Adam in the Archive: Discovering Named Entities in the Cold War Papers

Alexis Antracoli, Will Clements, and Cliff Wulfman

Society of American Archivists Research Forum, August 10, 2022
Background and Motivation

- NHPRC digitization grant, 2013 - 2015
- 350,000 pages of most typewritten documents now OCR’ed
- Began with the correspondence series in the Kennan papers.
Background and Motivation

Team wanted to determine:
○ Could using NER and dirty OCR produce useful information?
○ Can the use of automation conform with minimal processing approaches?
○ Can NER and OCR reveal named entities that are obscured by current description, especially individuals from marginalized communities?
Semantic Model

oa:hasTarget

annotation

oa:hasBody

canva

carries

inscription

depicts

appellation

identifies

person

has symbolic content

"Geo Kemon"

skos:prefLabel

Kennan, George Frost, 1904-2005

owl:sameAs

<http://viaf.org/viaf/66477608>
Technical Workflow

1. Finding Aid
2. Archival Collection
3. Machine-Readable Text
4. Local Authority KB
5. Named Entities
6. Names

Steps:
- Finding Aid → Archival Collection
- Archival Collection → Machine-Readable Text
- Machine-Readable Text → Named Entities
- Named Entities → Names
- Local Authority KB → Finding Aid

Symbols:
- IIIF
- OCR
- NER
- NEL
Discoveries

- Process finds names obscured by description that can be used to enhance the finding aid.
- People uncovered by NER usually not from marginalized communities.
- Dirty OCR results in a lot of incomplete names that require human review.
- Depending on how tightly focused a collection is, there may be a lot of “noise,” i.e. people who are not pertinent to the topics covered in the collection.
Next Steps

- Determine next corpus of documents to perform NER on.
- Assess whether different types of records and collections provide different or similar results.
- Enhance finding aids with names by creating Agent Records in ArchivesSpace, attaching to components within finding aids.

Icon by Afafat Uddin from Noun Project
Thank You!

Project Github:
https://github.com/pulibrary/finding_aid_enrichment