Natural Language Processing for Archival Description of Records: Potential and Priorities

Alexandra Chassanoff, Christopher Lee, Sunitha Misra, and Kam Woods
School of Information and Library Science, University of North Carolina Chapel Hill

INTRODUCTION

Libraries, archives, and museum (LAM) professionals are tasked with characterizing the content of disk images (or sets of files extracted from disk images).

Archival description and access are significantly driven by the ability to identify specific contextual entities (e.g. people, places, organizations, events). The identification and presentation of access points based on those entities has traditionally involved significant human effort. Text extraction from a wide range of formats and sources can be challenging, costly, and time-consuming. At the same time, end-user researchers want to be able to quickly locate specific entities of interest and find correlations among materials.

The use of natural language processing (NLP) to identify and extract such information is one promising approach under investigation. Applying NLP techniques has the potential to free up the time of archivists so they can focus more on processing tasks that require their judgement, as well as to expose access points that may never have been generated by an archivist.

Two existing systems -- the BitCurator environment[1] and BitCurator Access[2] WebTools -- can serve as a foundation for facilitating access to born-digital archival materials.

QUESTIONS FOR DISCUSSION

Points of Access

- **What are the most important contextual entities (e.g., people, places, organizations, events) to serve as access points for archival description and access?**
- **Why would you consider them to be the most important?**

Entities of Interest

- **What kinds of reporting will be most helpful for archivists and end users interested in meaningful characterization of the contents of files?**

POTENTIAL NLP ELEMENTS TO ENHANCE ACCESS AND USE

TEXT EXTRACTION

e.g., textrac, tika-python

CORPUS CONSTRUCTION & TEXT ANALYSIS

e.g., NLTK28, TextBlob29, python-ucto30

MACHINE-LEARNING TOOLS

e.g., Scikit-learn + spaCy

FUTURE WORK

The application of NLP tools and methods to born-digital materials has the potential to greatly enhance access services for "traditional" LAM users (such as genealogists, journalists, and historians) as well as for new audiences. The ability to identify and expose contextual information in large sets of heterogeneous files, including complex objects like disk images, would be beneficial for LAM staff (to carry out appraisal, description, reference services, discovery, and review for sensitivity) as well as end users who might care most about locating specific entities of interest. This work also offers one example of how LAMS can support the use cases that drive the Digital Humanities.

REFERENCES

1 The BitCurator Environment
http://wiki.bitcurator.net

2 The BitCurator Access Webtools
http://access.bitcurator.net

ACKNOWLEDGEMENTS

The BitCurator and BitCurator Access projects have been funded through the Andrew W. Mellon Foundation.