the Curation and Management of Digital Assets from the University of Maryland.