Challenges of Identifying and Managing Named Entities Found in Archival Finding Aids and Documents

Karen F. Gracy, Ph.D.
School of Information
Kent State University
kgracy@kent.edu

KENT STATE UNIVERSITY | School of Information

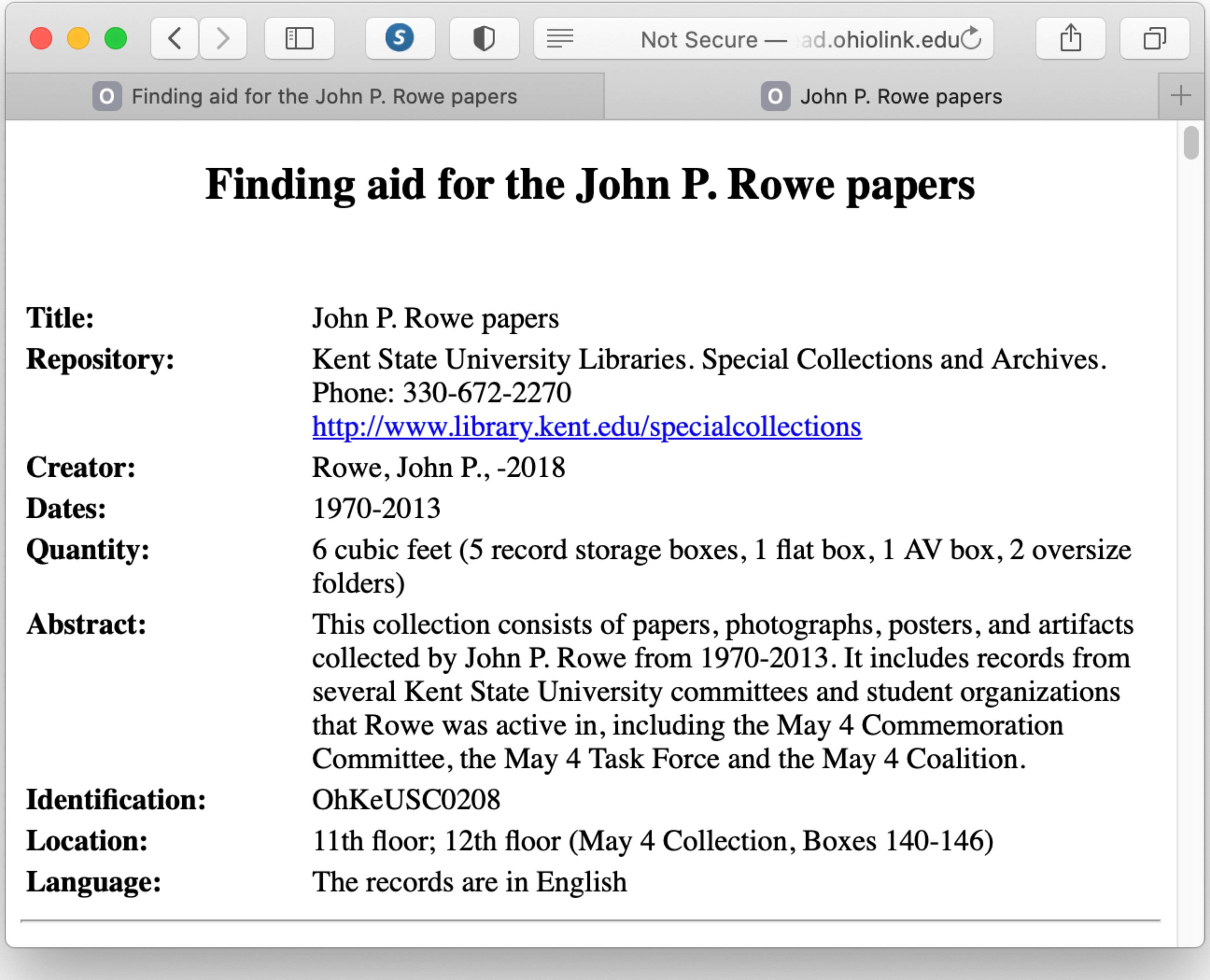
The Untapped Potential of Archival Resources

- Digitization of archival material = Huge increase in number of documents available to researchers/users
- But … discoverability of these resources has not kept pace with the increase in availability.
- Archival collections represent a tremendous source of untapped data, which is not discoverable without significant effort on the part of the researcher.

Problems Specific to Representation of Archival Materials

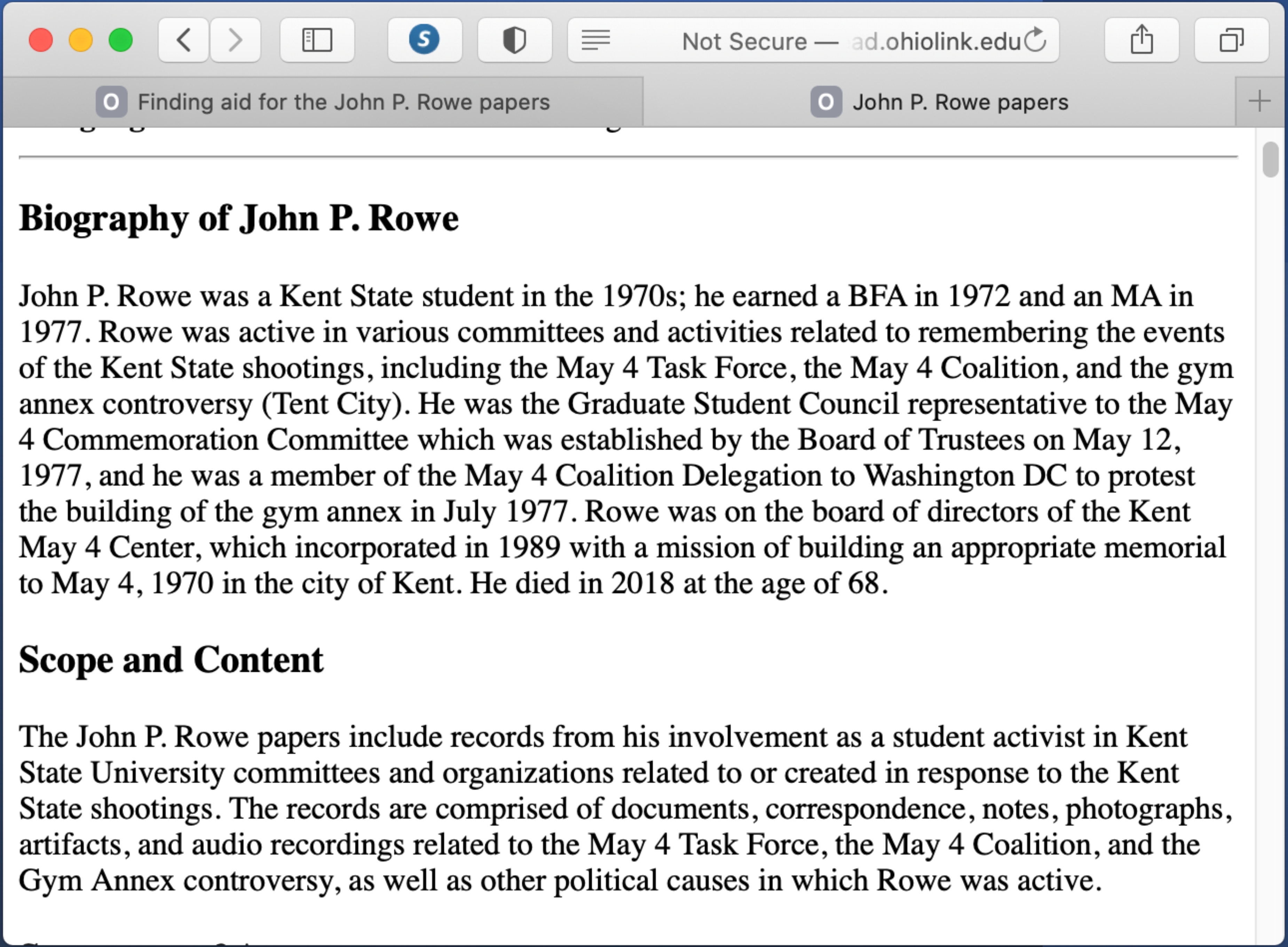
- Traditional methods for describing/ representing archival material in descriptive tools and catalogs make these resources and information found within them hard to discover. Problems include:
  - Unstructured data;
  - Multilevel description vs. single level description;
  - Inadequacy of existing authority files for archival entities (particularly names of persons);
  - Geographic names that may be historical in nature;
  - The lack of guidelines and best practices for the appropriate depth of indexing.

Finding Aid Example from May 4 Collections at Kent State Libraries


Not Secure — ead.ohiolink.edu
Finding aid for the John P. Rowe papers
John P. Rowe papers

Finding aid for the John P. Rowe papers

Title:	John P. Rowe papers
Repository:	Kent State University Libraries. Special Collections and Archives. Phone: 330-672-2270 http://www.library.kent.edu/specialcollections
Creator:	Rowe, John P., -2018
Dates:	1970-2013
Quantity:	6 cubic feet (5 record storage boxes, 1 flat box, 1 AV box, 2 oversize folders)
Abstract:	This collection consists of papers, photographs, posters, and artifacts collected by John P. Rowe from 1970-2013. It includes records from several Kent State University committees and student organizations that Rowe was active in, including the May 4 Commemoration Committee, the May 4 Task Force and the May 4 Coalition.
Identification:	OhKeUSC0208
Location:	11th floor; 12th floor (May 4 Collection, Boxes 140-146)
Language:	The records are in English



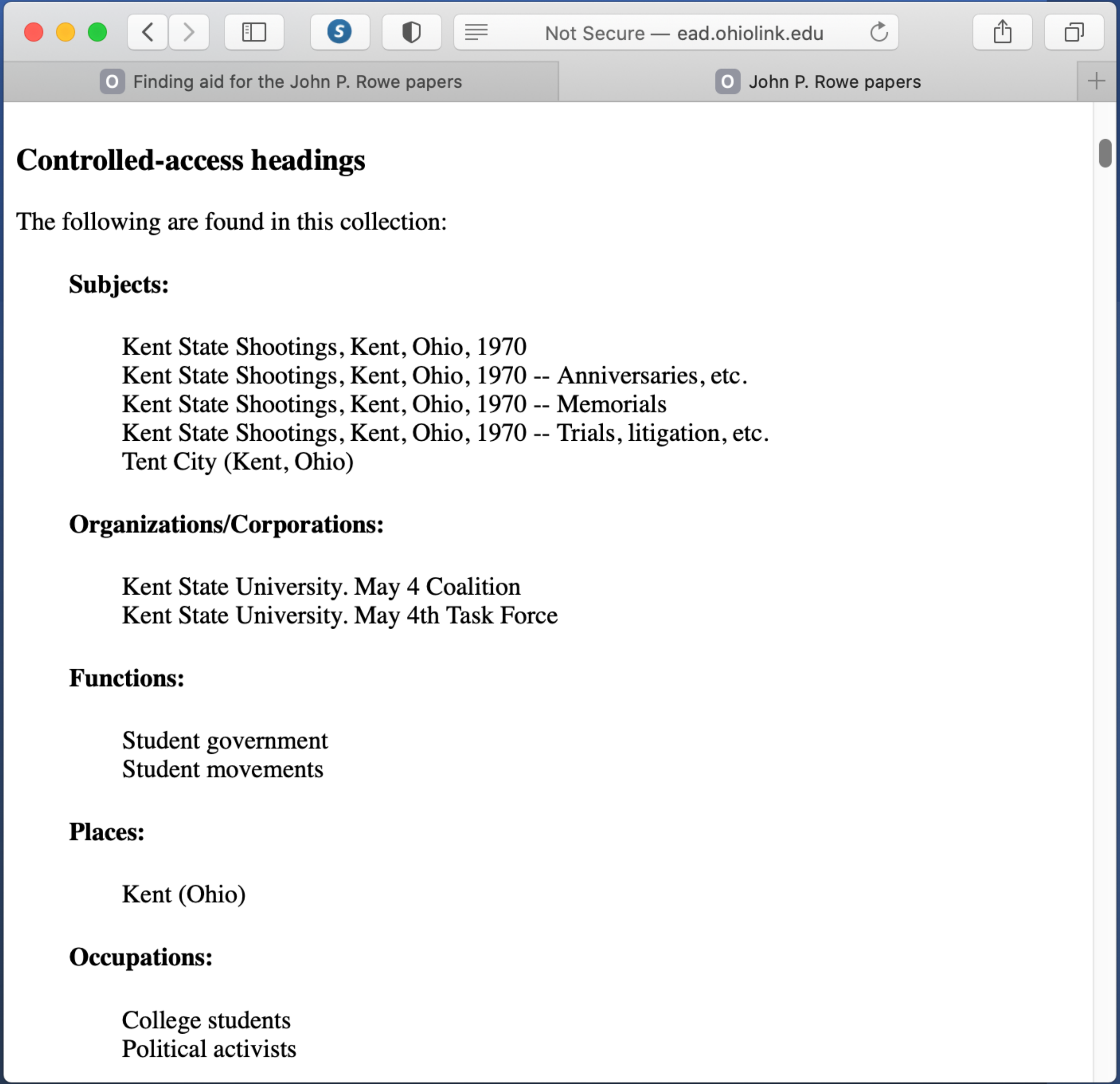
Not Secure — ad.ohiolink.edu
Finding aid for the John P. Rowe papers
John P. Rowe papers

Biography of John P. Rowe

John P. Rowe was a Kent State student in the 1970s; he earned a BFA in 1972 and an MA in 1977. Rowe was active in various committees and activities related to remembering the events of the Kent State shootings, including the May 4 Task Force, the May 4 Coalition, and the gym annex controversy (Tent City). He was the Graduate Student Council representative to the May 4 Commemoration Committee which was established by the Board of Trustees on May 12, 1977, and he was a member of the May 4 Coalition Delegation to Washington DC to protest the building of the gym annex in July 1977. Rowe was on the board of directors of the Kent May 4 Center, which incorporated in 1989 with a mission of building an appropriate memorial to May 4, 1970 in the city of Kent. He died in 2018 at the age of 68.

Scope and Content

The John P. Rowe papers include records from his involvement as a student activist in Kent State University committees and organizations related to or created in response to the Kent State shootings. The records are comprised of documents, correspondence, notes, photographs, artifacts, and audio recordings related to the May 4 Task Force, the May 4 Coalition, and the Gym Annex controversy, as well as other political causes in which Rowe was active.


Not Secure — ead.ohiolink.edu
Finding aid for the John P. Rowe papers
John P. Rowe papers

Controlled-access headings

The following are found in this collection:

Subjects:

Kent State Shootings, Kent, Ohio, 1970
Kent State Shootings, Kent, Ohio, 1970 -- Anniversaries, etc.
Kent State Shootings, Kent, Ohio, 1970 -- Memorials
Kent State Shootings, Kent, Ohio, 1970 -- Trials, litigation, etc.
Tent City (Kent, Ohio)

Organizations/Corporations:

Kent State University. May 4 Coalition
Kent State University. May 4th Task Force

Functions:

Student government
Student movements

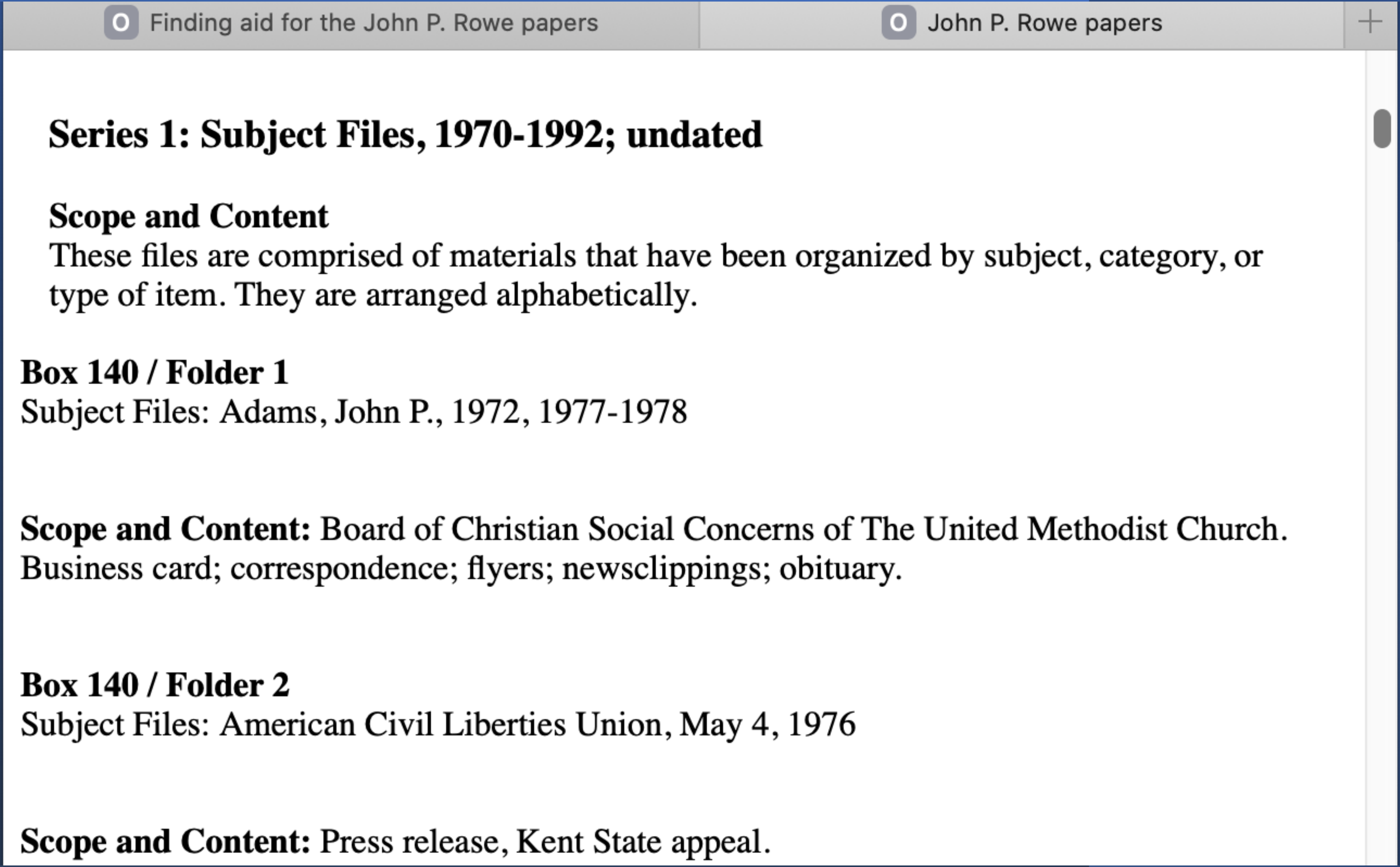
Places:

Kent (Ohio)

Occupations:

College students
Political activists

DCMI 2021



Series 1: Subject Files, 1970-1992; undated

Scope and Content

These files are comprised of materials that have been organized by subject, category, or type of item. They are arranged alphabetically.

Box 140 / Folder 1
Subject Files: Adams, John P., 1972, 1977-1978

Scope and Content: Board of Christian Social Concerns of The United Methodist Church. Business card; correspondence; flyers; newsclippings; obituary.

Box 140 / Folder 2
Subject Files: American Civil Liberties Union, May 4, 1976

Scope and Content: Press release, Kent State appeal.

Linked Data as Potential Solution

- Linked Data offers a potential solution for the problem of increasing discoverability of archival information:
  - Embed semantically structured data in collection inventories and textual documents
  - Interlink related information and make searchable through semantic queries

DCMI 2021

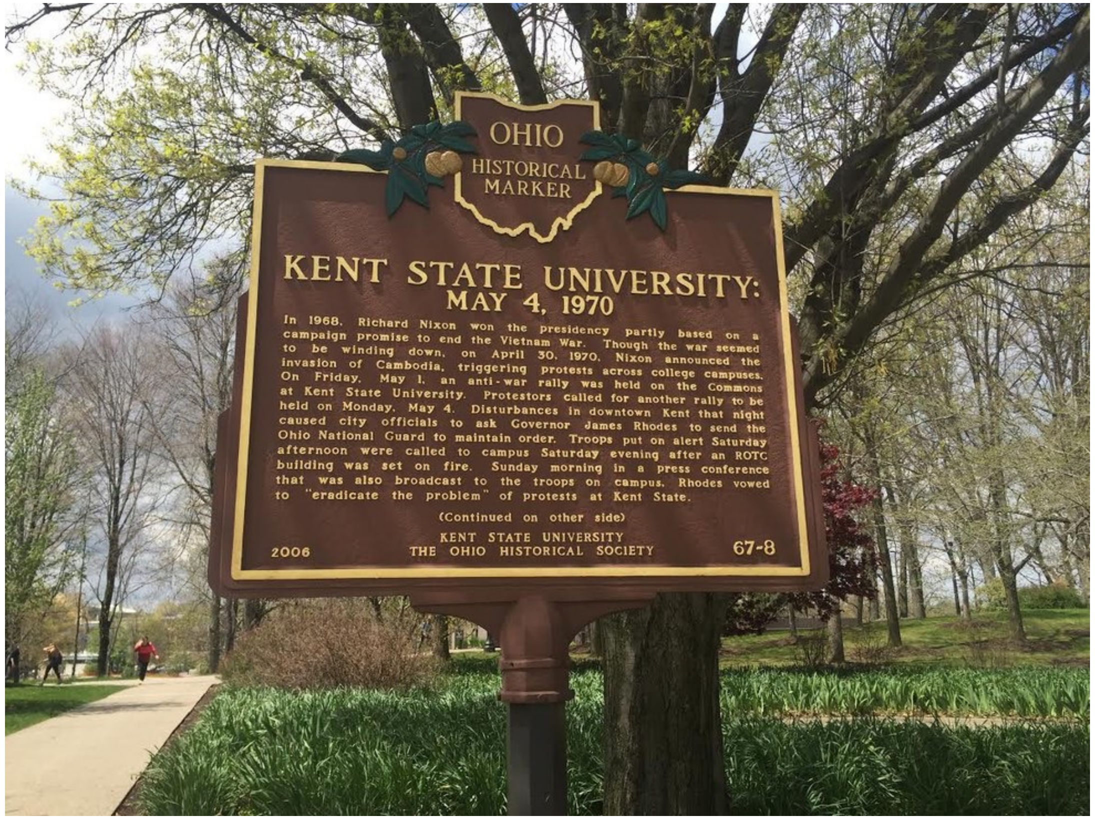
Crafting History: Using a Linked Data Approach to Support the Development of Historical Narratives of Critical Events

May 4 Collections

Kent State University Libraries

Background of Project

- This particular project focuses on the **difficulties of finding information on historical events in archival collections**.
- **Events are a special form of named entity**
  - **Event = nexus point that marks a relationship between specific agents, places, and points in time**
  - **Event = gathering mechanism for records of actions and are crucial aspects of archival information systems.**
- **Test case: May 4 tragedy at Kent State University**



In 1968, Richard Nixon won the presidency partly based on a campaign promise to end the Vietnam War. Though the war seemed to be winding down, on April 30, 1970, Nixon announced the invasion of Cambodia, triggering protests across college campuses. On Friday, May 1, an anti-war rally was held on the Commons at Kent State University. Protestors called for another rally to be held on Monday, May 4. Disturbances in downtown Kent that night caused city officials to ask Governor James Rhodes to send the Ohio National Guard to maintain order. Troops put on alert Saturday afternoon were called to campus Saturday evening after an ROTC building was set on fire. Sunday morning in a press conference that was also broadcast to the troops on campus, Rhodes vowed to "eradicate the problem" of protests at Kent State.

(Continued on other side)

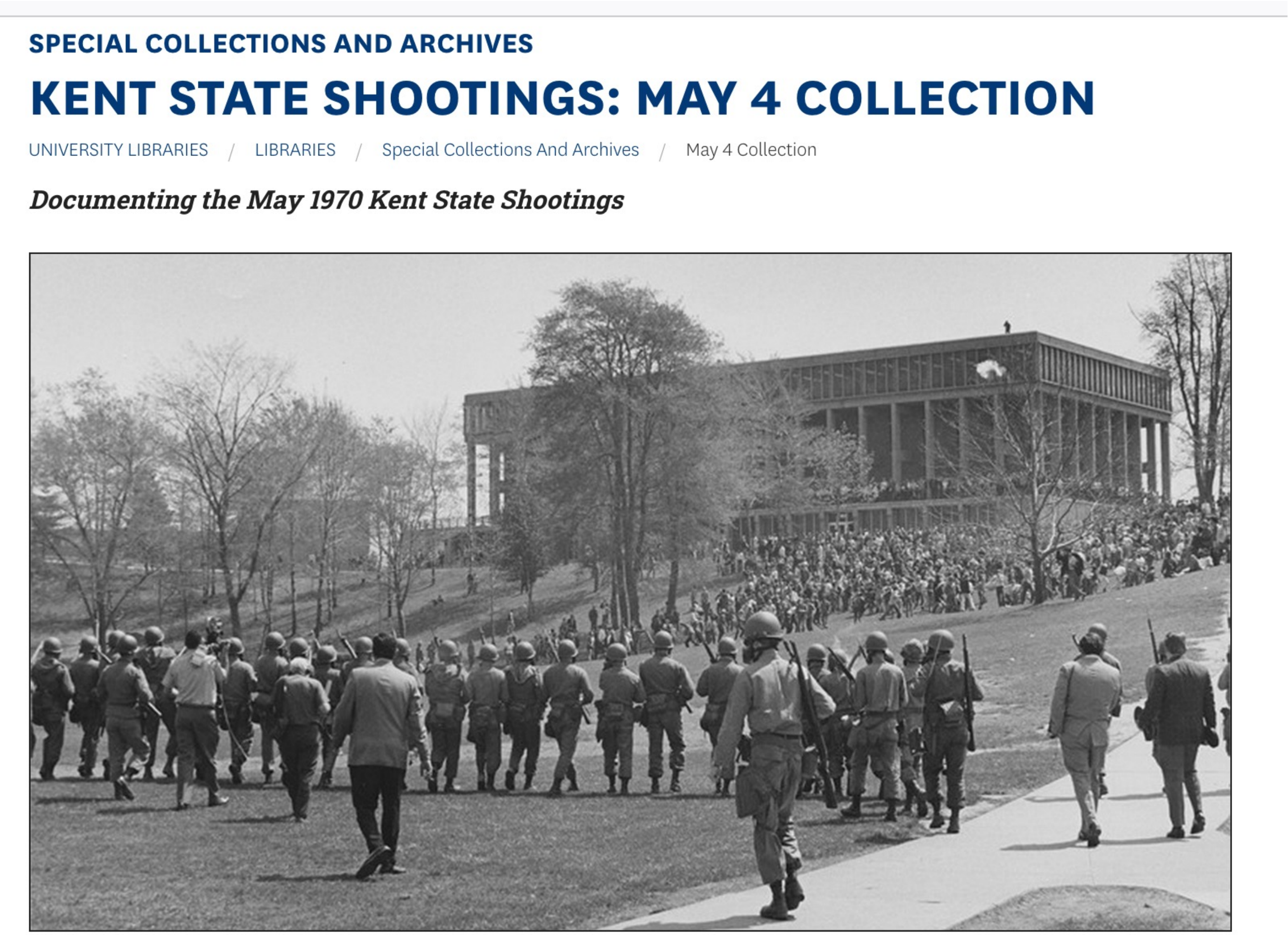
Project Goal

To explore how historians and other humanities scholars can most effectively access and use the data hidden in the silos of digital archival collections to craft narratives about significant developments and critical junctures in historical events, using Linked Data and event-based description.

Research Objectives

1. **To investigate the efficacy of an event-based model of description that will facilitate search across archival inventories and textual documents found in archival collections (current focus)**
2. To develop and test a software tool that will allow scholars to more easily discover and use these hidden nuggets of information about events and facilitate the construction of explanatory narratives about historical phenomena (future work).
3. To create historical event vocabularies and ontologies from entities derived from archival materials, and thus create and test an event-based model that encompasses spatio-temporal dimensions and agents associated with particular events (future work).

Methods: Data and Tools



SPECIAL COLLECTIONS AND ARCHIVES

KENT STATE SHOOTINGS: MAY 4 COLLECTION

UNIVERSITY LIBRARIES / LIBRARIES / Special Collections And Archives / May 4 Collection

Documenting the May 1970 Kent State Shootings

<u>Datasets</u>:

- Descriptions of archival materials covering period before, during, and after the May 4 events (1967-1975), including
  - finding aids for archival collections,
  - captions of photographs, and
  - oral history transcripts.

<u>Entity Extraction Methods</u>:

1. Open Calais and Cogito semantic analysis tools employed.
2. HTML files generated from extractions were converted to text and then processed using the homegrown Semantic Analysis Method (SAM) Tool to create CSV files (placing entities in separate columns).
3. Data clean-up included removal of extraneous characters and spaces, correcting miscategorizations, addition of geographic coordinates based on geographic location, and date normalization; clean-up involved a combination of Open Refine or manual correction.

DCMI 2021

Additional Data Post-Processing

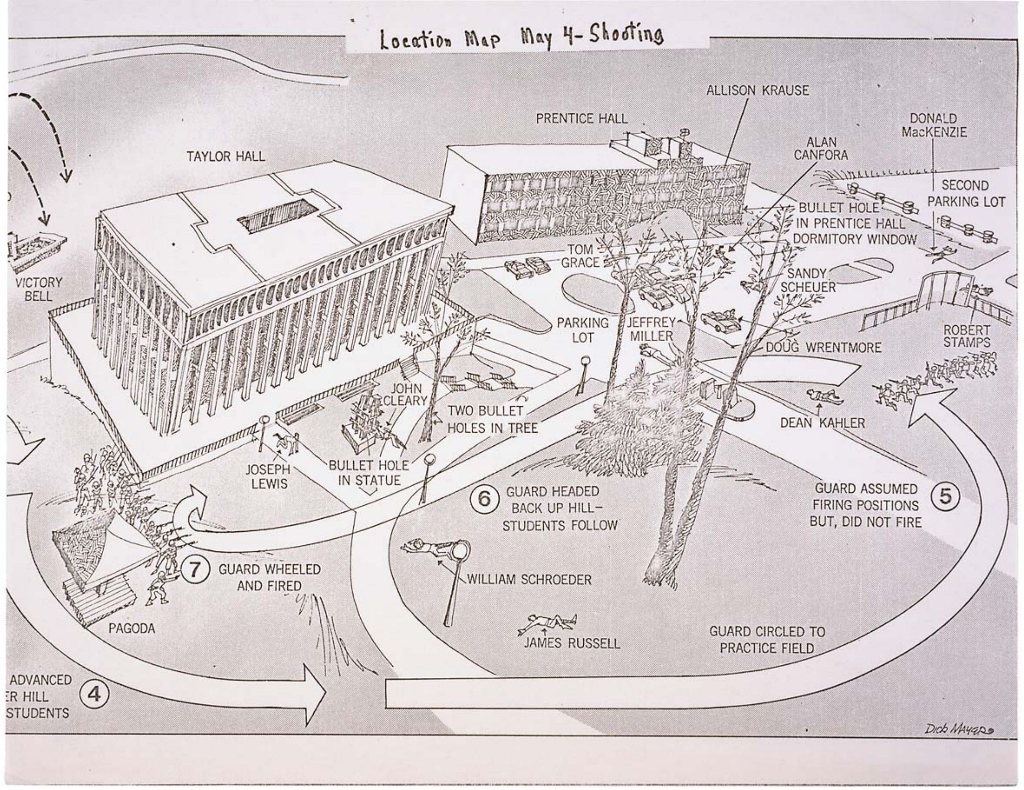
- For personal names, post-processing also included the addition of URIs from VIAF, LC Names, and Wikidata.
- Geographic names were enriched with URIs from Geonames and the Thesaurus of Geographic Names.
- Other data categories that were also captured include corporate names and media outlets (including newspaper titles and station names).
- All of these entities may be potential entry points to archival records.

Number of Place, Person, and Facility/Building Names Generated from Data Sources

	Geographic names	Person names	Facility/Building names	Dates/ Times
Finding aids (N = 111)	171	1,505	386	+2,000
Photograph captions (N = 30)	13	7	Not extracted	Not extracted
Oral history transcripts (N = 108)	165	504	153	+1,200

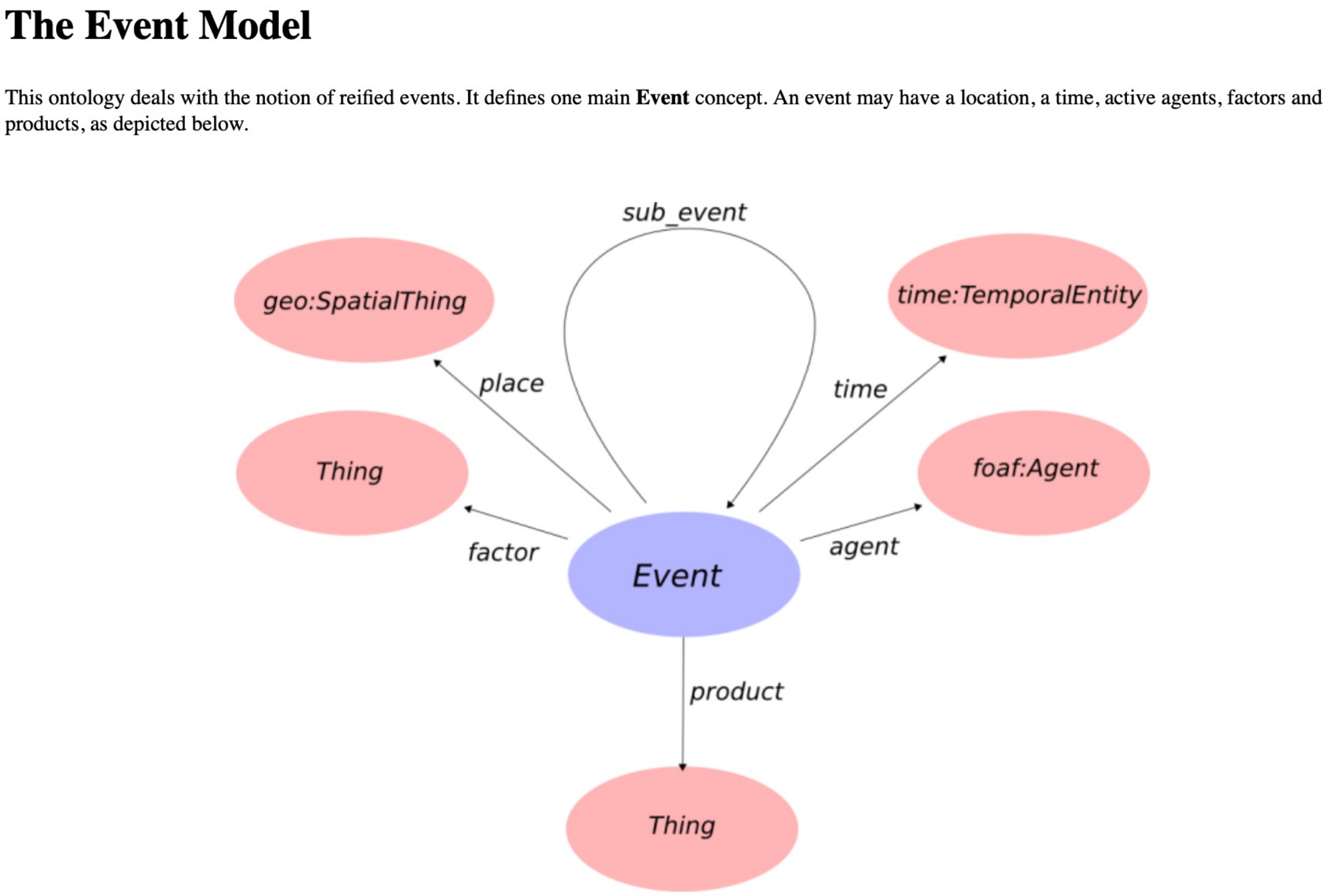
Additional Preliminary Findings

- Going beyond personal and geographic names for events:
  - The importance of building names became apparent early on, particularly for this event.
  - Placement of actors in a location may be needed for particular events.
- Representation of time
  - Can be expressed concretely or contextually
  - Need to find ways to express contextual time statements in our data model for this project.
- Increase in specificity required dramatically increased time needed for data clean-up and analysis.



Possible Influences on Event Model for Project

Event Ontology for a Music Production Environment



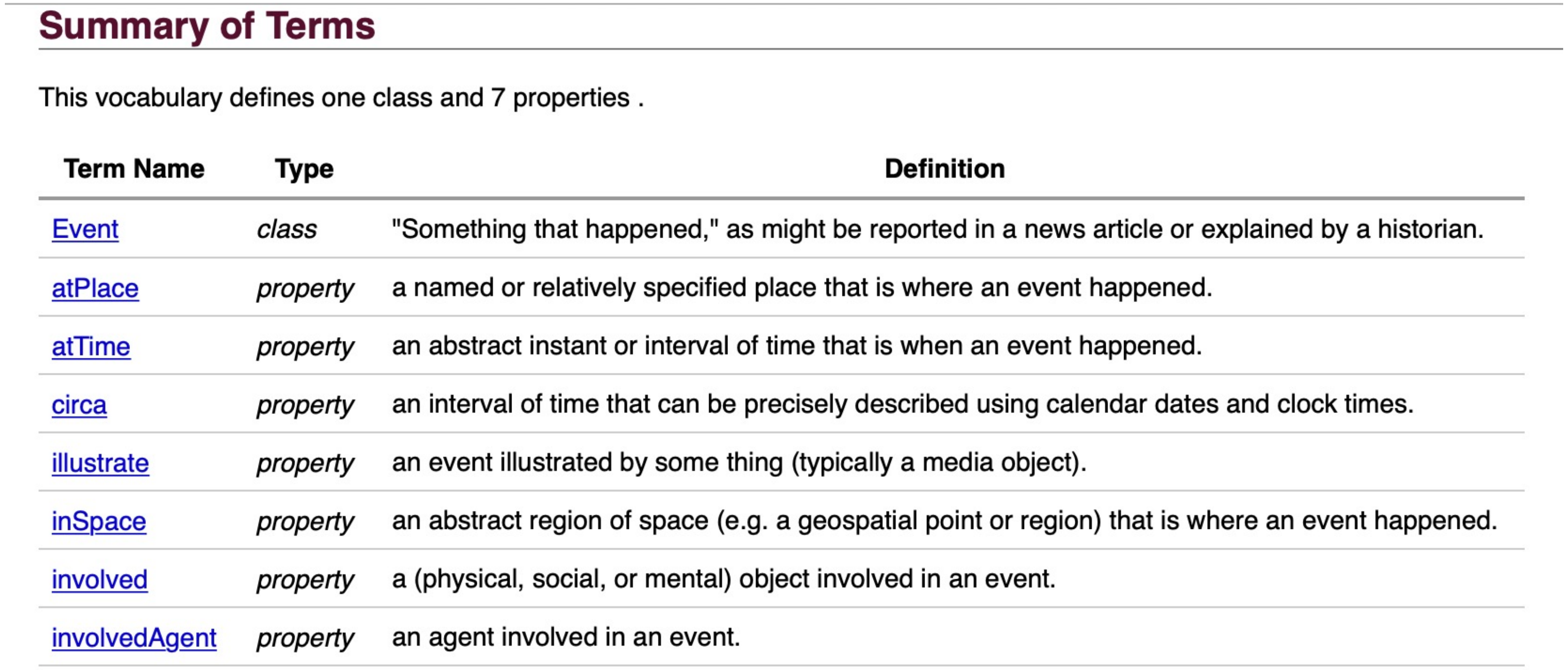
The Event Model

This ontology deals with the notion of reified events. It defines one main **Event** concept. An event may have a location, a time, active agents, factors and products, as depicted below.

Source: Yves Raimond, Samer Abdallah, The Event Ontology, http://motools.sourceforge.net/event/event.html#event

Possible Influences on Event Model for Project

Event Ontology for Publishing Descriptions of Historical Events



Summary of Terms

This vocabulary defines one class and 7 properties .

Term Name	Type	Definition
Event	*class*	"Something that happened," as might be reported in a news article or explained by a historian.
atPlace	*property*	a named or relatively specified place that is where an event happened.
atTime	*property*	an abstract instant or interval of time that is when an event happened.
circa	*property*	an interval of time that can be precisely described using calendar dates and clock times.
illustrate	*property*	an event illustrated by some thing (typically a media object).
inSpace	*property*	an abstract region of space (e.g. a geospatial point or region) that is where an event happened.
involved	*property*	a (physical, social, or mental) object involved in an event.
involvedAgent	*property*	an agent involved in an event.

Source: Ryan Shaw, LODE: An ontology for Linking Open Descriptions of Events, https://linkedevents.org/ontology/

Next Steps

- After completing initial design of event-based model, the project investigators will:
  - Develop and test a prototype tool for event information discovery and use with historians, students, and history enthusiasts.
  - Use tool to test the validity of the event-based model as a suitable approach for facilitating information discovery for archival materials.
- Desired outcome:
  - Provide model for empowering humanities researchers to build complex historical narratives from various primary and secondary sources. Model should be adaptable beyond current project.

Acknowledgements

- This work was partially funded by Research and Creative Activity Fund of the College of Communication and Information at Kent State University.
- Many thanks to my iSchool research team:
  - Partners, Dr. Marcia L. Zeng and Virginia Dressler (University Libraries)
  - iSchool postdoctoral scholar Dr. Tao Hu
  - Graduate research assistants Julaine Clunis and Charles Dupre.

Selected Sources

- **Barthes, R.** (1977). Introduction to the Structural Analysis of Narrative. In Heath, S. (trans), *Image, Music, Text*. New York: Hill & Wang, pp. 79-124.
- **Gracy, K.F.** (2015). Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges, *Archival Science* 15: 239-254. doi: 10.1007/s10502-014-9216-2
- **Hyvönen E., Lindquist T., Törnroos J., & Mäkelä E.** (2012). History on the Semantic Web as Linked Data—An Event Gazetteer and Timeline for the World War I. Proceedings of *CIDOC 2012, Enriching Cultural Heritage*, 10-14 June 2012, Helsinki, Finland. Retrieved from http://www.cidoc2012.fi/en/File/1609/hyvonen.pdf.
- **White, H.** (1984). The Question of Narrative in Contemporary Historical Theory, *History and Theory* 23(1): 1-33.