Lessons Learned in Building Linked Data for the American Art Collaborative

Craig Knoblock

University of Southern California

Information Sciences Institute

Pedro Szekely, Eleanor Fink, David Newbury, Robert Sanderson,
Duane Degler, Kate Blanch, Sara Snyder, Nilay Chheda, Nimesh Jain,
Ravi Raju Krishna, Nikhila Begur Sreekanth, and Yixiang Yao

Project Goals

- Launch the American Art Collaborative
  - Consortium of 14 American art museums
  - Explore the use of Linked Data to make their data available for research, education, and outreach
- Build 5* Linked Data for the museums
  - Map the data about artwork and artists to a common ontology
  - Link the data to other resources
  - Create/extend tools to support the construction of Linked Data
  - Create applications using the data

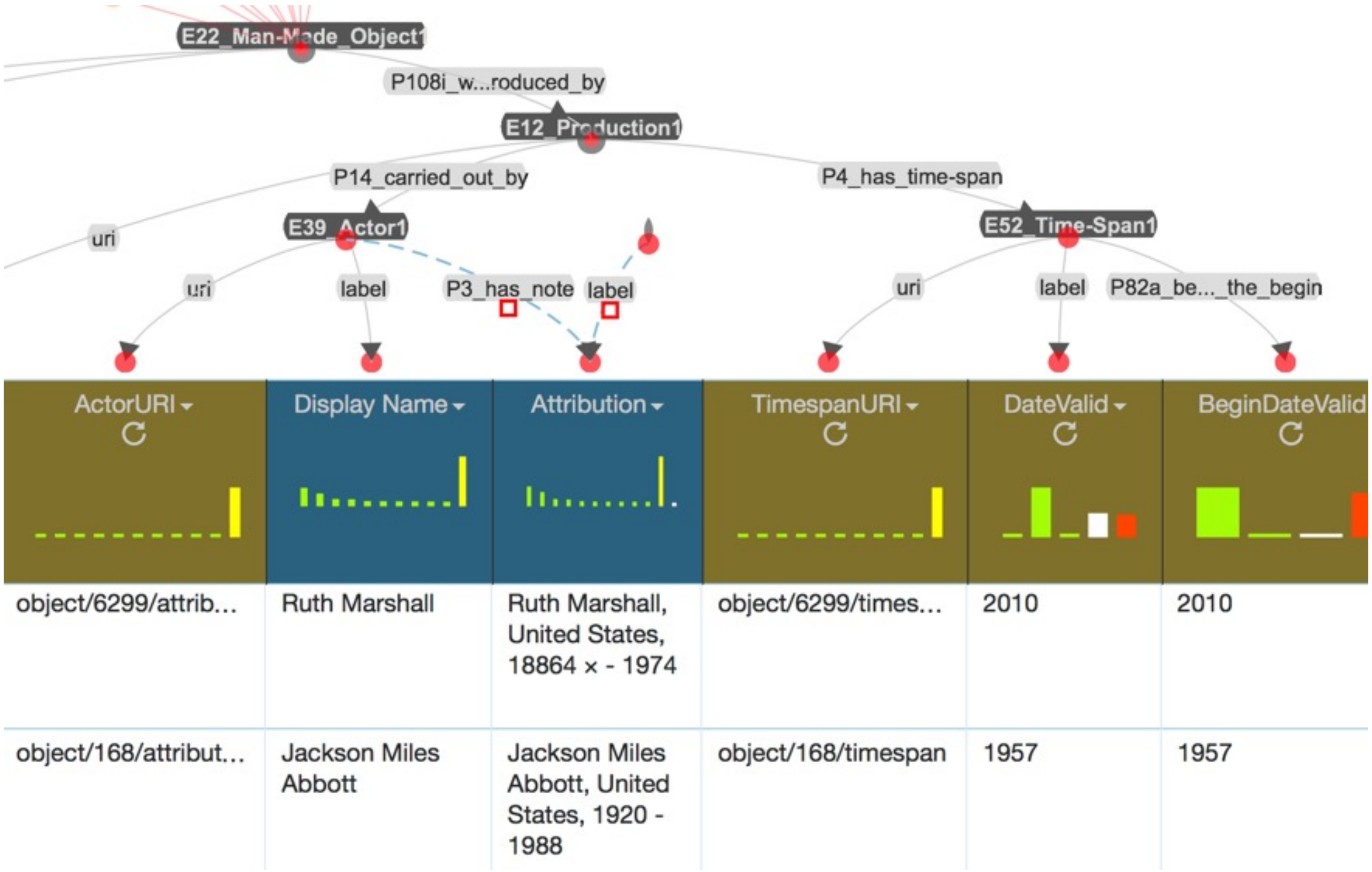
Outline

- Mapping the data
- Linking the entities
- Using the Linked Data
- Related Work & Discussion

Mapping the Data

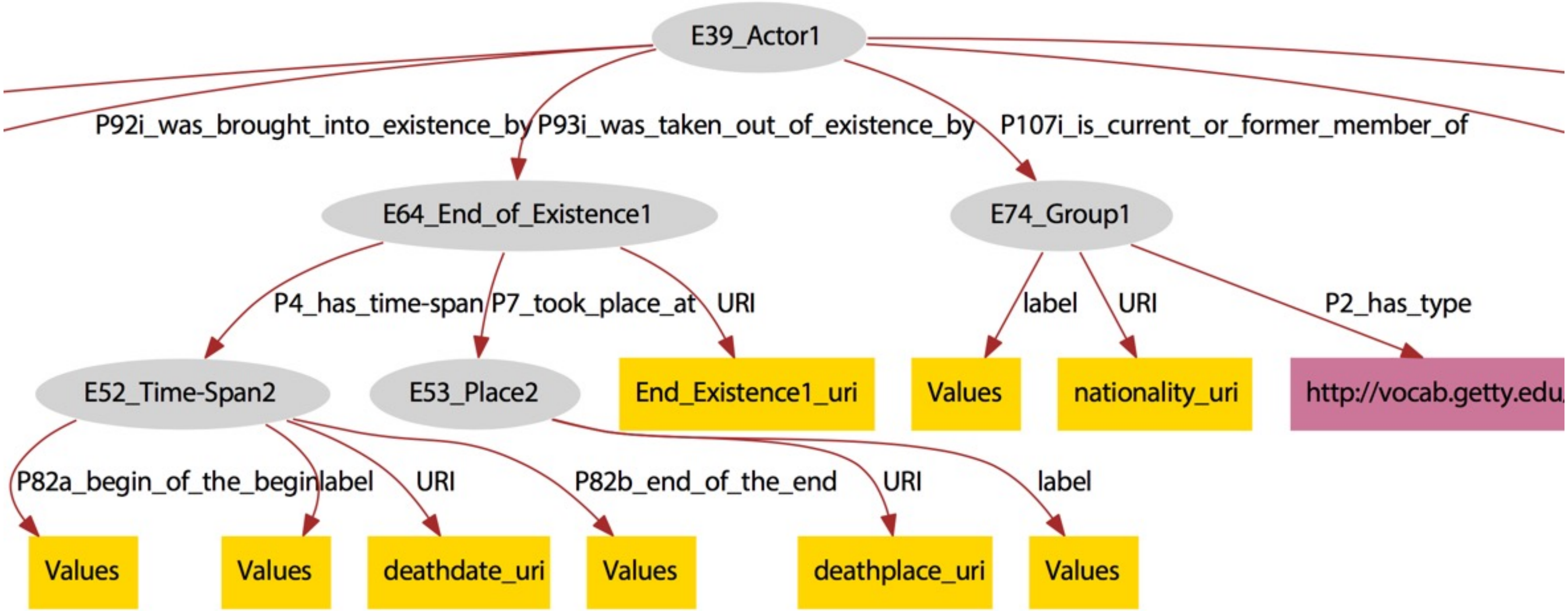
- Challenges
  - Museums have the data in wildly different formats and use different schemas
  - The CIDOC-CRM ontology is a large and complicated ontology
- Approach
  - Use Github to organize all of the data, mappings, and resulting RDF
  - Use Karma to create the mappings of each dataset
  - Trained a team of USC students to apply the tools to the datasets

Use Karma to Map the Data to the Ontology

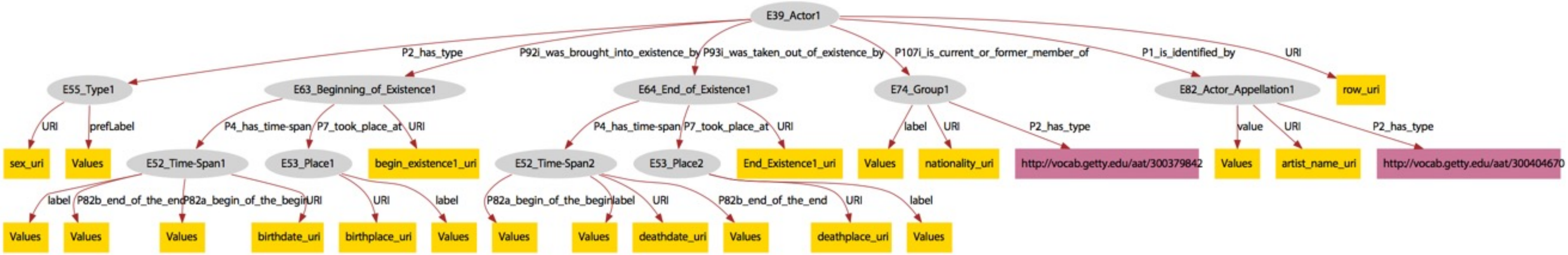


ActorURI	Display Name	Attribution	TimespanURI	DateValid	BeginDateValid
object/6299/attrib...	Ruth Marshall	Ruth Marshall, United States, 18864 × - 1974	object/6299/times...	2010	2010
object/168/attribut...	Jackson Miles Abbott	Jackson Miles Abbott, United States, 1920 - 1988	object/168/timespan	1957	1957

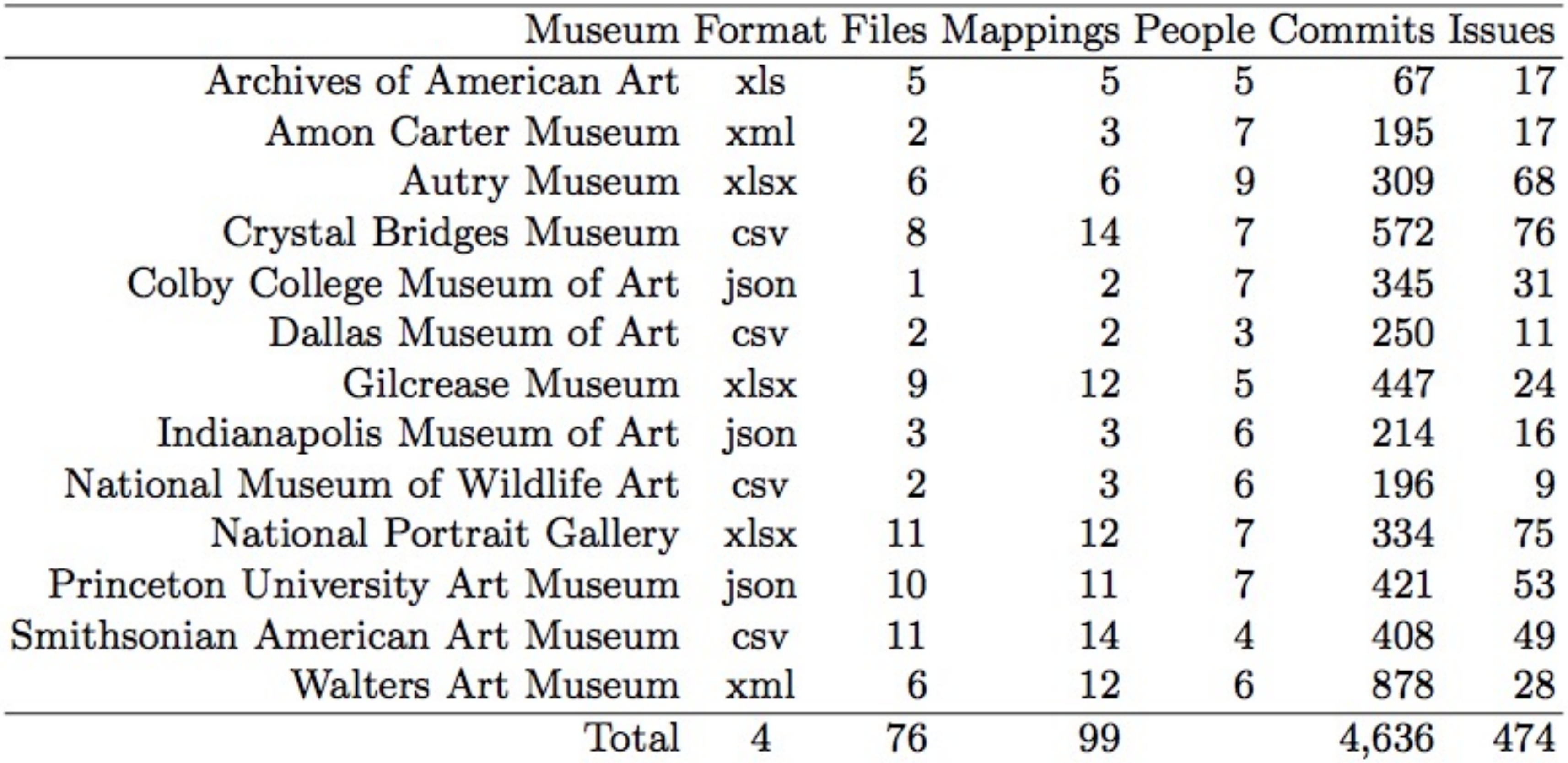
Example Model of Actor for Amon Carter



Complete Model of Actor for Amon Carter

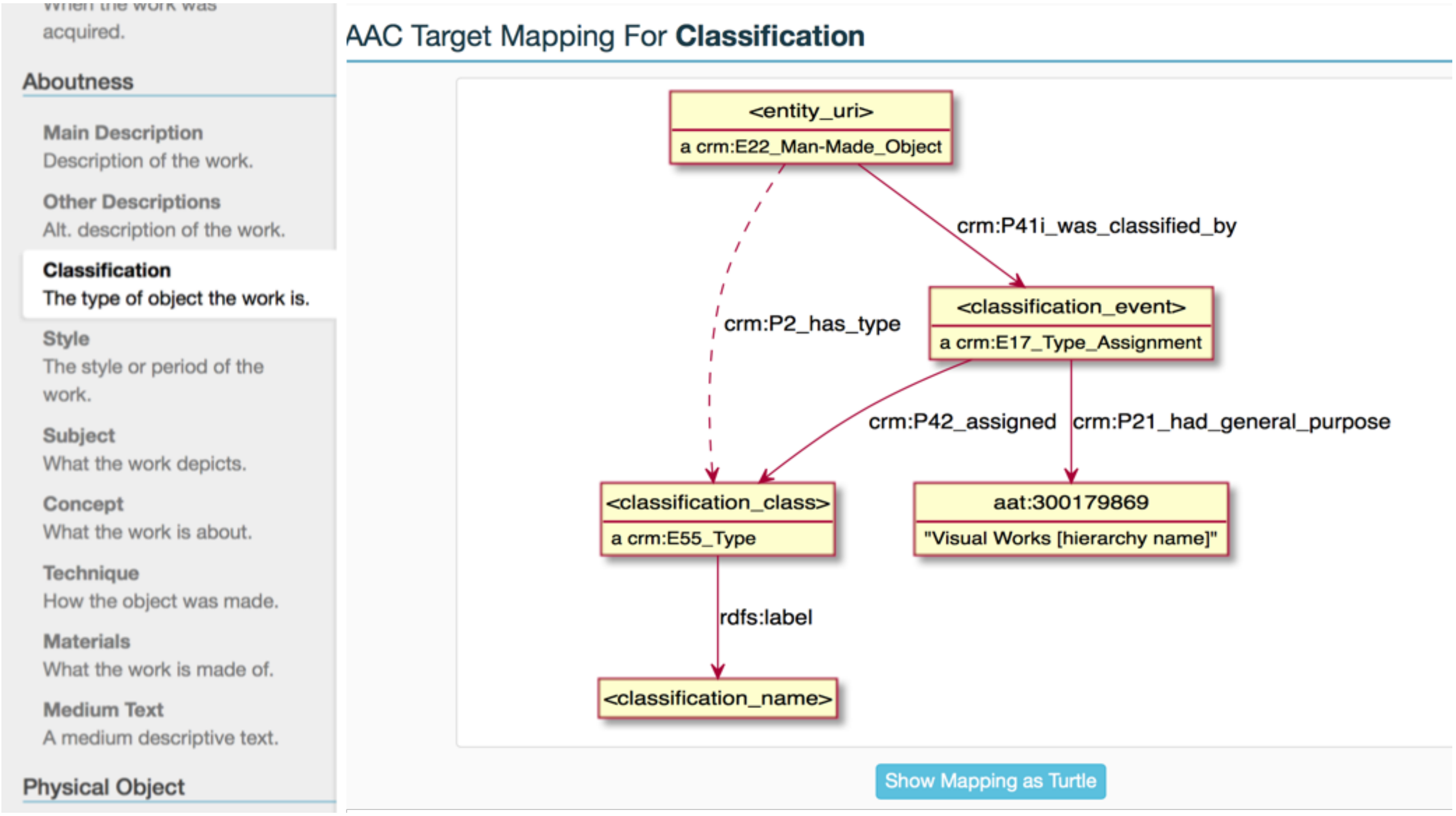


AAC Data Statistics



Museum	Format	Files	Mappings	People	Commits	Issues
Archives of American Art	xls	5	5	5	67	17
Amon Carter Museum	xml	2	3	7	195	17
Autry Museum	xlsx	6	6	9	309	68
Crystal Bridges Museum	csv	8	14	7	572	76
Colby College Museum of Art	json	1	2	7	345	31
Dallas Museum of Art	csv	2	2	3	250	11
Gilcrease Museum	xlsx	9	12	5	447	24
Indianapolis Museum of Art	json	3	3	6	214	16
National Museum of Wildlife Art	csv	2	3	6	196	9
National Portrait Gallery	xlsx	11	12	7	334	75
Princeton University Art Museum	json	10	11	7	421	53
Smithsonian American Art Museum	csv	11	14	4	408	49
Walters Art Museum	xml	6	12	6	878	28
Total	4	76	99		4,636	474

AAC Target Mappings



When the work was acquired.

Aboutness

Main Description
Description of the work.

Other Descriptions
Alt. description of the work.

Classification
The type of object the work is.

Style
The style or period of the work.

Subject
What the work depicts.

Concept
What the work is about.

Technique
How the object was made.

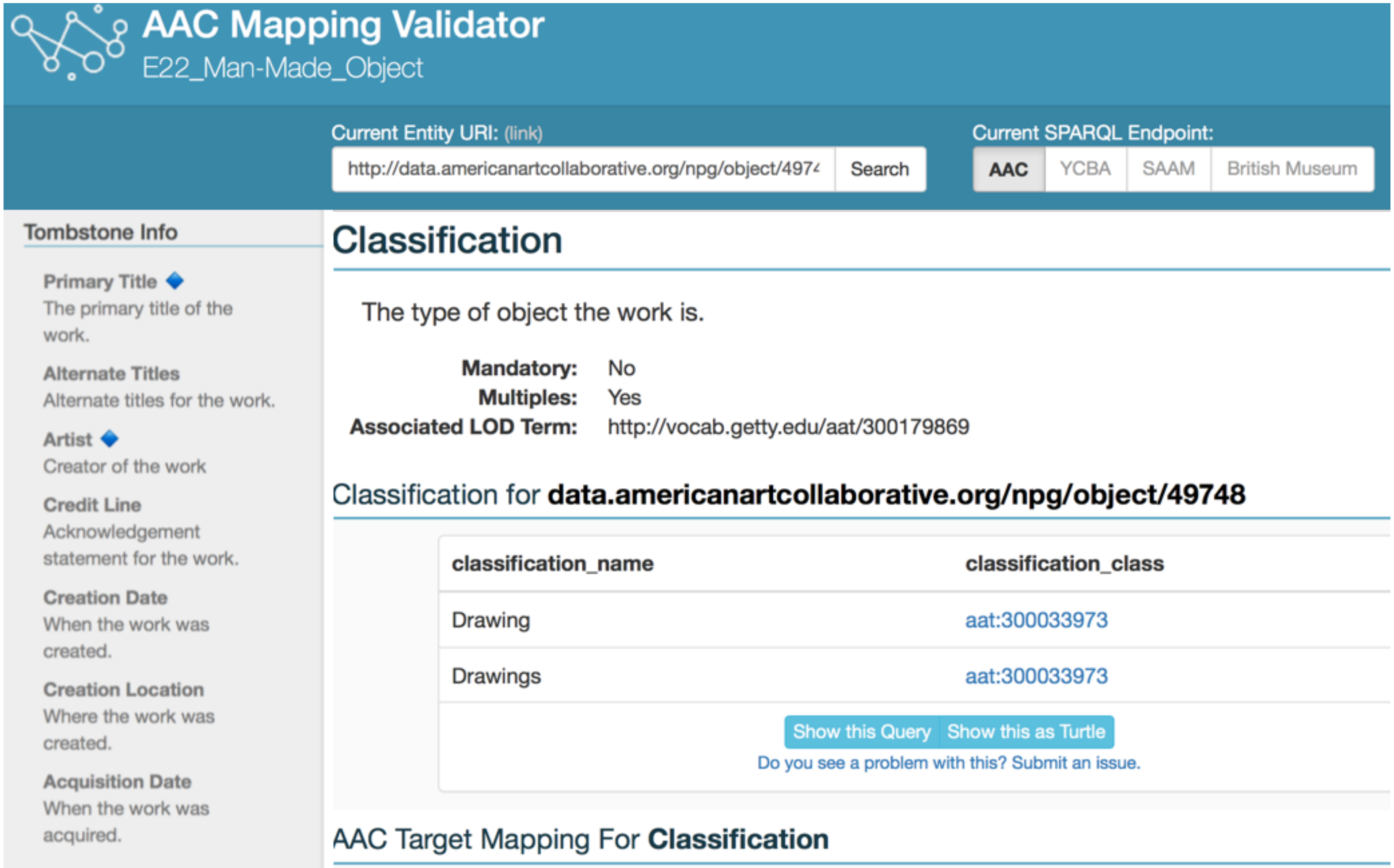
Materials
What the work is made of.

Medium Text
A medium descriptive text.

Physical Object

AAC Target Mapping For **Classification**

AAC Mapping Validator



AAC Mapping Validator

E22_Man-Made_Object

Current Entity URI: (link)

http://data.americanartcollaborative.org/npg/object/4974

Current SPARQL Endpoint:

Tombstone Info

Primary Title ◆
The primary title of the work.

Alternate Titles
Alternate titles for the work.

Artist ◆
Creator of the work

Credit Line
Acknowledgement statement for the work.

Creation Date
When the work was created.

Creation Location
Where the work was created.

Acquisition Date
When the work was acquired.

Classification

The type of object the work is.

Mandatory: No
Multiples: Yes
Associated LOD Term: http://vocab.getty.edu/aat/300179869

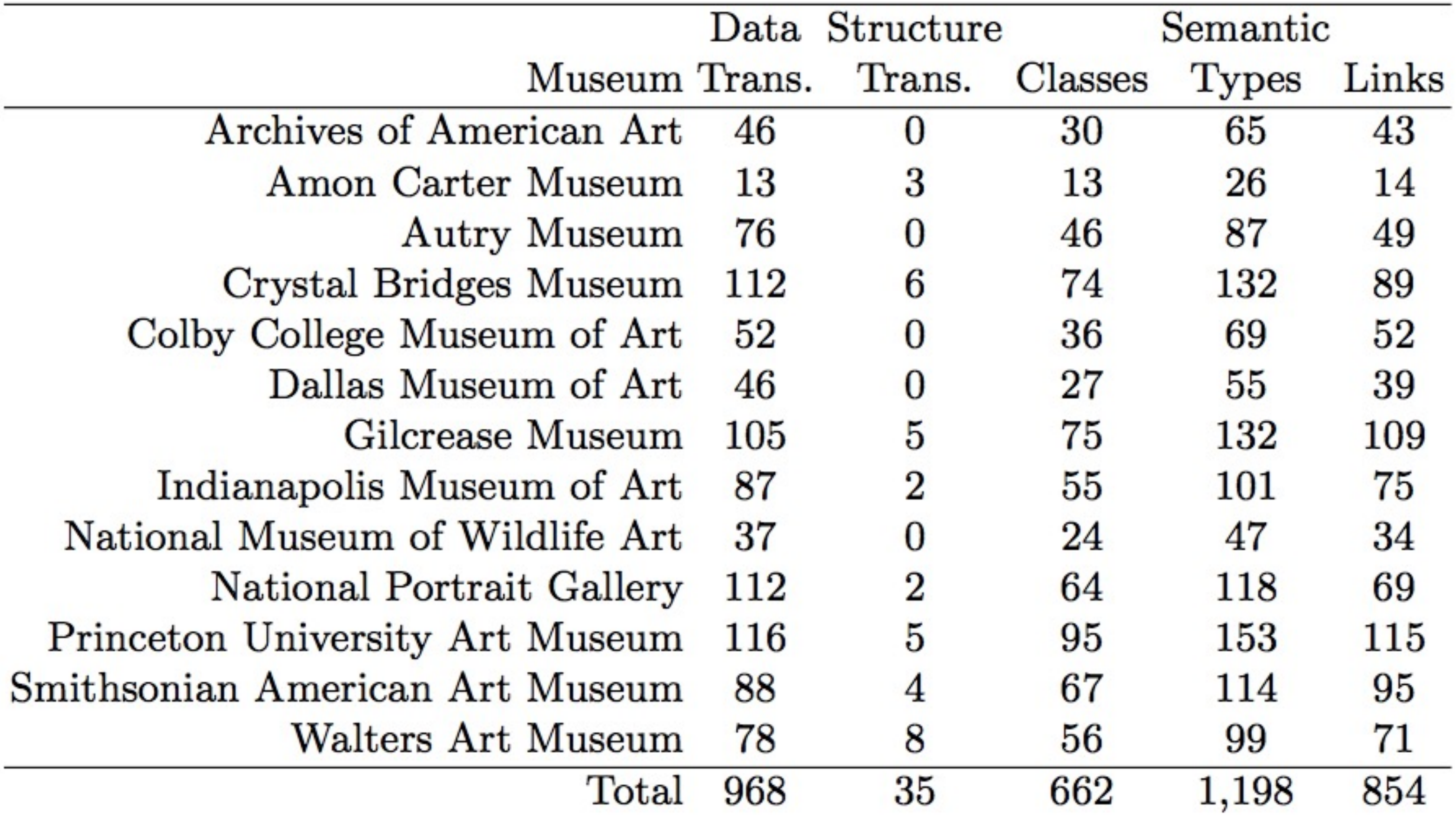
Classification for **data.americanartcollaborative.org/npg/object/49748**

classification_name	classification_class
Drawing	aat:300033973
Drawings	aat:300033973

Do you see a problem with this? Submit an issue.

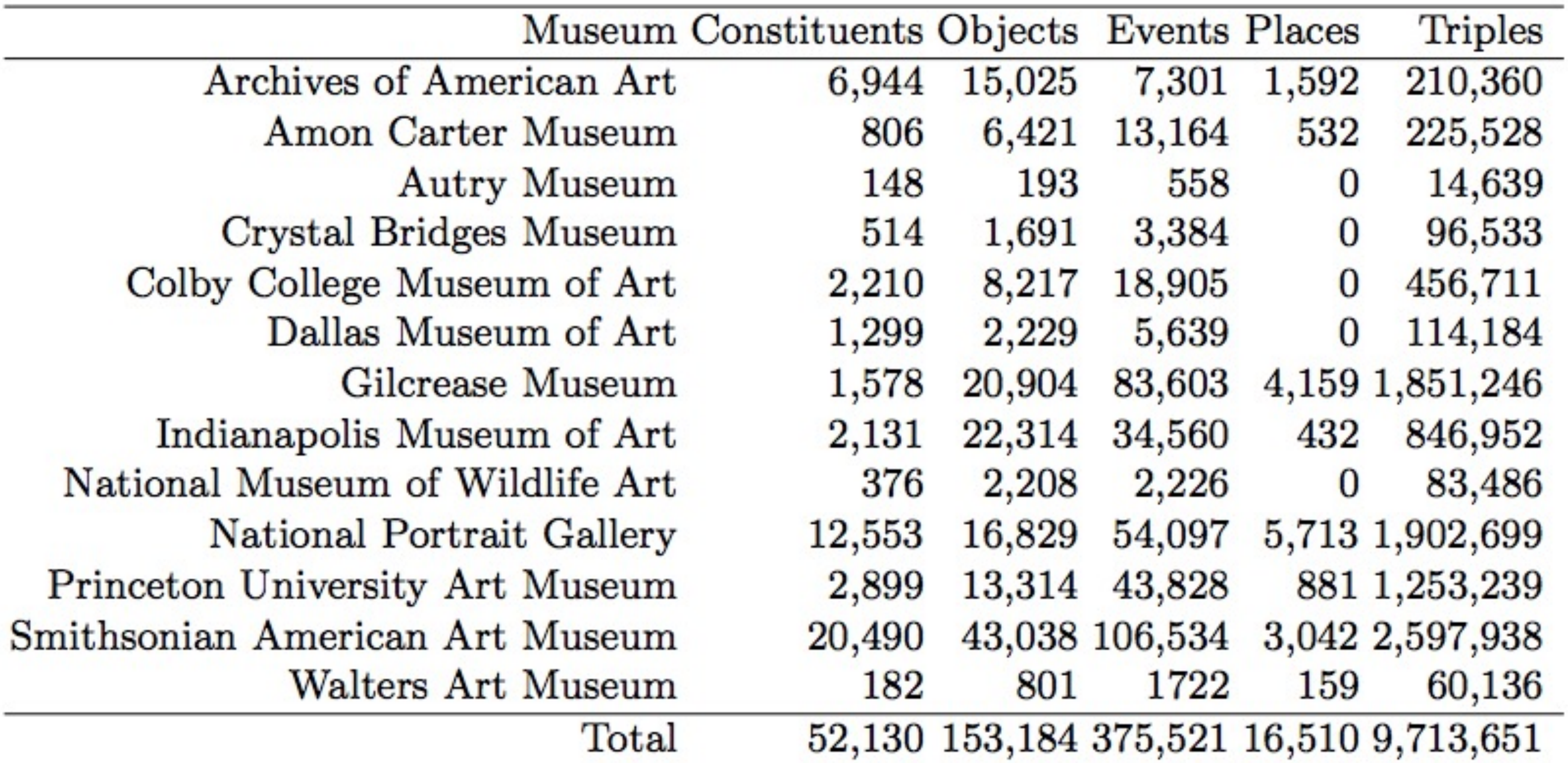
AAC Target Mapping For **Classification**

Statistics on the Mappings



	Data	Structure		Semantic	
Museum	Trans.	Trans.	Classes	Types	Links
Archives of American Art	46	0	30	65	43
Amon Carter Museum	13	3	13	26	14
Autry Museum	76	0	46	87	49
Crystal Bridges Museum	112	6	74	132	89
Colby College Museum of Art	52	0	36	69	52
Dallas Museum of Art	46	0	27	55	39
Gilcrease Museum	105	5	75	132	109
Indianapolis Museum of Art	87	2	55	101	75
National Museum of Wildlife Art	37	0	24	47	34
National Portrait Gallery	112	2	64	118	69
Princeton University Art Museum	116	5	95	153	115
Smithsonian American Art Museum	88	4	67	114	95
Walters Art Museum	78	8	56	99	71
Total	968	35	662	1,198	854

Statistics on What Was Mapped



Museum	Constituents	Objects	Events	Places	Triples
Archives of American Art	6,944	15,025	7,301	1,592	210,360
Amon Carter Museum	806	6,421	13,164	532	225,528
Autry Museum	148	193	558	0	14,639
Crystal Bridges Museum	514	1,691	3,384	0	96,533
Colby College Museum of Art	2,210	8,217	18,905	0	456,711
Dallas Museum of Art	1,299	2,229	5,639	0	114,184
Gilcrease Museum	1,578	20,904	83,603	4,159	1,851,246
Indianapolis Museum of Art	2,131	22,314	34,560	432	846,952
National Museum of Wildlife Art	376	2,208	2,226	0	83,486
National Portrait Gallery	12,553	16,829	54,097	5,713	1,902,699
Princeton University Art Museum	2,899	13,314	43,828	881	1,253,239
Smithsonian American Art Museum	20,490	43,038	106,534	3,042	2,597,938
Walters Art Museum	182	801	1722	159	60,136
Total	52,130	153,184	375,521	16,510	9,713,651

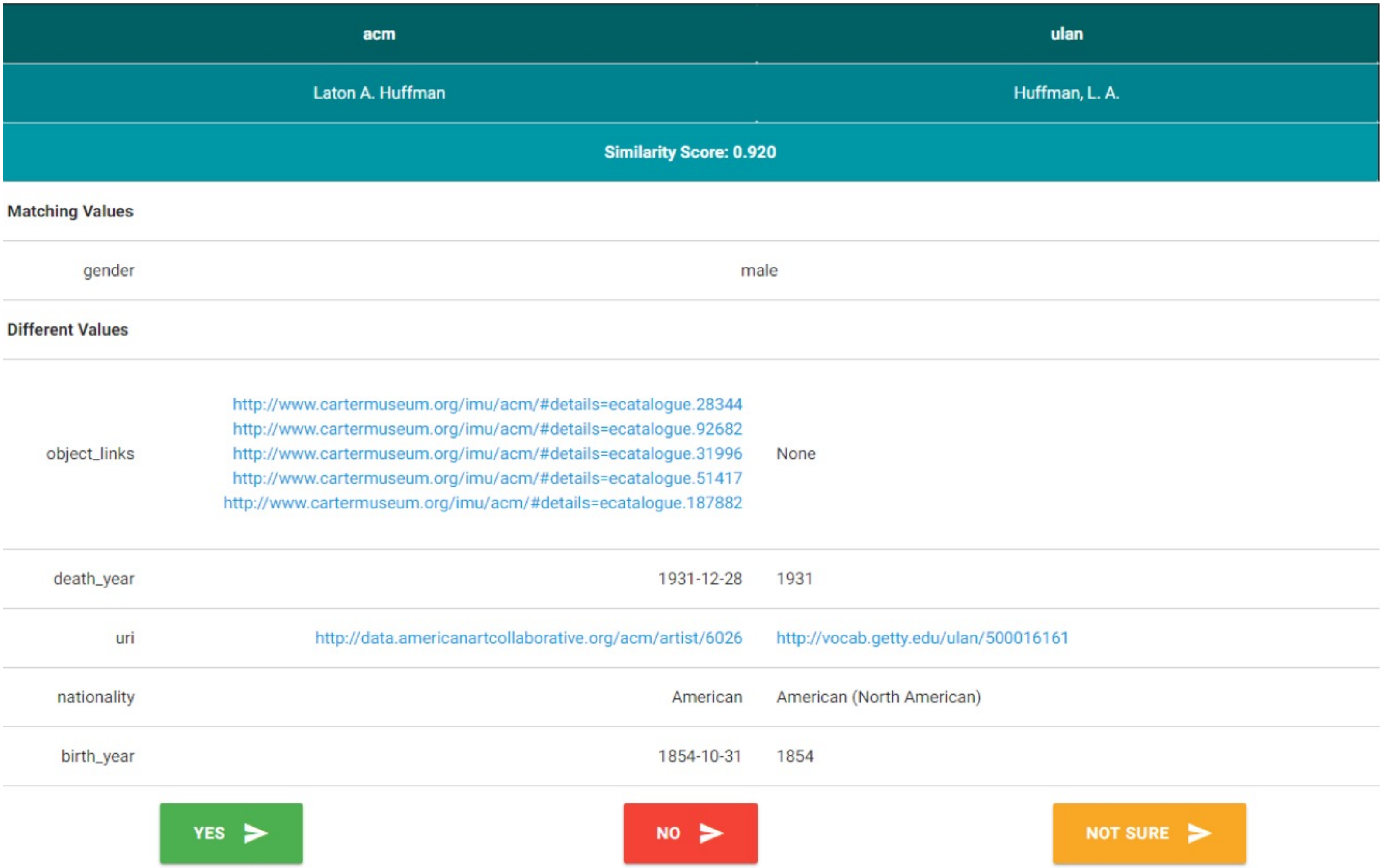
Mapping Lessons

- Lesson 1: Reproducible Workflows
  - Allow museums to export raw data from their collection management systems
- Lesson 2: Shared Repository
  - Github was invaluable for managing all the data and mapping files
- Lesson 3: Data Cleaning
  - Significant data cleaning was required
  - Integrated as part of the data processing workflow in Karma
- Lesson 4: Mapping Inconsistencies
  - Validation tool was critical to completing a consistent set of mappings
- Lesson 5: Expert Review
  - The outside review was crucial in identifying and resolving mapping inconsistencies

Linking the Data

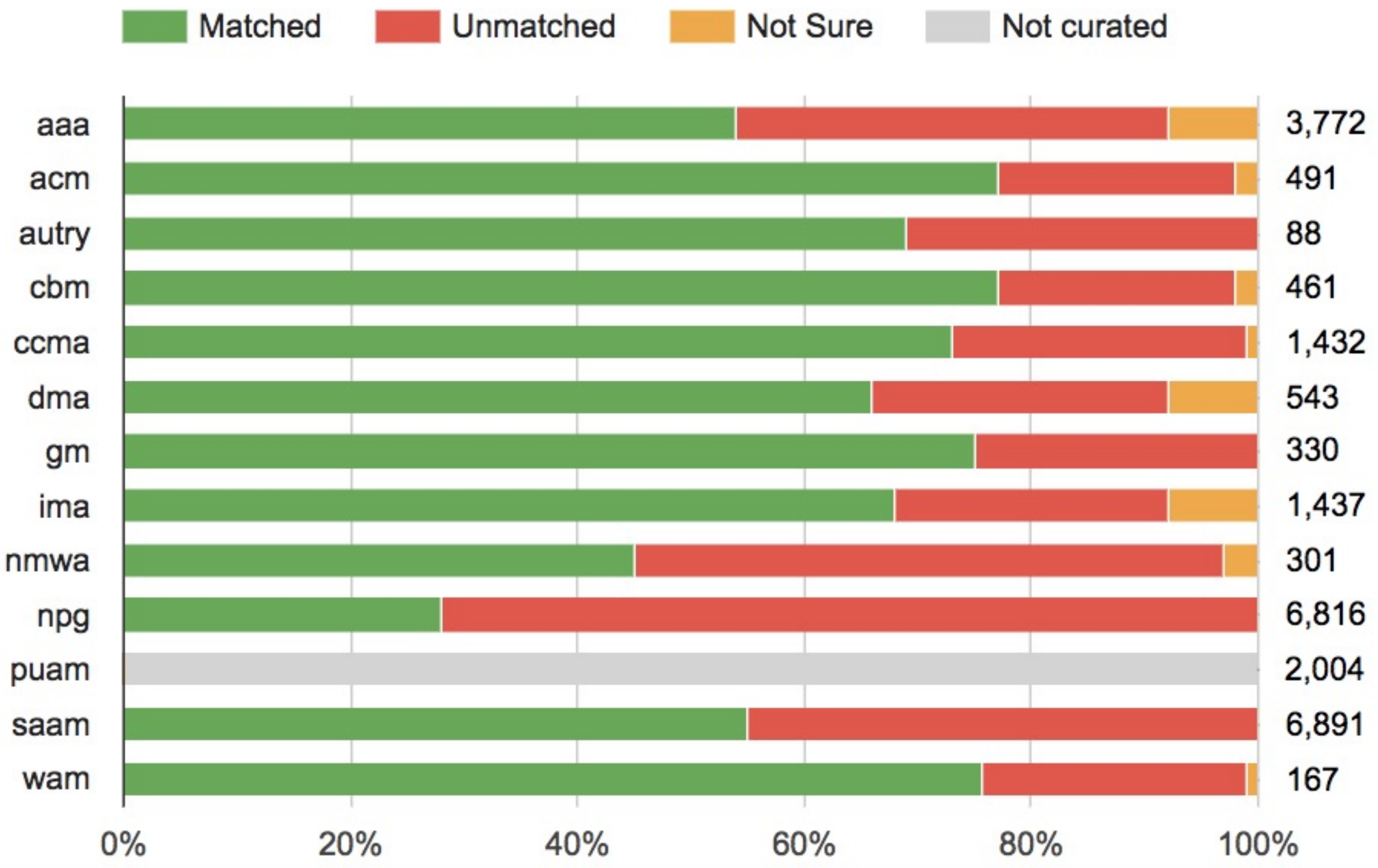
- Goals
  - Link the entities in the museum data to other resources
  - Capability for museums to curate the automatically generated links
  - Demonstration: linking artists to Getty ULAN
- Approach
  - Attempted to use existing linking tools, but they either didn’t scale or students found them difficult to configure
  - Wrote a specialized script to generate high recall & precision candidates
  - Built a link review tool that the museums used to curate their links

Link Review Screenshot



	acm	ulan
	Laton A. Huffman	Huffman, L. A.
	Similarity Score: 0.920	
Matching Values		
gender	male	
Different Values		
object_links	http://www.cartermuseum.org/imu/acm/#details=ecatalogue.28344 http://www.cartermuseum.org/imu/acm/#details=ecatalogue.92682 http://www.cartermuseum.org/imu/acm/#details=ecatalogue.31996 http://www.cartermuseum.org/imu/acm/#details=ecatalogue.51417 http://www.cartermuseum.org/imu/acm/#details=ecatalogue.187882	None
death_year	1931-12-28	1931
uri	http://data.americanartcollaborative.org/acm/artist/6026	http://vocab.getty.edu/ulan/500016161
nationality	American	American (North American)
birth_year	1854-10-31	1854

Screenshot with Linking Review Results



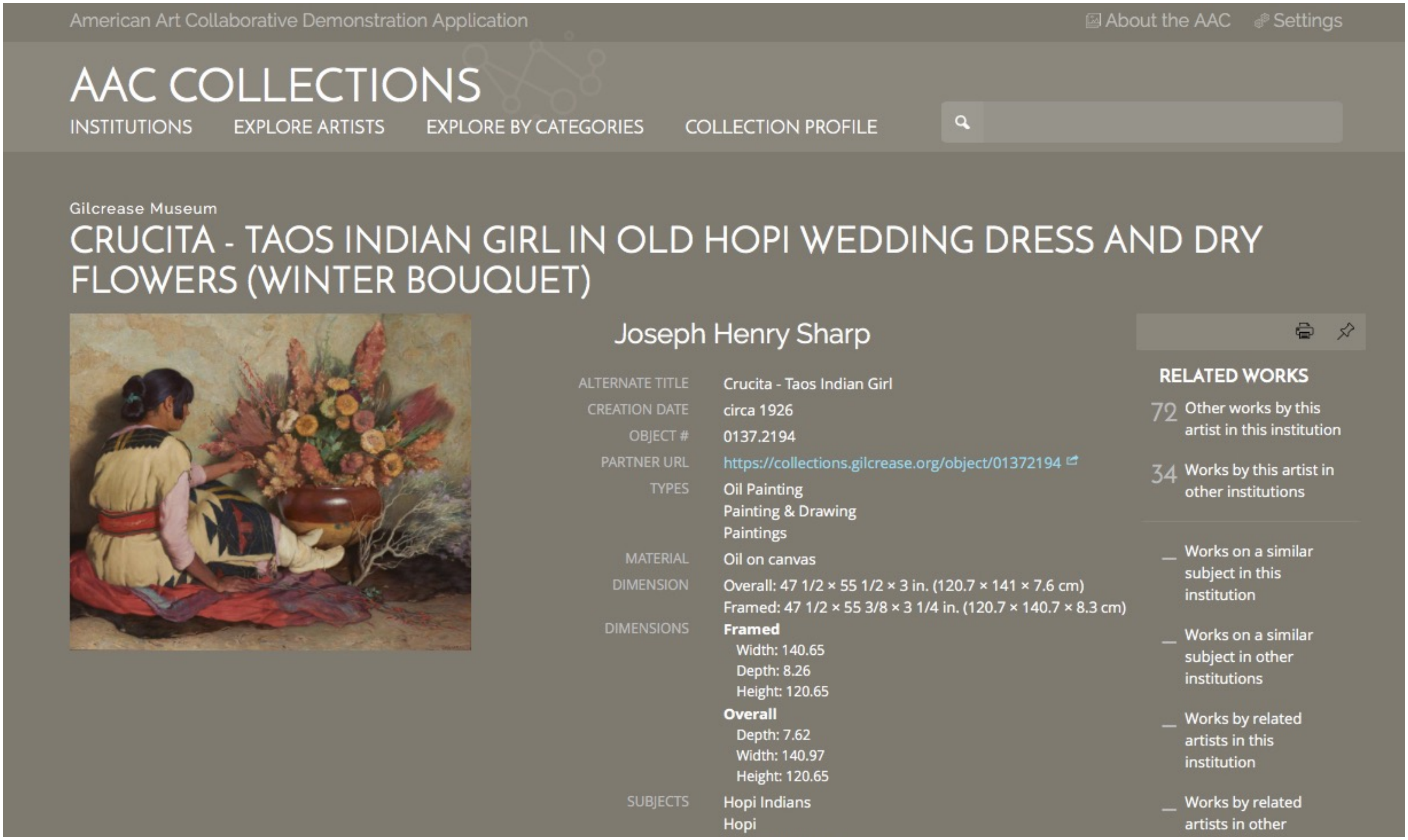
Summary Statistics on the Linking Process

- Number of constituents in data: 42,685
- Previously existing links to ULAN: 3,349
- Linking based on previously existing links:
  - Precision: .96
  - Recall: .88
  - F1-measure: .92
- Candidate matches: 24,733
- Incorrect links in museums datasets: 19
- New links to ULAN: 9,357
- Incorrect links after review: 2
- Previously existing links not discovered: 136

Lessons Learned on Linking

- Lesson 6: Linking Tools
  - Difficult to configure and use the existing linking tools and get them to scale to large datasets (e.g., ULAN, DBPedia, & VIAF)
  - We need easy to work with and scalable libraries for linking tasks
- Lesson 7: Manual Review
  - Users are willing to invest significant time and energy to ensure the final data is accurate
  - The museums reviewed almost 25K links!
  - A few weeks of effort almost tripled the number of links to ULAN

Using the Data: The Browse Application



Gilcrease Museum

CRUCITA - TAOS INDIAN GIRL IN OLD HOPI WEDDING DRESS AND DRY FLOWERS (WINTER BOUQUET)

Joseph Henry Sharp

ALTERNATE TITLE	Crucita - Taos Indian Girl
CREATION DATE	circa 1926
OBJECT #	0137.2194
PARTNER URL	https://collections.gilcrease.org/object/01372194
TYPES	Oil Painting Painting & Drawing Paintings
MATERIAL	Oil on canvas
DIMENSION	Overall: 47 1/2 × 55 1/2 × 3 in. (120.7 × 141 × 7.6 cm) Framed: 47 1/2 × 55 3/8 × 3 1/4 in. (120.7 × 140.7 × 8.3 cm)
DIMENSIONS	**Framed** Width: 140.65 Depth: 8.26 Height: 120.65 **Overall** Depth: 7.62 Width: 140.97 Height: 120.65
SUBJECTS	Hopi Indians Hopi

RELATED WORKS

72 Other works by this artist in this institution

34 Works by this artist in other institutions

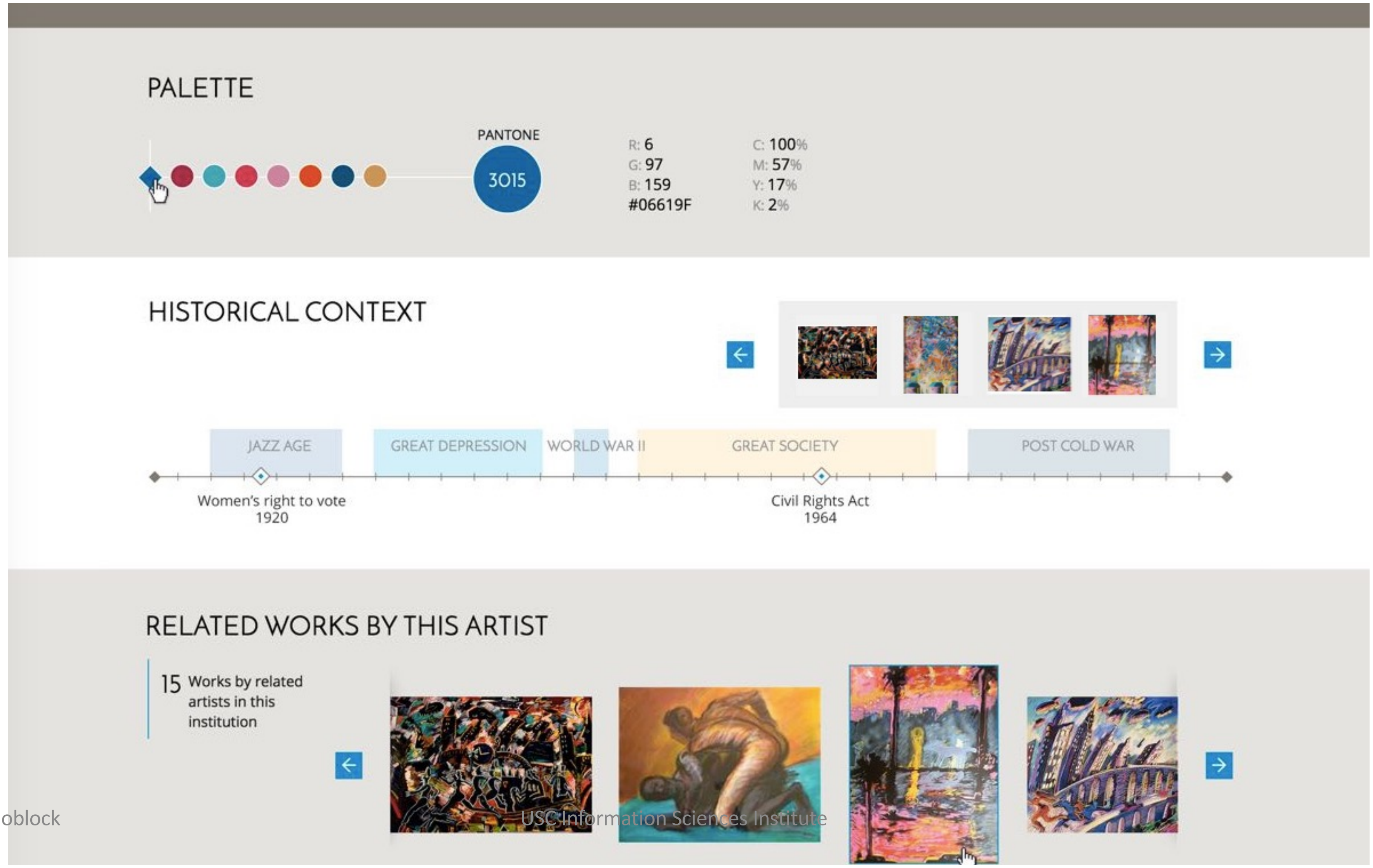
_ Works on a similar subject in this institution

_ Works on a similar subject in other institutions

_ Works by related artists in this institution

_ Works by related artists in other

Using the Data: Under Development



PALETTE

PANTONE 3015

R: 6
G: 97
B: 159
#06619F

C: 100%
M: 57%
Y: 17%
K: 2%

HISTORICAL CONTEXT

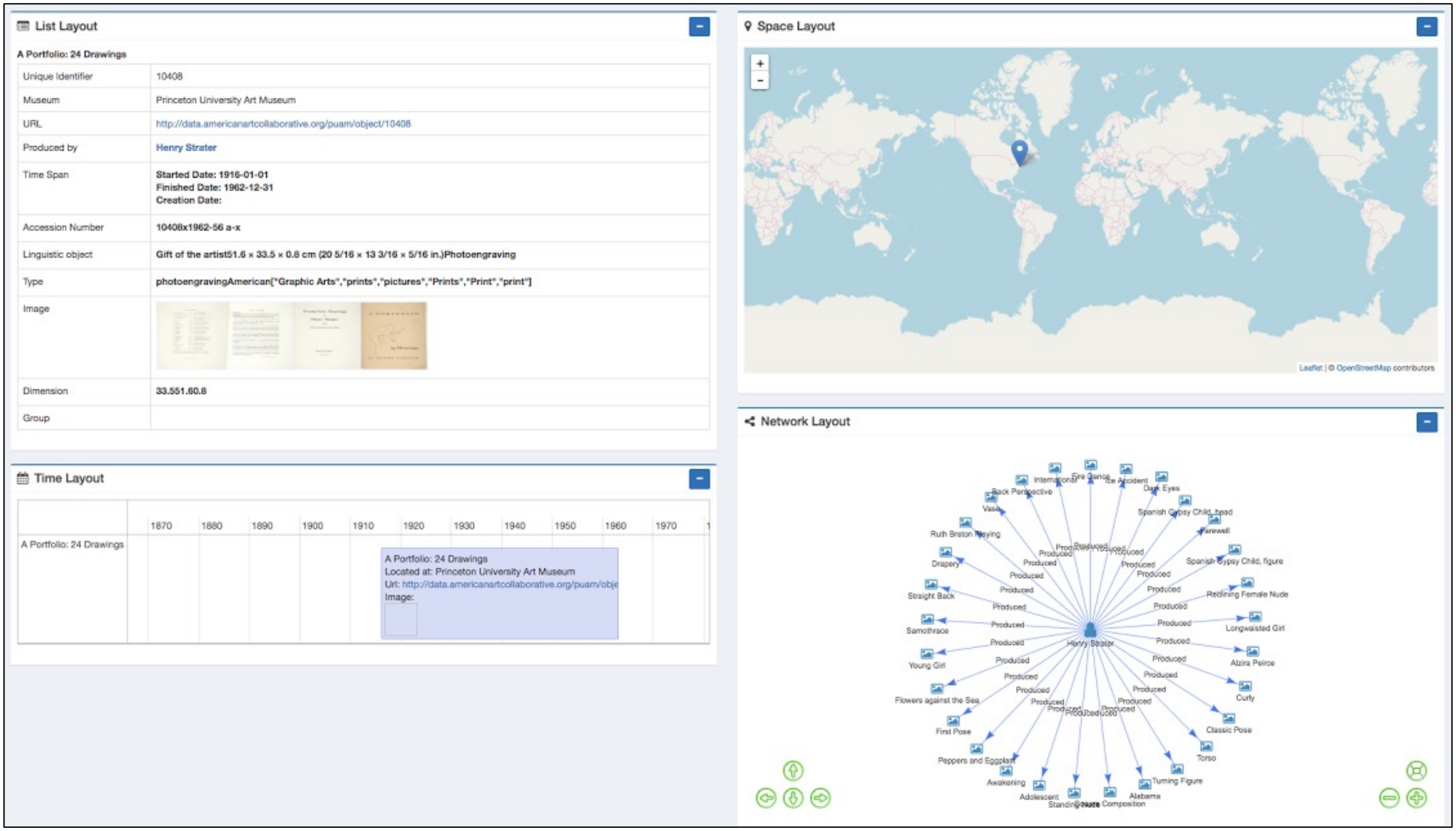
Women's right to vote
1920

Civil Rights Act
1964

RELATED WORKS BY THIS ARTIST

15 Works by related artists in this institution

SemUI Visualization [Giunchiglia, Ojha, & Das, ICSC 2017]



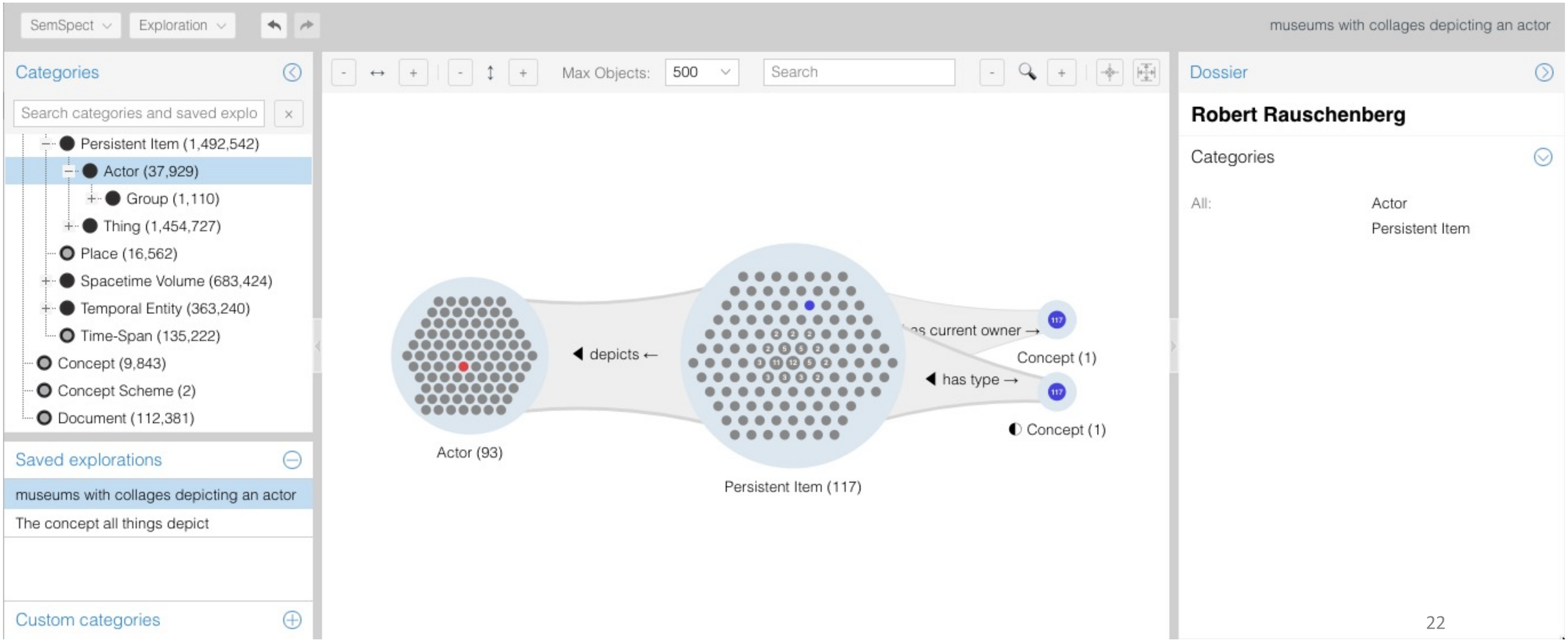
A Portfolio: 24 Drawings

Unique Identifier	10408
Museum	Princeton University Art Museum
URL	http://data.americanartcollaborative.org/puam/object/10408
Produced by	Henry Strater
Time Span	Started Date: 1916-01-01 Finished Date: 1962-12-31 Creation Date:
Accession Number	10408x1962-56 a-x
Linguistic object	Gift of the artist51.6 × 33.5 × 0.8 cm (20 5/16 × 13 3/16 × 5/16 in.)Photoengraving
Type	photoengravingAmerican["Graphic Arts","prints","pictures","Prints","Print","print"]
Image	
Dimension	33.551.60.8
Group	

A Portfolio: 24 Drawings

A Portfolio: 24 Drawings
Located at: Princeton University Art Museum
Url: http://data.americanartcollaborative.org/puam/obje
Image:

SemSpect (http://aac.semspect.de/)
Thorsten Liebig (derivo)



museums with collages depicting an actor

Max Objects: 500

Robert Rauschenberg

Categories

All: Actor, Persistent Item

Saved explorations:
- museums with collages depicting an actor
- The concept all things depict

Lessons Learned in Using the Data

- Lesson 8: Visualization
  - Easy to understand visualization is needed for non-technical users

- Lesson 9: Simple Schema
  - CRM ontology may be useful for research, but challenging for applications
  - Created a set of SPARQL queries to create JSON objects
    - Loaded the objects into Elasticsearch for complex analysis

Related Work

- Consortiums of museums
  - Europeana – 1500 cultural heritage institutions, 17 million items
  - CHIN – 8 Canadian museums, 85,000 items
  - LODAC – 114 museums in Japan
  - Published a fixed schema and mapped all museums/institutions to the schema
- Organizations using the CRM ontology
  - Research space (British museum & Yale Center for British Art)
  - Pharos – 14 historical photo archives
  - Each organization is responsible for publishing their own data to CRM
- Mapping data to CRM
  - X3ML – maps XML data to CRM using manually written rules
- Linking
  - Silk, Dedup, etc. – focus is on automatic linking, but no curation of the links
  - Mix’n’match, OpenRefine – support link review, but targets highly technical users

Discussion

- Collaborated with 14 American art museums to publish 5* Linked Data
- Created a set of tools to create the linked data
  - Karma – clean and map the data, publish directly to Gitub
  - Mapping Validation tool – review the mappings to ensure consistency
  - Karma execution tool – applies Karma mappings and published both RDF and JSON-LD directly to Github
  - Link Review tool – allows non-technical users to quickly and easily review links to other sources
  - Browse application – allows museum staff, art historians, and the general public to verify and explore the data

Success?

- 14 additional museums have now released their data as linked data
- Three museums have already learned how to use the tools to create their own mappings
  - Indianapolis Museum of Art
  - Smithsonian Archives of American Art
  - Colby College Museum of Art
- Researchers outside the project have applied their visualization tools to the data
  - Sajan Raj Ojha (Univ. of Trento): SemUI
  - Thorsten Liebig (derivo): SemSpect http://aac.semspect.de/

Future Work

- Automate the addition of new museums to the AAC
  - Gather, map, and link the data directly form their online web pages
- Extend the types of information supported
  - E.g., exhibition data & bibliographies
  - Improve the ability in Karma to automate complex mappings
- Link the existing data to other sources
  - E.g., VIAF, Geonames, & DBpedia
  - Build a library of linking functions to support easy and scalable linking

More Info

Karma: karma.isi.edu

AAC: americanartcollaborative.org

Github: github.com/american-art

- Thanks to the Mellon Foundation & Institute of Museum and Library Services for their financial support of this project

Thanks!