Semantic Analysis Method (SAM): A Tool for Identifying Potential Access Points in Unstructured Text

NKOS 2014 (London, UK)

September 11-12, 2014

Karen F. Gracy, Marcia Lei Zeng, and Sammy Davidson

School of Library and Information Science

Kent State University

The Problem

- Many legacy descriptions in library, archival, and museum (LAM) information systems contain numerous unstructured text blocks.
- Many untapped potential access points can be found in this unstructured data.
- To implement linked data applications in LAM environments, potential access points must be semantically defined and mapped to other vocabularies, such as name authority files and external data sources.
- LAM professionals need a tool to help them solve the challenge of converting unstructured textual descriptions of cultural heritage material into linked data.

Features of Archival Description

- Can occur at multiple levels:
  - The same collection can be described in whole or in part (e.g., a description of subgroupings and individual items).
- Descriptions appearing in bibliographic catalogs are often abbreviated collection-level descriptions (top of the hierarchy), and may have some controlled vocabulary terms attached by catalogers.
- Multi-level finding aids are often generated by processing archivists and may or may not contain controlled vocabulary terms.
- Finding aids can be separated into two major sections,
  - Prefatory notes describing the creator of the materials and the scope and contents of the collection
  - Detailed descriptions at multiple levels, which may or may not contain location information of the material (e.g., Box 3, folder 17)
  - Both sections can be characterized by large blocks of unstructured text.
- Full understanding of a particular entity's importance to the collection as a whole is often reliant on the position of that entity within the larger hierarchy of documents.

Sample Finding Aid : Pearl Harbor Attack (Dec 6-Dec 8, 1941)

Page: 1 of 6 140%

Finding Aid to the Artificial Collection:

Pearl Harbor Attack (Dec 6 – Dec 8, 1941)

Size: (.5 cu.ft.)

Dates: December 6, 1941 – December 8, 1941

Location of Repository: Franklin D. Roosevelt Presidential Library

Name of Finding Aid Author: Ali Caron & Georgina Garcia

Date of Creation: Summer 2011

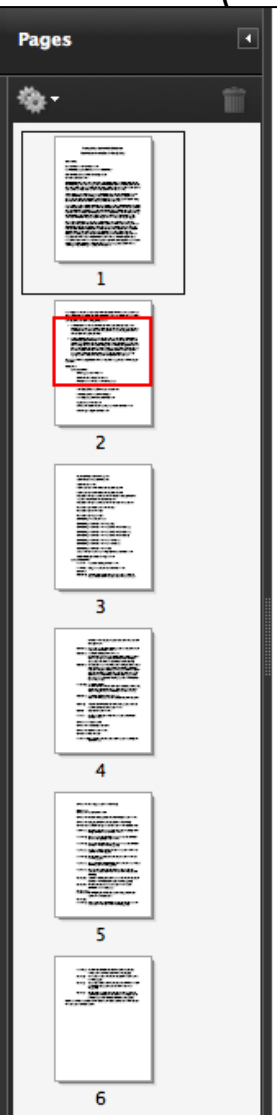
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the public domain. The official writings of United States government officials within this collection are in the public domain. The writings of Eleanor Roosevelt within this collection are subject to Mrs. Roosevelt's literary estate. All other materials are subject to the United States Copyright law, 17 U.S.C. 101, et seq.

Administrative Note: Franklin D. Roosevelt Presidential Library is the first presidential library and the only presidential library used by a sitting president. The library houses documents, objects, and photographs pertaining and relating to President Franklin D. Roosevelt (FDR). FDR was in office during the events of Pearl Harbor and the library houses materials about the attack.

Scope and Content: This artificial collection is composed of photocopies of original documents identified

Source:
http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf

(cont.) Sample Finding Aid : Pearl Harbor Attack (Dec 6-Dec 8, 1941)



Series Descriptions: The collection is organized in 2 series:

- Series I: Documents – The items selected for this series remain within December 6, 1941- December 8, 1941, date range. The items reflect the Pearl Harbor attack or events relevant to that incident. The contents found under the container list portion are arranged by: collection title; and folder title, found in quotations.

- Series II: Still Photographs – Images comprising this series are selected from the FDR Library's General Photograph Collection, folder: WWII: Hawaii: Attack on Pearl Harbor. Listed here are original captions taken from the photographs themselves, along with a Library control number. Unless copyright information is stated in the image caption, all of the photographs in this series belong in the public domain. This means that, to the best of our knowledge, the materials may be freely used by the researcher. However, for copyrighted materials, it is the researcher's responsibility to determine the limits of Fair Use as defined by sections 107 to 118 of the copyright law and to obtain permission from the copyright holder for further use.

Arrangement: Series I is arranged alphabetically after the president's papers and Series II is arranged numerically.

Container Lists:

SERIES I: DOCUMENTS

OF400: Appointments; Hawaii, 1941

OF4675: World War II; General, 1941-1942

PPF200b: Nov. 11, 1941- Jan. 6, 1942; Public Reactions

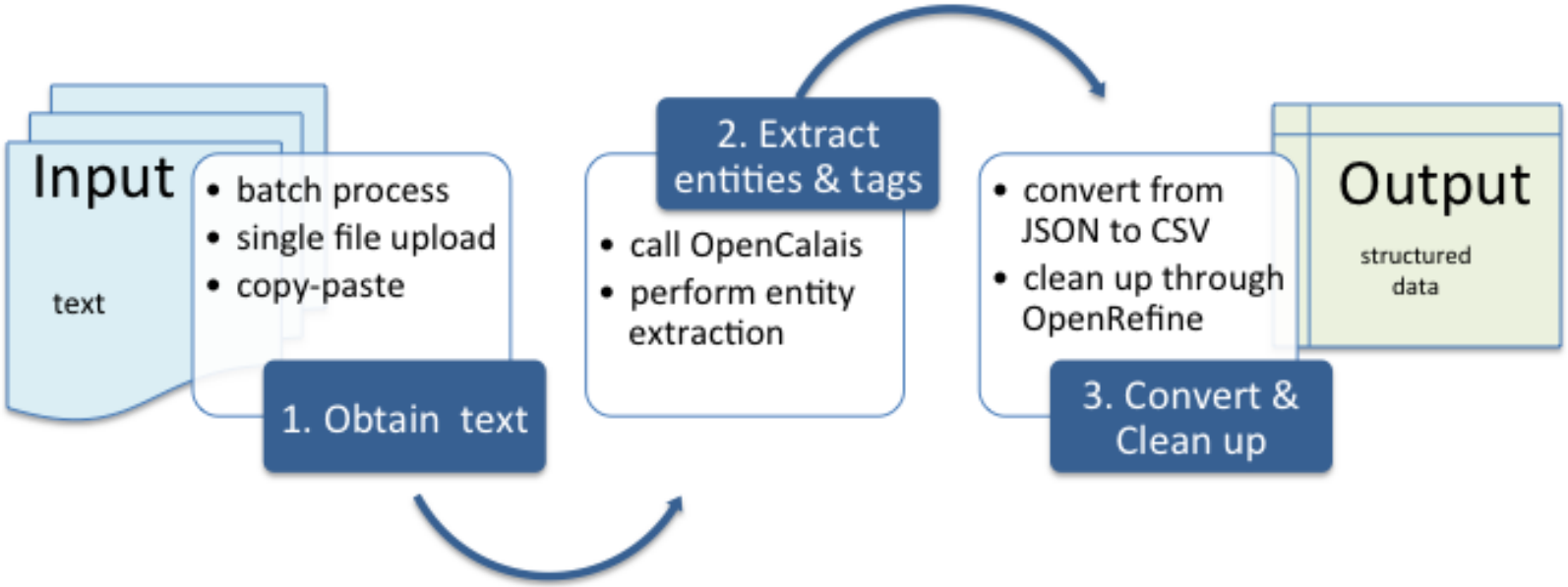
Source:
http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf

The Proposed Solution

The Semantic Analysis Method (SAM) tool provides a bridge from unstructured descriptions and narratives to semantically-enhanced descriptions containing identified and tagged access points.

The SAM tool accomplishes the following:

- Identifies name entities and topics via a semantic analysis engine (OpenCalais);
- Produces an initial output in the form of a JSON data file, which is then converted to the comma-separated-value (CSV) format.
- Resulting CSV file can then be imported into a data cleanup application such as OpenRefine for further editing and removal of misidentified entities.



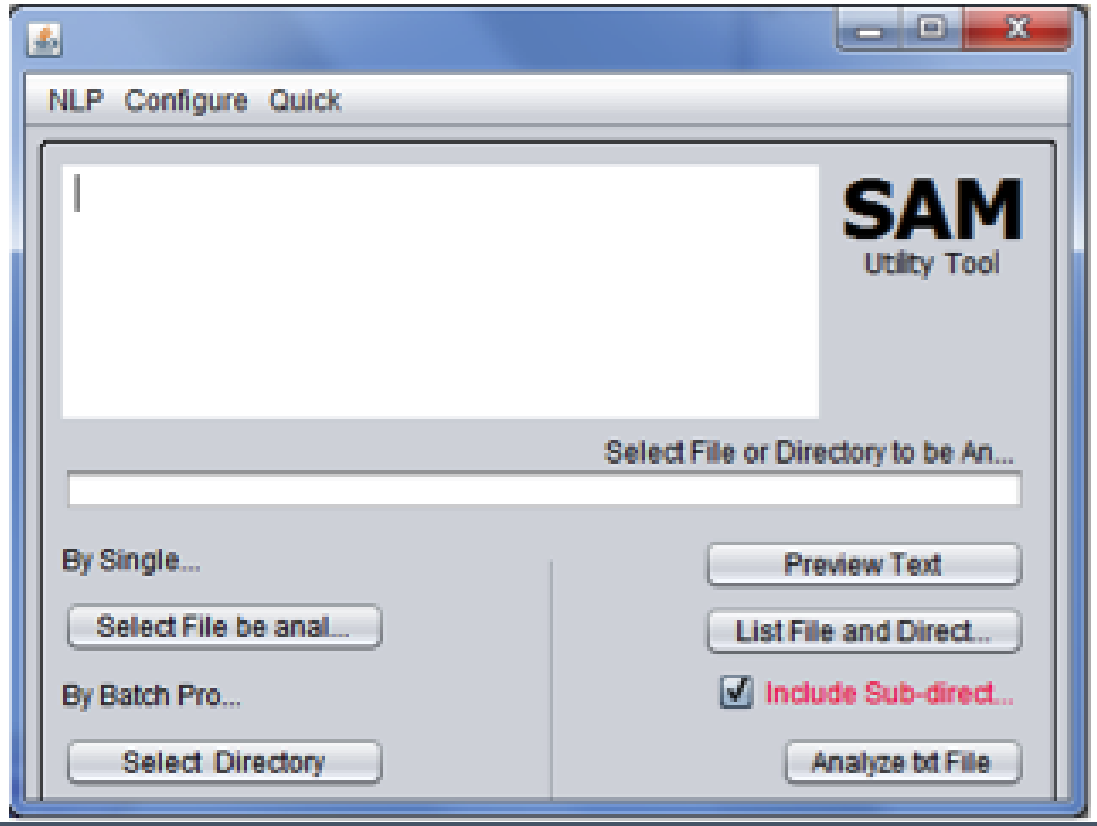
Overview of SAM Tool Functionality

The Semantic Analysis Method (SAM) Tool automates identification and extraction of potential access points and parses the resulting data into a database for further cleanup and editing.

SAM Tool Development

The SAM Tool integrates:

- **Open Calais semantic analysis API service;**
- **j-calais, a third-party library that provides a Java interface to the OpenCalais API; and,**
- **Additional scripts in Java to streamline the tasks of:**
  1. **Obtaining text files from a finding aid data repository;**
  2. **Calling the OpenCalais web service API;**
  3. **Performing the tasks of access point extraction and social tagging through the Open Calais service;**
  4. **Converting the resulting data to the CSV database format.**



SAM Tool

Step 1: Obtaining Text

The Calais initiative is about enabling semantic applications by providing a metadata generation web service, sample applications using that service to jumpstart development efforts, and support for developers.

The Calais Web Service

The Calais web service automatically attaches rich semantic metadata to the content you submit. Using natural language processing, machine learning and other methods, Calais categorizes and links your document with entities (people, places, organizations, etc.), facts (person "x" works for company "y"), and events (person "z" was appointed chairman of company "y" on date "x").

* The Calais Viewer works with Firefox and Internet Explorer - other browsers may yield unpredictable results

Enter text here:

Submit

OpenCalais Viewer

- Open source, free version of semantic analysis engine.
- Creates semantic metadata (lists of entities and social tags), generated in RDF, that can be used for news aggregators and blogs, as well as other linked data applications.
- Users can copy and paste text from PDFs, websites, databases, etc. directly into the window.
- The SAM Tool automates this process of inserting text into the window.

Inputting Text into OpenCalais Semantic Analysis Engine Using the SAM Tool

- Options for inputting text for analysis in SAM Tool include:
  - Manual copy and paste from existing document
  - Single file upload
  - Batch file upload

Finding Aid to the Artificial Collection:

Pearl Harbor Attack (Dec 6 – Dec 8, 1941)

Size: (.5 cu.ft.)

Dates: December 6, 1941 – December 8, 1941

Location of Repository: Franklin D. Roosevelt Presidential Library

Name of Finding Aid Author: Ali Caron & Georgina Garcia

Date of Creation: Summer 2011

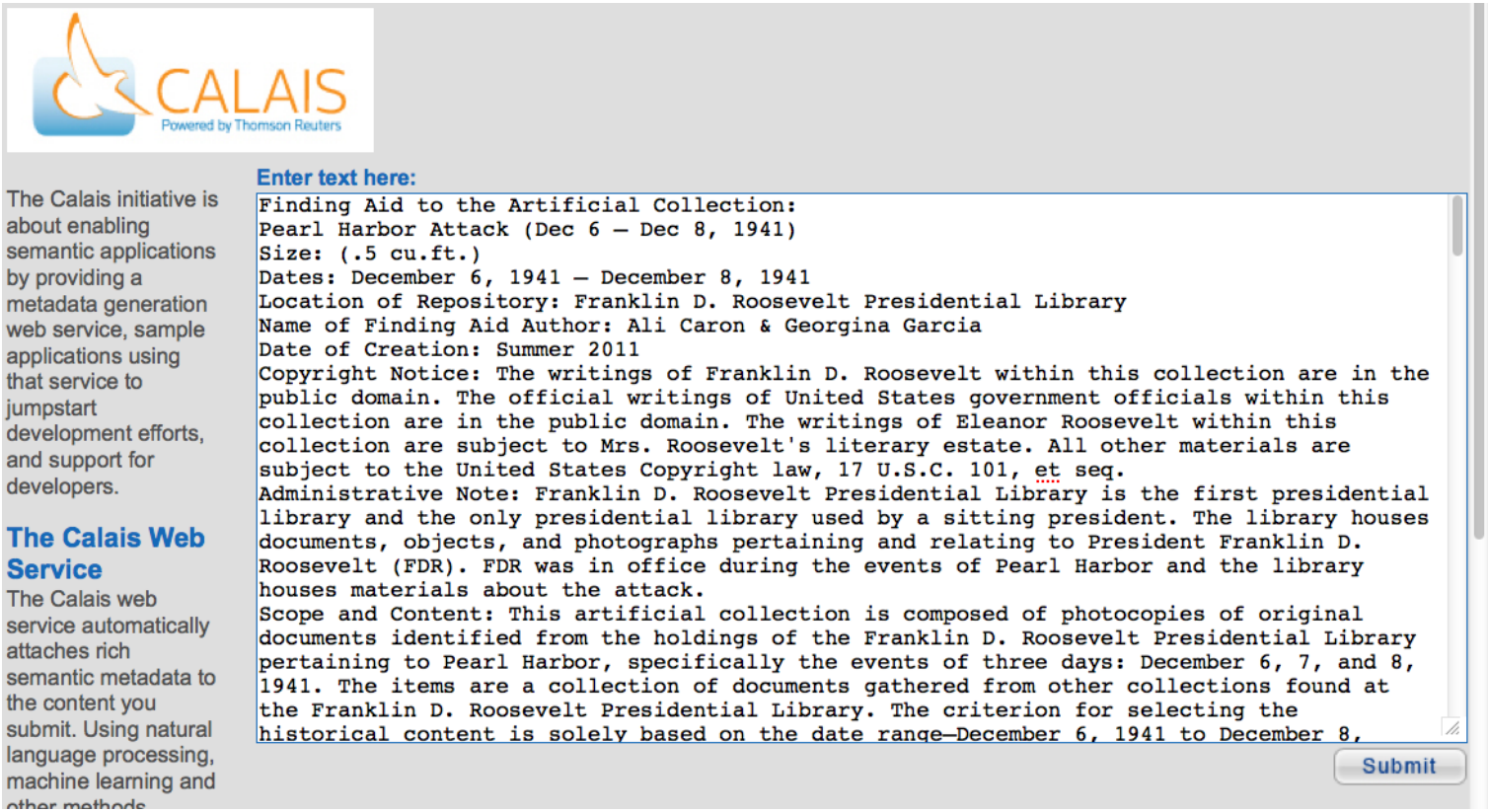
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the public domain. The official writings of United States government officials within this collection are in the public domain. The writings of Eleanor Roosevelt within this collection are subject to Mrs. Roosevelt's literary estate. All other materials are subject to the United States Copyright law, 17 U.S.C. 101, et seq.

Administrative Note: Franklin D. Roosevelt Presidential Library is the first presidential library and the only presidential library used by a sitting president. The library houses documents, objects, and photographs pertaining and relating to President Franklin D. Roosevelt (FDR). FDR was in office during the events of Pearl Harbor and the library houses materials about the attack.

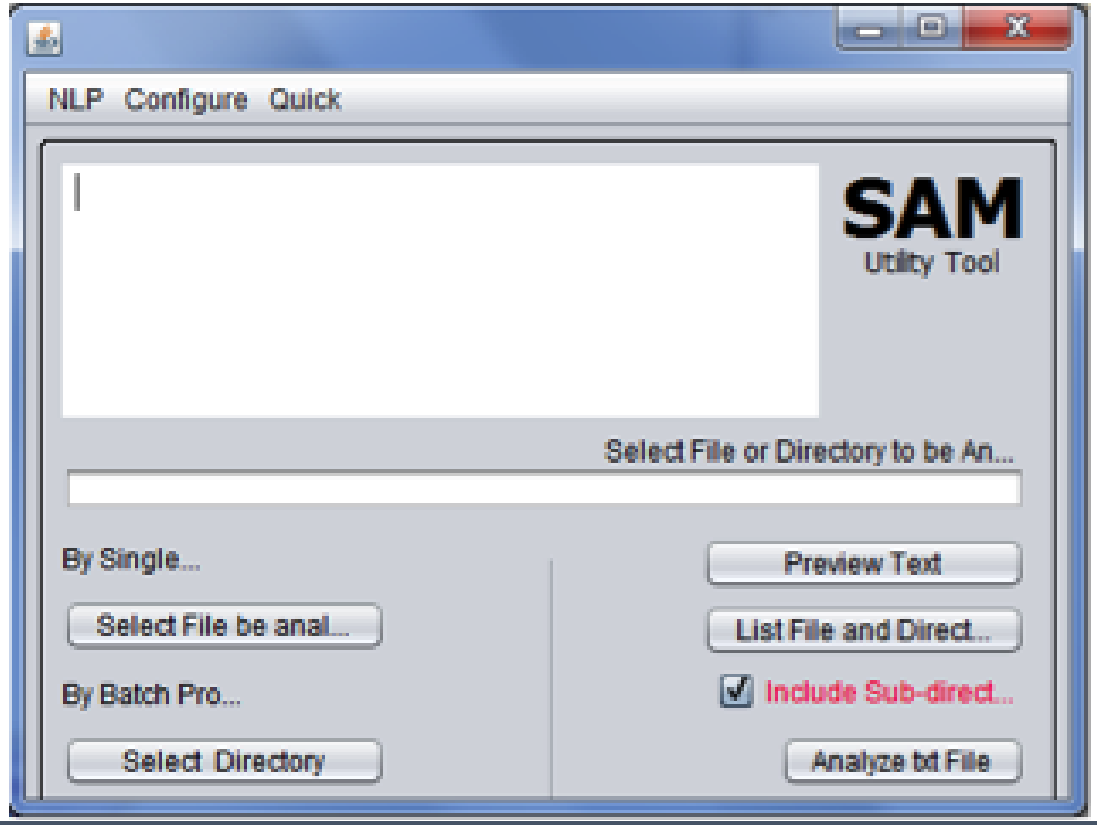
Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. Roosevelt Presidential Library pertaining to Pearl Harbor, specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections found at the Franklin D. Roosevelt Presidential Library. The criterion for selecting the historical content is solely based on the date range—December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs.

Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of materials in the Roosevelt Library documenting the events of Pearl Harbor. There is a vast amount of documents relating to the lead up to Pearl Harbor attack itself, and the aftermath. To simplify matters only the immediate before and after dates are available within this collection. The Pearl Harbor Guide is available for researchers seeking additional information, including documents related to: Indo-China, Japanese-U.S. Relations, Magic, and WWII.

Provenance: The Pearl Harbor artificial collection includes: President's Official File (OF), President's Personal File (PPF), President's Secretary File (PSF), Master Speech File (MSF), Map Room Papers, Francis Biddle Papers, Charles Fahy Papers, William Hasset Papers, Henry Morgenthau Jr. Diaries, Frank A. Schuler Papers, John Toland Papers, and Claude Wickard Papers.

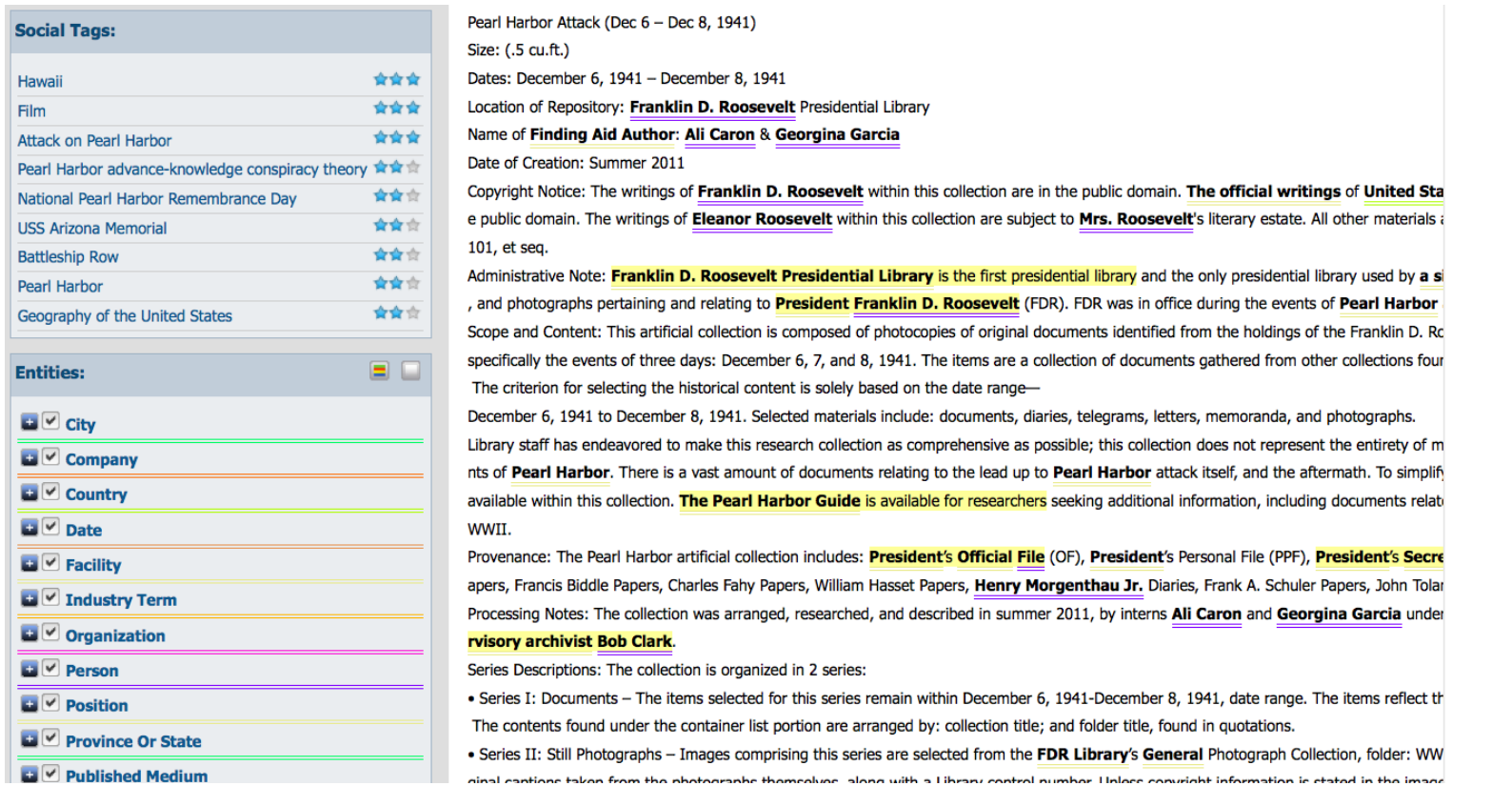


OpenCalais with Input Unstructured Text



SAM Tool

Step 2: Extracting Entities and Tags



Social Tags:

- Hawaii
- Film
- Attack on Pearl Harbor
- Pearl Harbor advance-knowledge conspiracy theory
- National Pearl Harbor Remembrance Day
- USS Arizona Memorial
- Battleship Row
- Pearl Harbor
- Geography of the United States

Entities:

- City
- Company
- Country
- Date
- Facility
- Industry Term
- Organization
- Person
- Position
- Province Or State
- Published Medium

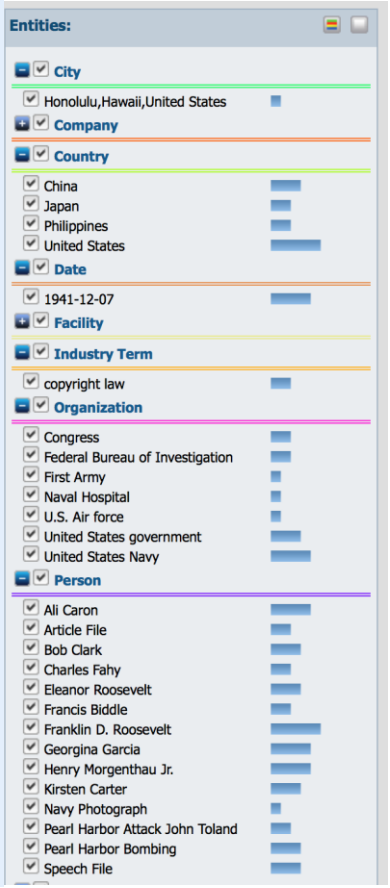
Pearl Harbor Attack (Dec 6 – Dec 8, 1941)
Size: (.5 cu.ft.)
Dates: December 6, 1941 – December 8, 1941
Location of Repository: **Franklin D. Roosevelt** Presidential Library
Name of **Finding Aid Author**: **Ali Caron** & **Georgina Garcia**
Date of Creation: Summer 2011
Copyright Notice: The writings of **Franklin D. Roosevelt** within this collection are in the public domain. **The official writings** of **United Sta**
e public domain. The writings of **Eleanor Roosevelt** within this collection are subject to **Mrs. Roosevelt**'s literary estate. All other materials
101, et seq.
Administrative Note: **Franklin D. Roosevelt Presidential Library** is the first presidential library and the only presidential library used by **a s**
, and photographs pertaining and relating to **President Franklin D. Roosevelt** (FDR). FDR was in office during the events of **Pearl Harbor**
Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. Ro
specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections four
The criterion for selecting the historical content is solely based on the date range—
December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs.
Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of m
nts of **Pearl Harbor**. There is a vast amount of documents relating to the lead up to **Pearl Harbor** attack itself, and the aftermath. To simplify
available within this collection. **The Pearl Harbor Guide** is available for researchers seeking additional information, including documents relat
WWII.
Provenance: The Pearl Harbor artificial collection includes: **President**'s **Official File** (OF), **President**'s Personal File (PPF), **President**'s **Secre**
apers, Francis Biddle Papers, Charles Fahy Papers, William Hasset Papers, **Henry Morgenthau Jr.** Diaries, Frank A. Schuler Papers, John Tolar
Processing Notes: The collection was arranged, researched, and described in summer 2011, by interns **Ali Caron** and **Georgina Garcia** under
rvisory archivist Bob Clark.
Series Descriptions: The collection is organized in 2 series:
• Series I: Documents – The items selected for this series remain within December 6, 1941-December 8, 1941, date range. The items reflect th
The contents found under the container list portion are arranged by: collection title; and folder title, found in quotations.
• Series II: Still Photographs – Images comprising this series are selected from the **FDR Library**'s **General** Photograph Collection, folder: WW
ginal captions taken from the photographs themselves, along with a Library control number. Unless copyright information is stated in the image

Example of Results from OpenCalais Semantic Analysis

Entities Generated by OpenCalais

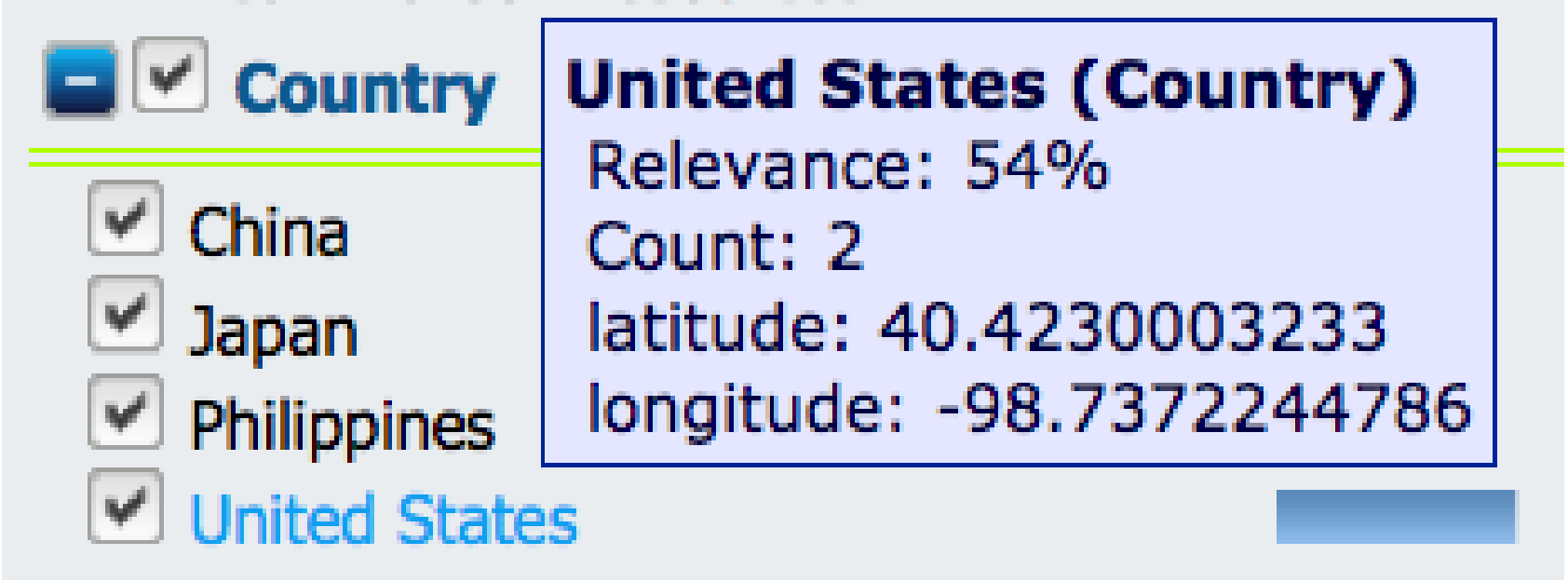
A Few of the More Useful OpenCalais Entity Types

- Person
- Company, Facility, Organization, Product (see also Topics)
- City, Continent, Country, NaturalFeature, ProvinceOrState, Region
- MusicAlbum, Movie, PublishedMedium, RadioProgram, TVShow
- IndustryTerm, Position, Product (see also corporate body names), Technology



OpenCalais Entity Types Mapped to Types of Common LAM Access Points

OpenCalais Entity Types	Entity Groupings	Example Matches to LAM Vocabularies
Person	Personal names	MARC: 100/700 EAD: <persname>
Company, Facility, Organization, Product (see also Topics)	Corporate body names	MARC: 110/710 EAD: <corpname>
City, Continent, Country, NaturalFeature, ProvinceOrState, Region	Geographic names	MARC: 651 EAD: <geogname>
MusicAlbum, Movie, PublishedMedium, RadioProgram, TVShow	Publications (Titles)	MARC: 240; EAD: <title>
IndustryTerm, Position, Product (see also corporate body names), Technology	Topics	MARC: 650 EAD: <subject>

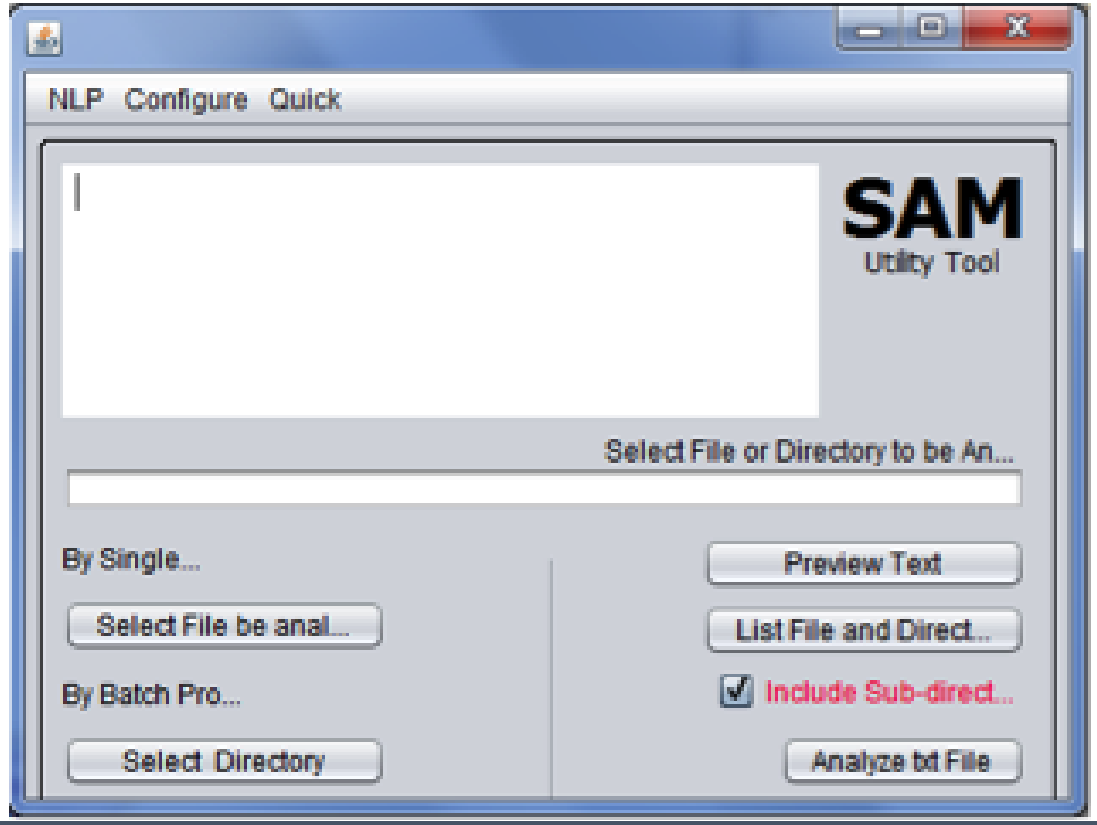


Relevance Rankings

"The relevance scoring takes into account the disambiguation of companies and geographies so that each unique entity will get a single relevance score, even if it is referenced in various ways throughout the text."—OpenCalais website

Social Tags Generated by OpenCalais

- "SocialTags … attempts to emulate how a person would tag a specific piece of content … isn't true semantic extraction."
- "A topic extracted by Categorization with a score higher than 0.6 will also be extracted as a SocialTag. If its score is higher than 0.8, its importance (as a SocialTag) will be set to 1. If the score is between 0.6 and 0.8 its importance is set to 2." – OpenCalais website

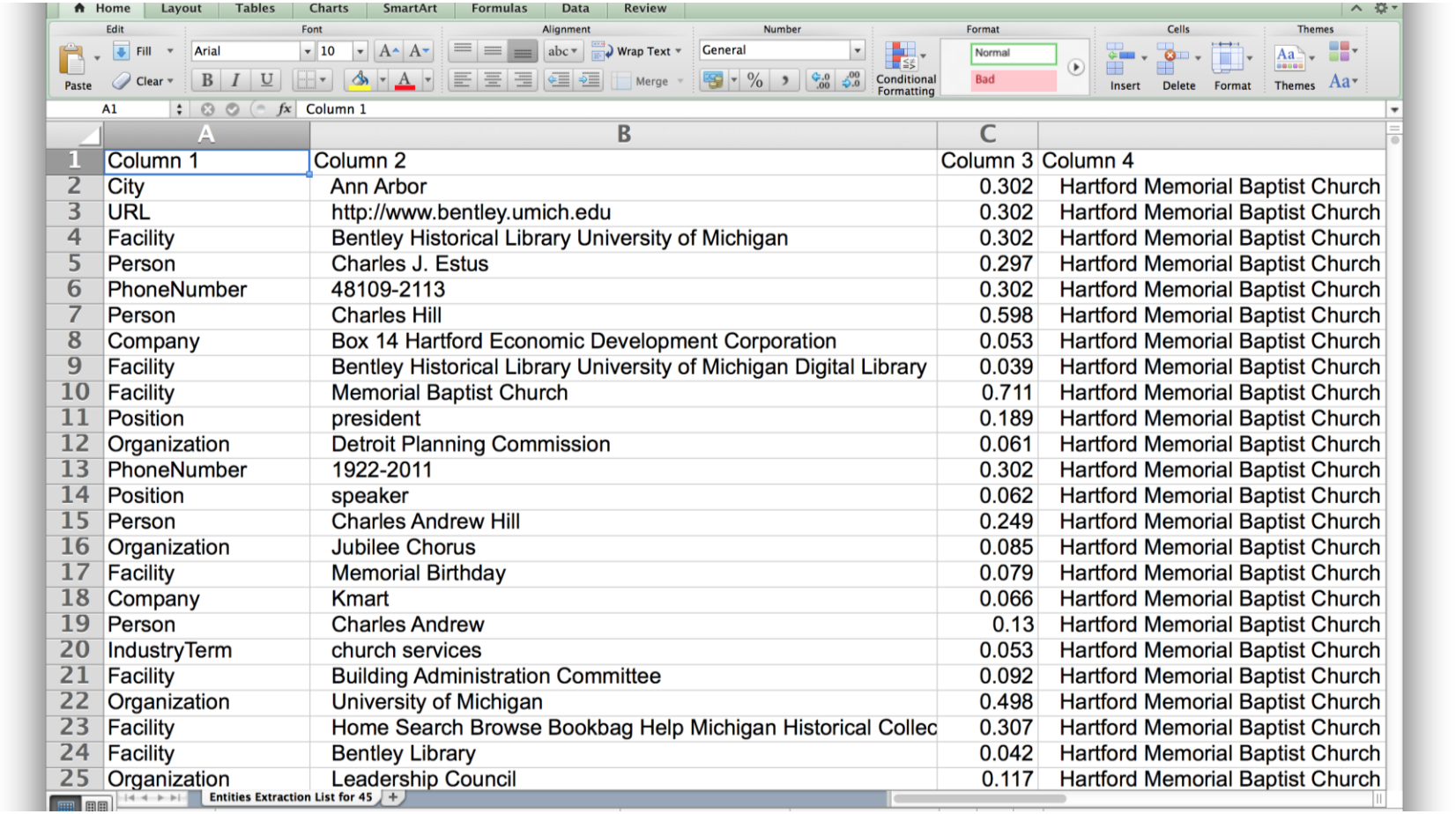


SAM Tool

Step 3: Converting and Clean-Up

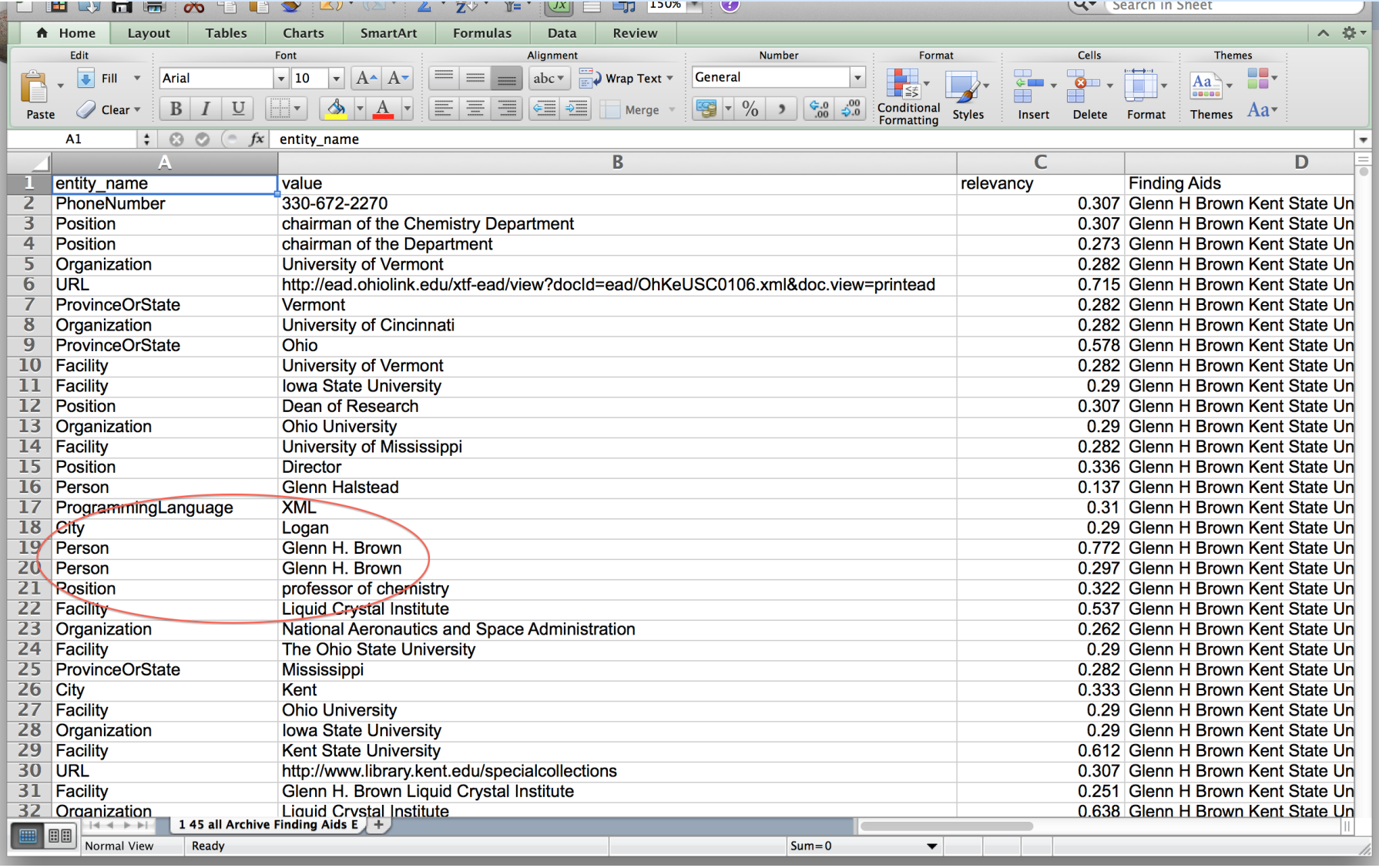
The Resulting Database

- JSON → CSV
- CSV table has four fields:
  - Entity-type
  - Entity-name
  - Relevance-ratio
  - File-source



	A	B	C	
1	Column 1	Column 2	Column 3	Column 4
2	City	Ann Arbor	0.302	Hartford Memorial Baptist Church
3	URL	http://www.bentley.umich.edu	0.302	Hartford Memorial Baptist Church
4	Facility	Bentley Historical Library University of Michigan	0.302	Hartford Memorial Baptist Church
5	Person	Charles J. Estus	0.297	Hartford Memorial Baptist Church
6	PhoneNumber	48109-2113	0.302	Hartford Memorial Baptist Church
7	Person	Charles Hill	0.598	Hartford Memorial Baptist Church
8	Company	Box 14 Hartford Economic Development Corporation	0.053	Hartford Memorial Baptist Church
9	Facility	Bentley Historical Library University of Michigan Digital Library	0.039	Hartford Memorial Baptist Church
10	Facility	Memorial Baptist Church	0.711	Hartford Memorial Baptist Church
11	Position	president	0.189	Hartford Memorial Baptist Church
12	Organization	Detroit Planning Commission	0.061	Hartford Memorial Baptist Church
13	PhoneNumber	1922-2011	0.302	Hartford Memorial Baptist Church
14	Position	speaker	0.062	Hartford Memorial Baptist Church
15	Person	Charles Andrew Hill	0.249	Hartford Memorial Baptist Church
16	Organization	Jubilee Chorus	0.085	Hartford Memorial Baptist Church
17	Facility	Memorial Birthday	0.079	Hartford Memorial Baptist Church
18	Company	Kmart	0.066	Hartford Memorial Baptist Church
19	Person	Charles Andrew	0.13	Hartford Memorial Baptist Church
20	IndustryTerm	church services	0.053	Hartford Memorial Baptist Church
21	Facility	Building Administration Committee	0.092	Hartford Memorial Baptist Church
22	Organization	University of Michigan	0.498	Hartford Memorial Baptist Church
23	Facility	Home Search Browse Bookbag Help Michigan Historical Collec	0.307	Hartford Memorial Baptist Church
24	Facility	Bentley Library	0.042	Hartford Memorial Baptist Church
25	Organization	Leadership Council	0.117	Hartford Memorial Baptist Church

Example of Extracted Entities from Finding Aids



	A	B	C	D
1	entity_name	value	relevancy	Finding Aids
2	PhoneNumber	330-672-2270	0.307	Glenn H Brown Kent State Un
3	Position	chairman of the Chemistry Department	0.307	Glenn H Brown Kent State Un
4	Position	chairman of the Department	0.273	Glenn H Brown Kent State Un
5	Organization	University of Vermont	0.282	Glenn H Brown Kent State Un
6	URL	http://ead.ohiolink.edu/xtf-ead/view?docId=ead/OhKeUSC0106.xml&doc.view=printead	0.715	Glenn H Brown Kent State Un
7	ProvinceOrState	Vermont	0.282	Glenn H Brown Kent State Un
8	Organization	University of Cincinnati	0.282	Glenn H Brown Kent State Un
9	ProvinceOrState	Ohio	0.578	Glenn H Brown Kent State Un
10	Facility	University of Vermont	0.282	Glenn H Brown Kent State Un
11	Facility	Iowa State University	0.29	Glenn H Brown Kent State Un
12	Position	Dean of Research	0.307	Glenn H Brown Kent State Un
13	Organization	Ohio University	0.29	Glenn H Brown Kent State Un
14	Facility	University of Mississippi	0.282	Glenn H Brown Kent State Un
15	Position	Director	0.336	Glenn H Brown Kent State Un
16	Person	Glenn Halstead	0.137	Glenn H Brown Kent State Un
17	ProgrammingLanguage	XML	0.31	Glenn H Brown Kent State Un
18	City	Logan	0.29	Glenn H Brown Kent State Un
19	Person	Glenn H. Brown	0.772	Glenn H Brown Kent State Un
20	Person	Glenn H. Brown	0.297	Glenn H Brown Kent State Un
21	Position	professor of chemistry	0.322	Glenn H Brown Kent State Un
22	Facility	Liquid Crystal Institute	0.537	Glenn H Brown Kent State Un
23	Organization	National Aeronautics and Space Administration	0.262	Glenn H Brown Kent State Un
24	Facility	The Ohio State University	0.29	Glenn H Brown Kent State Un
25	ProvinceOrState	Mississippi	0.282	Glenn H Brown Kent State Un
26	City	Kent	0.333	Glenn H Brown Kent State Un
27	Facility	Ohio University	0.29	Glenn H Brown Kent State Un
28	Organization	Iowa State University	0.29	Glenn H Brown Kent State Un
29	Facility	Kent State University	0.612	Glenn H Brown Kent State Un
30	URL	http://www.library.kent.edu/specialcollections	0.307	Glenn H Brown Kent State Un
31	Facility	Glenn H. Brown Liquid Crystal Institute	0.251	Glenn H Brown Kent State Un
32	Organization	Liquid Crystal Institute	0.638	Glenn H Brown Kent State Un

Example of Cleanup Activity in Resultant Database

Testing the SAM Tool

- Test collection consisted of 45 archival finding aids drawn from 16 repositories.
- Collections were selected to provide a variety of types of archival materials, including:
  - Personal papers
  - Corporate records
  - Government records
  - "Artificial collections," i.e., materials from multiple provenances gathered to document a particular person, family, corporate body, topic, or event.
- OpenCalais raw analysis of the finding aids for these collections resulted in:
  - 8,096 individual entities
  - 336 suggested social tags

Testing the SAM Tool (cont.)

- ***Number of potential access points into collection descriptions identified by semantic analysis was a significant increase over number of controlled vocabulary terms assigned to the same collections by catalogers in collection-level MARC records.***
  - In test collection, the median number of assigned corporate body names in MARC collection-level records was 0-2 names (depending on type of collection)
  - For same collections, analysis of full text of finding aids (describing full extent of collection at all levels), the median number of uncontrolled corporate body entities could range from 0-71, depending on type of collection, and the place in the finding aid (detailed descriptions of series, subseries, files, and items provided the most potential entities).

Testing the SAM Tool (cont.)

- Data clean up will reduce the number of unique entities through the processes of:
  - Deduplication;
  - Collapse of synonyms into single data points;
  - Removal of incorrect extractions.

Errors Generated by the Semantic Analysis Process

- Entity Duplication
- Entity Variants
- Entity Miscategorization
- Inclusion of Unrelated Text as Part of Entity Name

Entity duplication

- Common in archival finding aids, where the same entity can be mentioned in multiple places (history and scope notes, the container listings, series descriptions, etc.)
- Example:
  - New York, N.Y. (extracted and listed five times from the same finding aid)

Entity variants

- Finding aids can contain multiple variants of names, particularly personal and corporate body names.
- The biography or administrative history are the most likely places for entity variants to appear, as names can change over a person's life or the life of a corporate body.
- It can be particularly difficult to resolve names in archival descriptions, as these names are less likely to appear in national/international authority lists.
- Example below, from the Alexander Pope Papers finding aid (three variants found):

Biography / Administrative History

Alexander Hillhouse Pope's grandfathers were both lawyers, and his grandmother was active in a forerunner of the ACLU, which may explain why Alex Pope was walking precincts for Roosevelt and Truman at a young age. He skipped his last two years of high school to attend the University of Chicago where his political activism continued with the National Students Association and Americans for Democratic Action. Pope's college advisor told him law school was the "surest ticket to public office," and in 1952 Pope received his Juris Doctor from the University of Chicago Law School.

Entity miscategorization

- Examples:
  1. *Two Gentleman of Verona* (title miscategorized as Movie, should be Published Medium)
  2. Sandy Hook, Virginia Key (geographic names miscategorized as Persons)

- Entities in finding aids are particularly dependent upon the context within which they are found.
  - From Example #1, *Two Gentleman of Verona* is a work studied by an English professor (title was a folder title within research materials);
  - For Example #2, the geographic names were locations mentioned in several places in collection of materials relating to shore erosion in the United States.

- Archival finding aids rarely use qualifiers within the finding aid to differentiate among entities that may be used in multiple contexts, which will complicate name resolution.

Inclusion of unrelated text

- The formatting of finding aids that are not encoded (in PDF or a word processing format) can often trip up semantic analysis engines.
- Example below:
  - Entity identified by OpenCalais as "Box 9 Traveling Pictures Animation Company" includes a location reference ("Box 9" is not part of the corporate body name)

Box 9	System Installations, 1960s-1970s
Box 9	Those Amazing Animals, 1981
Box 9	Traveling Pictures Animation Company
Box 9	UBS Detroit
Box 9	UBS (Emerson College, Muntz TV, Pontiac/Chrysler, Yale)

OpenRefine Capabilities

Error Type	OpenRefine Resolution?
Entity duplication	✔
Entity variants	✔
Entity miscategorization	No*
Inclusion of unrelated text as part of entity name	No†

- * = Requires human judgment to correct miscategorization.
- † = Reduction of this error would involve pre-processing to remove certain text (such as physical location information).

Challenges of Current Processes for Entity Extraction and Name Resolution

- Limitations of OpenCalais for analysis of archival description
  - OpenCalais optimized for current news and events, not historical people, places, and events;
  - Procedure for inputting text into OpenCalais API results in errors that may be avoided with some pre-processing of documents prior to analysis;
  - Other semantic analysis engines may be more helpful for analyzing archival description; further testing is needed with other tools.

Challenges of Current Processes (cont.)

- Name resolution not successful for entities not found in commonly used name authority files such as Library of Congress Name Authority File, Virtual International Authority File, DBpedia)
  - Enrichment of these files with names from local authority files could significantly increase success at identifying and extracting entities.
  - Social Networks and Archival Context (SNAC) Project is attempting to establish a "sustainable international cooperative program for archival description":
    - Prototype contains over 2.6 million identity descriptions of persons, families, and organizations drawn from OCLC WorldCat and the British Library, which are then linked to holdings in over 3,000 repositories.
    - Works well for people with "strong" identities (well-documented in primary sources), but not so well for people who are "weakly identified."

The Barriers to Establishing Archival Authority Files

- National/international archival authority files may be very difficult to establish and maintain.
  - Unanswered questions about:
    - Who will be responsible for management of the master file, including merging and validating the entries?
- Local or field-specific authority files may be more feasible initially
  - A smaller number of institutions with shared interests and related collections could create and maintain authority records, and may be more able to handle merging/validation of records.
  - Example: American Numismatics Society biographies (http://numismatics.org/authorities/)

Encoded Archival Context (EAC-CPF) and Encoded Archival Description (EAD)

- The recently revised EAD standard for archival descriptions and the quickly growing adoption of the EAC-CPF standard for archival authority descriptions should push further development of linked data-ready archival information.
  - URI's can be embedded tags for the current version of EAC-CPF and the new version of EAD.
  - Ability to link directly to other data sources will encourage more interest in data exchange among and beyond traditional LAM bibliographic and authority files.

Current Status and Future Directions

- We are looking to refine our analysis and name resolution processes:
  - Preprocessing of finding aids to remove "noise" that might lead to inadvertent inclusion of unrelated text;
  - Ways to incorporate contextual information in name resolution;
  - Testing of finding aids encoded in EAD descriptions, in addition to plain text;
  - Direct query of linked data sets such as LCNAF, VIAF, and DBpedia to find proposed matches to established access points;
  - Exploration of new data sources to improve accuracy of name resolution, such as the Social Networks and Archival Context (SNAC) dataset.

Current Status and Future Directions (2)

- Planned additions of new features to the SAM tool include:
  - Generating RDF instead of JSON during OpenCalais analysis and extraction to remove intermediate step of generating a CSV file;
  - Incorporating other semantic analysis engines as options, such as:
    - Jetlore (http://dev.jetlore.com)
    - Machine Linking (http://www.machinelinking.com/wp/)
    - Zemanta (http://www.zemanta.com/api/)
  - Further modularization of processes and procedures to make future updates easier.

For More Information …

- For more information about activities and publications of the LOD-LAM Research Group, please visit the website and contact team members:
  - Website:
    - http://lod-lam.slis.kent.edu/
  - Email:
    - Marcia Lei Zeng (Principal Investigator), mzeng@kent.edu
    - Karen F. Gracy (Co-Principal Investigator), kgracy@kent.edu
    - Sammy Davidson (Graduate Assistant/Software Designer), sdavids6@kent.edu
- To download the most recent source code for the SAM Tool, go to:
  - https://github.com/sammysemantics/SAM

Acknowledgements

Funding for the MV-Junction Project was provided by the generous support of the IMLS National Leadership Grant program and Kent State University School of Library and Information Science.