On Leveraging AI and NLP to Create an Open Source Workflow for the Rapid Creation of Archival Linked Data for Digital Collections

Jennifer Proctor
PhD Student, University of Maryland College Park
AI Collaboratory
ARLIS
BENEFITS OF LINKED DATA

Visibility for Cultural Value

These images depict amazing things but get lower use than they should because they're hard to find and information necessary for some uses, like journalism (names and permission of those depicted) aren't known.

- Connecting LAM images to larger information networks can help more people find those images
- It can also help identify lost information about the content, like who is depicted, and fill in details, like what they went on to do in life

Information & Systems Create Connections
FIND THE SCRIPTS

https://github.com/JeProctor/Metadata_to_LinkedData

DEPENDENCIES
Python, OpenCV, Tensorflow, NLTK, Selenium, RDFlib
PROCESSES

IMAGE SEGMENTATION
OPENCV

METADATA TO TRIPLES
REGEX, RDFLIB

CONNECTION AND DISAMBIGUATION
OPENREFINE

AUTOMATED BOXING AND CROPPING
OPENCV

URI QUERY AND CREATION
HIVE, RDFLIB
Testing the algorithms was important to ensure its ethicality, accuracy, and relevance

- Person recognition was good
- OpenCV/DVT has a tie fetish
- Context awareness is very poor, especially considering the dataset has ontological tags (cups vs beakers in a lab)
- Strong consistency
- Inference: trained on lots of historical portraiture - great on men in suits and people in portrait poses, confused by casual poses and dance
Getting started [edit]

In this tutorial, we are going to import data about universities in the UK. Our source is a portal run by the UK government which lets us download its data as a CSV file. Open this file with OpenRefine (you can copy the link address and OpenRefine will download the file for you).
Making Connections

HIVE
Vocabulary Server
Helping Interdisciplinary Vocabulary Engineering

Vocabularies Search Index

Search one or more vocabularies for a term or phrase.

1. Select vocabularies
   - AGROVOC
   - ChEBI
   - Diabetes
   - Gastroenterology
   - MeSH
   - Pediatrics
   - SAREF
   - Asthma
   - Chambers
   - CHMO
   - ChambersTest
   - FASTgeo
   - FASTpersonal
   - LCSH
   - LCSH1910
   - Metals
   - Oncology
   - ROO
   - RADLEX
   - UAT
   - USGS

2. Enter search term: African American

Search

FASTpersonal
   - African American

LCSH
   - Africa--Civilization--African American influences
   - African American Bahais
   - African American Baptists
   - African American Biography
From text

- Item 01154 - [has] dc:title - “Wine in the Wilderness, February 1989”
- Item 01154 - [has] dc:date - Feb 1989
- Item 01154 - foaf:depicts - foaf:person
- [said] foaf:person - [has been given] foaf:openid - 214
- 214 - [is] foaf:gender - female
- 214 - foaf:knows - Spelman College etc.
Evaluation

Future automation of semantic connections like these should be highly achievable because Gold Standard case studies show:

- There are only a small number of connections to be followed in order to move from Existing to After

- The relevance of the facts in other sources can be verified by the connections between individuals and Spelman Collections, and

- The connections were found to be both frequent across many websites, and contained within well maintained and official digital archives, the former in a newspaper with full digitized text and the latter in many US National Archives military records, as well as research and reports on Black history and women's history in the military.
Evaluation

Future automation of semantic connections like these should be highly achievable because Gold Standard case studies show:

This not only validates the future potential of the Linking component but also illustrates the importance of providing a framework like this which makes it easier for Archives and Cultural Heritage organizations to adopt Linked Data approaches, since the benefit of wider adoption to the potential for automation is exponential.
Conclusions

There was so much to find!
The information that wasn't in the metadata turned out to be fascinating, a huge asset to the archive, and more entry points

Rules-based Parsers are it for now
Metadata is so information-dense that models trained on more extended writing don't work

Ontologies have Potential
There's so many photos with human-assigned subject tags already organized in ontologies, maybe a context-aware classifier would be a smart contribution for LIS to make to wider AI

+1 Archive = +10-s of connections
The connections that can be made automatically grow more rapidly the more institutions participate