

Lessons Learned in Building Linked Data for the American Art Collaborative

Craig A. Knoblock¹, Pedro Szekely¹, Eleanor Fink², Duane Degler³,
David Newbury⁴, Robert Sanderson⁴, Kate Blanch⁵, Sara Snyder⁶,
Nilay Chheda¹, Nimesh Jain¹, Ravi Raju Krishna¹,
Nikhila Begur Sreekanth¹, and Yixiang Yao¹

¹ University of Southern California

² American Art Collaborative

³ Design for Context

⁴ J Paul Getty Trust

⁵ The Walters Art Museum

⁶ Smithsonian American Art Museum

Abstract. Linked Data has emerged as the preferred method for publishing and sharing cultural heritage data. One of the main challenges for museums is that the defacto standard ontology (CIDOC CRM) is complex and museums lack expertise in semantic web technologies. In this paper we describe the methodology and tools we used to create 5-star Linked Data for 14 American art museums with a team of 12 computer science students and 30 representatives from the museums who mostly lacked expertise in Semantic Web technologies. The project was completed over a period of 18 months and generated 99 mapping files and 9,357 artist links, producing a total of 2,714 R2RML rules and 9.7M triples. More importantly, the project produced a number of open source tools for generating high-quality linked data and resulted in a set of lessons learned that can be applied in future projects.

Keywords: Linked Data, data mapping, linking, lessons learned

1 Introduction

There is growing interest in Linked Open Data (LOD) among museums and the cultural heritage sector. In recent years it has gained traction because most museums are interested in using technology to reach new audiences, collaborate with other museums, deepen research, and help audiences of all ages experience, learn about, appreciate, and enjoy art. In fact, these concepts and others that characterize features of LOD inspired 14 art museums to form a collaborative to learn about and implement LOD within their respective museums and set the stage for the broader art-museum community to explore LOD.

The goals of the American Art Collaborative (AAC)⁷ are to learn about LOD; create and publish a critical mass of LOD drawn from the collections of the 14

⁷ <http://americanartcollaborative.org/>

museums that will be made available on the Internet for researchers, educators, developers, and the general public; test LOD reconciliation methods; develop open-source production and reconciliation tools; demonstrate the value of LOD through a prototype browse application; and publish good practices guidelines to help the broader museum community learn about and implement LOD.

Towards these goals, we built 5-star Linked Data (actually, we built 7-star linked data [1], which is 5-star data with an explicit meta-data schema and data validation) for 13 museums⁸ by applying existing tools and developing new tools where needed to map and link the data. The project involved a number of different communities of users: about 30 representatives from the various museums, very knowledgeable about art, but inexperienced in Linked Data or ontologies, 5 Semantic Web experts who provided guidance and direction on the project, 12 USC students, inexperienced in art and the Semantic Web, who both helped develop new tools and applied the tools to the provided data, and 3 experts in the CIDOC CRM ontology who reviewed the data mappings at various stages of the project. In every stage of the project, all of these different communities were engaged in one way or another.

The three main thrusts of the project are mapping the data to a common cultural heritage ontology, linking the data to other resources, and then using the results to allow users to explore the data. For mapping the data, the AAC chose the CIDOC Conceptual Reference Model (CRM)⁹ as its ontology. The CRM (ISO 21127:2006) is an extensive cultural heritage ontology containing 82 classes and 263 properties, including classes to represent a wide variety of events, concepts, and physical properties. The USC students used Karma [2] to align the museum data to the ontology. Given the complexity of the CRM ontology, the AAC project developed a Mapping Validation tool to guide the students in performing the mapping and validating them using queries to the actual data. For linking, the project focused on linking the artists to the Getty Union List of Artist Names (ULAN), a widely-used knowledge base of artists. To ensure high-quality links, we developed a link review tool that allowed museum representatives to review candidate links to make a final decision about whether each entity was the same as an entity listed in ULAN. To use the data, we developed an application that allows a user to explore the data by museum, artist, or artwork.

In the remainder of this paper, we will present the details of each thrust of the project (mapping, linking, and using) along with the lessons learned. We also compare the project to other related work and conclude with a discussion of the results, impact, and future work.

2 Mapping the Data

Managing the data: The first step in the project was to map the data from each museum to the CRM ontology. Since each museum had as many as 14 data

⁸ The Yale Center for British Art had already mapped and linked their data, so we only needed to build Linked Data for 13 of the 14 museums in the AAC.

⁹ <http://www.cidoc-crm.org/>

Table 1. The AAC mapping process

| | 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 | | |

files, all in different formats, just managing the data from the 13 museums was a challenge. We addressed the data management problem by using the GitHub source control system to manage all project data. We set up one repository for each museum and taught the museums how to upload their data. In addition, each GitHub repository organizes and stores all the resources associated with each data set: the mappings that specify how each data set is mapped to the CRM ontology; visualizations of the mappings that enable non-technical museum personnel to review the mappings; the resulting RDF data that is then loaded into the triplestore; and the issues identified by CRM experts, and discussions that led to their resolution.

Table 1 provides the details of the use of GitHub for the data that we received from the museums, including the format of the data provided, the number of files (we only counted the ones we actually mapped), the number of mappings (each file can have more than one mapping to different classes), the number of people involved in creating and refining the mappings, the total number of GitHub commits, and the number of issues identified and discussed on the GitHub issue tracker. Note that all of this information is available online.¹⁰

Mapping the data: After the raw data was uploaded to GitHub, the next and most difficult challenge, was mapping the data to the CRM domain ontology. This was challenging for several reasons. First, there was a lot of data, much more than we originally expected since it included data about artists, artwork, exhibitions and bibliographies, as well as collection data from the Smithsonian Archives of American Art. Second, the CRM ontology is very complicated and requires significant expertise to understand and use. Third, the pool of students that we had available to work on this project were skilled undergraduate and masters students in computer science, but they were not experts in the Semantic Web, cultural heritage data, or the CRM ontology.

¹⁰ <http://github.com/american-art>

In previous work, we developed the Karma information integration system, a semi-automated tool for mapping data sources to a domain ontology [2, 3]. Karma supports a wide variety of source types, has a machine learning capability to provide recommendations on the mappings to an ontology, and has an intuitive graphical interface for visualizing and refining mappings. Figure 1 provides a fragment of a screen shot of the use of Karma to map one of the datasets to the CRM ontology. In the screenshot, the graph showing the mapping is shown at the top of the figure, the attribute names and an analysis of the distribution of the data for that attribution are shown in the blue and green rectangles, and the data is shown at the bottom.

After completing a mapping or updating it to address an issue, users can publish the mapping (R2RML file) and associated resources (report of all data transformations and visualization of the mapping) to GitHub using Karma. Figure 2 shows a visualization of a mapping. In addition, the R2RML mapping is applied to the raw data to create RDF triples, which are subsequently loaded into the triplestore and posted on GitHub.

We started the mapping process with a team of USC computer science students in January 2016. The students quickly became proficient in Karma and worked closely with a local CRM expert to begin the mapping process for the data from the National Portrait Gallery. Because of the complexity of the CRM ontology, it took several iterations to create a mapping that satisfied the expert. Other students worked on mapping the data from other museums and by the end of the spring semester 2016, we had built mappings for a half a dozen museums.

Review by a different CRM expert revealed many issues with the mappings. Some issues resulted from inconsistencies in the mappings produced by different students. Interestingly, some issues revealed disagreements among CRM experts who had previously worked together. Students updated the mappings according to guidance provided in the discussion forum associated with each issue. The

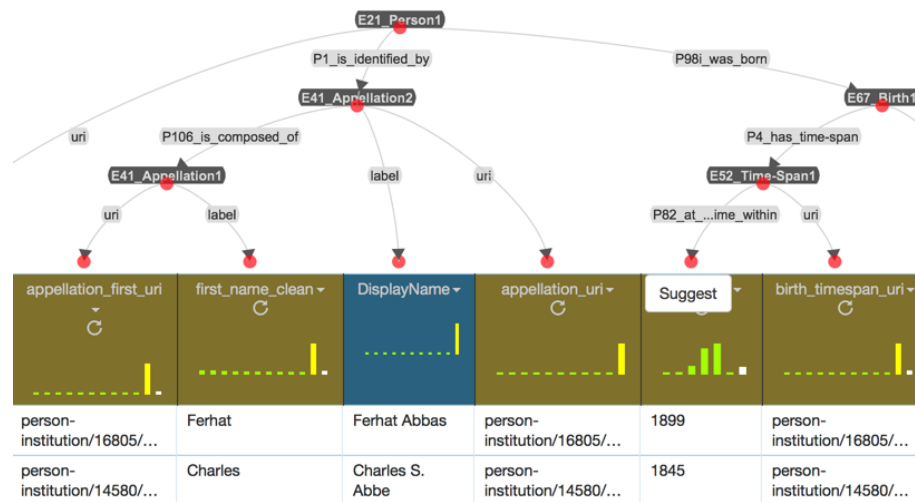


Fig. 1. Screen shot of Karma building a mapping of the National Portrait Gallery data

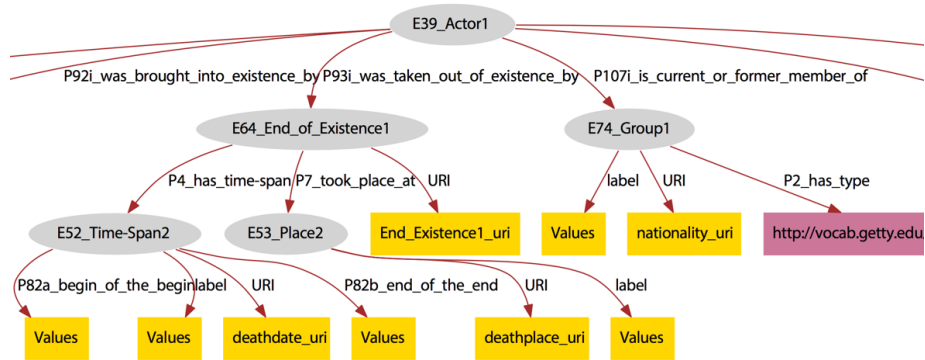


Fig. 2. Mapping of artist data from the Amon Carter Museum to the CRM ontology

process was slowed-down by lengthy discussions among the experts on the correct way to map various attributes (e.g., several issues had over 20 replies by different CRM experts). After a significant number of weeks spent on numerous revisions of the mappings and lack of convergence, we decided to suspend the mapping process until the CRM experts could provide clear and consistent guidance.

Validating the mappings: To address the challenges in creating consistent and correct mappings, we developed the AAC Mapping Validator,¹¹ (Figure 3). In this tool, two of our Semantic Web experts defined a target mapping for each of the relevant pieces of information from the museums. The figure shows the target mapping for Classification. The Mapping Validator implements this target mapping as a query, which can then be run against any one of several SPARQL Endpoints. The upper half of the figure shows that the AAC Endpoint has been selected, and a specific object from the National Portrait Gallery is specified. The tool runs that SPARQL query for the target mapping against the AAC triplestore and displays the result for the specified object (#49748).

The validation tool led to a dramatic improvement in the efficiency of the mapping process. The validation tool diagram showed students how to map the data, and the query enabled students to test their mappings after loading the RDF in the triplestore.

At this point, a year into the project, more than half of the datasets had been mapped twice. For many attributes, the templates defined in the validation tool required mapping the data a third time. In order to meet the project deadlines, we recruited a team of 6 M.S. students to participate in an intense 2-week “Karma-fest” after their final exams. Using the validation tool for guidance, and Karma to build the mappings, the students re-mapped the datasets for 12 of the 13 AAC museums. The two museums that were left out were the Archives of American Art, whose data was very different from the other museums since it is an archive, and the Yale Center for British Art, which had already mapped their data to the CRM. Then in the spring 2017 we had one student

¹¹ <http://review.americanartcollaborative.org>

The screenshot displays the AAC Mapping Validator interface for the object `data.americanartcollaborative.org/npg/object/49748`. The interface is divided into several sections:

- Tombstone Info:** A sidebar on the left containing metadata fields such as Primary Title, Alternate Titles, Artist, Credit Line, Creation Date, Creation Location, Acquisition Date, Main Description, Other Descriptions, Style, Subject, Concept, Technique, Materials, Medium Text, and Physical Object.
- Classification:** A central section titled "Classification" with the description "The type of object the work is." It lists properties: **Mandatory:** No, **Multiples:** Yes, and **Associated LOD Term:** `http://vocab.getty.edu/aat/300179869`.
- Classification for data.americanartcollaborative.org/npg/object/49748:** A table showing specific classifications:

| classification_name | classification_class |
|---------------------|----------------------|
| Drawing | aat:300033973 |
| Drawings | aat:300033973 |
- AAC Target Mapping For Classification:** A diagram illustrating the semantic relationships between classes. It shows:
 - `<entity_uri>` (a `crm:E22_Man-Made_Object`) connected to `<classification_event>` (a `crm:E17_Type_Assignment`) via `crm:P411_was_classified_by`.
 - `<entity_uri>` connected to `<classification_class>` (a `crm:E55_Type`) via `crm:P2_has_type`.
 - `<classification_event>` connected to `<classification_class>` via `crm:P42_assigned`.
 - `<classification_event>` connected to `aat:300179869` ("Visual Works [hierarchy name]") via `crm:P21_had_general_purpose`.
 - `<classification_class>` connected to `<classification_name>` via `rdfs:label`.

Fig. 3. The Mapping Validator specifies the target mapping and queries the triplestore based on the mapping to verify that the mapping is done correctly

refine the mappings based on a new review, complete some missing pieces, and map the Archives data. Table 2 shows the details of mappings created for each museum, including the number of data transformations, structure transformations, classes, semantic types, and links between classes. Table 3 provides details of the data produced by the mappings, including the number of constituents (e.g., artists), object, events, places, and the total number of triples.

In this process, we learned a number of important lessons:

Lesson 1 - Reproducible Workflows To enable construction of reproducible workflows, allow museums to submit the raw data exported from their

collection management systems, and implement the necessary data cleaning as part of the mapping workflows. We found line-based data formats, such as JSON Lines, CSV, or XLS, to be much easier to work with compared to large document formats, such as XML or JSON dictionaries.

Lesson 2 - Shared Repository: GitHub proved invaluable for managing multiple data submissions from museums, multiple versions of the mappings and associated resources, and the issues raised during the mapping process. Karma was extended to support the GitHub-based workflow, providing a one-click *publish mapping* command to publish into Github the R2RML file along with a visualization of the mapping, which is automatically created using Graphviz.¹²

Lesson 3 - Data Cleaning: The data submitted was of varied quality because most museum data has legacy data issues that have not been resolved for decades. For example, museums do not consistently record dates, dimensions, or have consistent ways of referring to an "unknown" work of art. A significant amount of data cleaning is necessary to produce high quality RDF. Karma supports arbitrary data transformations as part of the mapping process (using Python scripts), which made it possible to address an open-ended set of data cleaning scenarios.

Lesson 4 - Mapping Inconsistencies: Even though the CRM is a very prescriptive ontology, different CRM experts may map the same data differently, making it difficult to write reliable SPARQL queries. A template-based validation tool makes it easy to enforce consistency.

Lesson 5 - Expert Review: A formal review process by an outside consultant was very effective in identifying and resolving problems and inconsistencies in the mapping of the data. The USC students conducting the mapping were

¹² <http://www.graphviz.org>

Table 2. The AAC mappings

| | Data Structure | | Semantic | | |
|---------------------------------|----------------|--------|----------|---------|-------|
| | Museum | Trans. | Trans. | Classes | Types |
| 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 |

Table 3. The results of applying the mappings

| 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 |

not art experts or CRM experts and at times made assumptions that did not work for the museums or were incorrect mappings, but were not identified by the validation tool.

3 Linking and Reviewing the Data

An important aspect of producing high quality Linked Data is to link subjects to external datasets. In the cultural heritage community, the Getty Union List of Artist Names (ULAN) is an authoritative reference dataset, containing over 650,000 names for over 160,000 artists. The goal of our linking effort was to discover links to ULAN for artists in the museum datasets.

Museums take enormous pride on the quality of their data, so they want 100% correct links. They were willing to manually review *every* link before publication, so we developed a workflow where an automated algorithm first proposes links (pairs of museum actors and ULAN artists), and a human curator verifies each link. Given the large number of artists (52,130), the review effort is significant, so it was important to use a high precision algorithm to propose candidate links for review. A natural approach was to use existing tools, as the students working on the project are not experts in entity resolution algorithms.

We explored several different approaches to generate candidate links. We first assigned a student to work with Dedupe,¹³ a popular entity resolution library based on Bilenko’s work [4]. Initial results with a subset of the data were good, but then we ran into problems getting the algorithm to scale to the full ULAN dataset, running out of memory. Next we assigned a student to use SILK [5], a popular entity resolution tool. But the student struggled to configure the software to generate good results and we decided to abandon this approach after several weeks of effort.

¹³ <https://github.com/dedupeio/dedupe>

| acm | ulan | |
|-------------------------|---|---|
| Leton 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 |

YES >
NO >
NOT SURE >

Fig. 4. Screenshot of the Link Review Tool

We then assigned a student to implement a simple blocking scheme using birth year and the first two characters of the names (for records that didn't include birth year). He compared the names using Hybrid Jaccard with Jaro Winkler string similarity. Reusing software found on GitHub, he was able to implement this simple algorithm in a couple of weeks and tune the similarity thresholds in a few more weeks. The final algorithm uses different thresholds for records where birth or death dates are available, using a stricter string comparison threshold for records without birth or death years. Although not efficient, the algorithm produces links for the entire dataset in 20 hours running on a laptop.

The automated algorithm produced 24,733 links that needed to be reviewed by museum personnel. Some museums wanted links to be independently reviewed by more than one person and published if at least two reviewers approved them. We developed a generic link review tool optimized to support efficient and accurate comparison of pairs of records. The tool (Figure 4) requires the two datasets to be represented in the same schema to enable building a simple card that shows values side by side. Each row shows the values of one record field, placing the values side-by-side to support rapid assessment. The card segregates fields with identical values from fields with different values so users can quickly see the differences. When multiple candidates exist for a single record, the tool shows all cards for the record in a single page. Even though the number of links is large, all museums reviewed their links in a week or two, sometimes with multiple personnel conducting the review. There was a range of time reported for

reviewing candidate links, ranging from 18 to 51 seconds to review each one. Not surprisingly, museums with fewer candidates spent more time on each candidate.

Figure 5 shows the statistics chart present in the home page of the link review tool. For each museum it shows the total number of links in need of review, and progress towards completion. The “Not Sure” category is the result of sparse records, containing values for too few fields to enable confident as-

essment. The large number of “Unmatched” records for NPG (National Portrait Gallery) is the result of a large number of constituent records for people depicted in the painting with names similar to that of artists.

Table 4 summarizes the data and results of the linking process. Column G shows the number of links to ULAN records present in the datasets provided by the museum. We used these links as ground-truth to evaluate our automated algorithm: columns P , R and F shows the precision, recall and F1-measure of our algorithm. We hypothesize that the set of links provided do not represent a random sample of artists given that the number of rejected links shown in

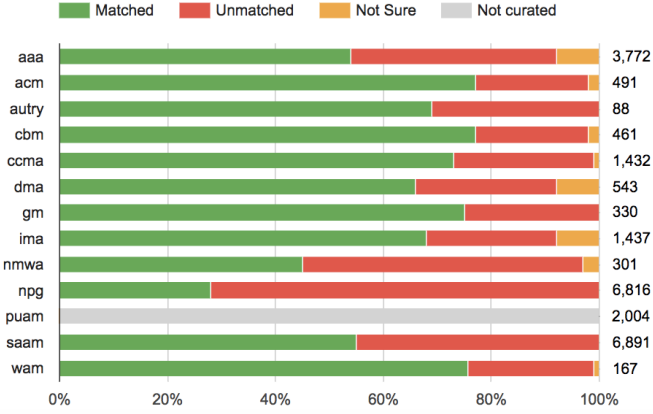


Fig. 5. Status chart from the Link Review Tool showing the work completed for each museum

Table 4. Statistics on the linking process.

| Museum | Artists | $ G $ | P | R | F | Pairs | $ G \cap A $ | $ G^* $ | $ A \setminus G $ | $ A^* $ | $ G \setminus A $ |
|--------|---------|-------|-----|-----|-----|--------|--------------|---------|-------------------|---------|-------------------|
| AAA | 6,944 | 0 | - | - | - | 3,772 | - | - | 2,038 | - | - |
| ACM | 772 | 0 | - | - | - | 491 | - | - | 377 | - | - |
| AM | 114 | 73 | .93 | .75 | .83 | 88 | 55 | 0 | 6 | 0 | 18 |
| CBM | 513 | 0 | - | - | - | 461 | - | - | 354 | - | - |
| CCMA | 2,005 | 1,060 | .96 | .85 | .90 | 1,432 | 1,043 | 1 | 0 | 0 | 17 |
| DMA | 649 | 0 | - | - | - | 543 | - | - | 358 | - | - |
| GM | 1,198 | 266 | 1.0 | .94 | .97 | 330 | 229 | 0 | 16 | 1 | 37 |
| IMA | 2,077 | 671 | .96 | .92 | .94 | 1,437 | 596 | 17 | 359 | 1 | 58 |
| NMWA | 375 | 0 | - | - | - | 301 | - | - | 135 | - | - |
| NPG | 12,552 | 0 | - | - | - | 6,816 | - | - | 1,919 | - | - |
| PUAM | 2,866 | 1,174 | .95 | .90 | .93 | 2004 | - | - | - | - | 1,174 |
| SAAM | 12,439 | 0 | - | - | - | 6,891 | - | - | 3,769 | - | - |
| WAM | 181 | 105 | .98 | .95 | .97 | 167 | 99 | 1 | 26 | 0 | 6 |
| Total | 42,685 | 3,349 | .96 | .88 | .92 | 24,733 | 2,022 | 19 | 9,357 | 2 | 1,310 |

Figure 5 is much larger than the precision numbers in the table suggest. The Pairs column represents the number of candidate links generated by our algorithm. The number of pairs is always smaller than the number of artists because ULAN is incomplete. Column $G \cap A$ shows the number of approved links that were part of the museum data. The reduced number is a result of sub-optimality of the blocking algorithms and recall failures of the matching algorithm. Column G^* shows the count of incorrect ULAN links present in the museum dataset. We identified the links where the museum and our review tool disagree, and we evaluated the links by looking at the Web pages in the museum and Getty websites. Similarly, A^* shows the count of incorrect links produced by our review tool. Column $A \setminus G$ contains the counts of new links produced with our linking workflow. The number of new links is more than double the number of links present in the museum databases. The last column shows the counts of links present in the museum databases that were not discovered by our workflow. The numbers are relatively small, except for PUAM, which opted out of the review.

Lesson 6 - Linking Tools: We found it difficult to configure and use existing semantic web linking tools to generate links against a large dataset, such as ULAN, DBPedia, and VIAF. We need to have scalable, easy-to-configure, easy to work with libraries for creating the links.

Lesson 7 - Manual Review: Users are willing to invest significant time and effort to ensure that the final data is accurate (a few weeks of effort by museum personnel more than tripled the number of existing links).

4 Using the Data

The goals of the American Art Collaborative include finding ways to foster collaboration among multiple institutions over time to support exploration, scholarship, and information access. Thus, establishing methods that allow federated, linked information to grow over time along with the commitment of all the people who use and manage the information within institutions is critical to success. The development of the prototype Browse Application described in this section is important to make Linked Data real to museum users.

The project established a Browse Working Group, involving 6 of the 14 institutions, to design and develop a usable application for exploring the AAC data. The group began the process by identifying the goals of the partner institutions, as well as gathering ideas for the types of explorations that were difficult to do currently on the web. An analysis survey was developed to gather qualitative input from curators, registrars, educators, and outside researchers. One finding from this analysis was, as expected, that people find it easy to identify barriers in their existing work processes, yet struggle to imagine alternative approaches. This validated the projects commitment to a browse application that would present information in new ways.

The Mapping Validator and the Browse application (Figure 6) were designed and developed in parallel, making sure that the data to be displayed was valuable and available. Iterative design focused on both presenting the primary entities

from the data (people, artworks, museums, and possibly also locations and subjects) and exposing relationships between the entities. Pages for artworks and artists feature a small panel on the right side that provides links to related objects and are described with short phrases that describes the relationships.

To help with search and to improve performance across the growing collection of data, the browse application is populated by querying the triplestore and generating JSON-LD documents for each primary entity. These documents include all the associated link references for the entity, to allow rich cross-referencing within the browse application and to allow populating dynamic JavaScript features that need rich data. The JSON-LD documents are stored in Elasticsearch so that they are easily searchable by the application, which is useful when generating lists of related artworks based on specific parameters. The triplestore remains available for more complex and ad hoc queries.

Each page was designed to incorporate small visualizations, aggregations, and tools that help expose interesting aspects of the data about the pages artwork or artist. These tools were nicknamed toys in the toybox and presented below the artwork's data on the page (Figure 7).¹⁴ Each individual toy has its own horizontal row, and has a profile that expresses what data it needs from the linked data to be able to present a usable representation. As the entity loads into the page, the available data is checked against the profiles, and each toy is

¹⁴ The depicted screen is under development.


American Art Collaborative Demonstration Application About the AAC Settings

AAC COLLECTIONS

INSTITUTIONS EXPLORE ARTISTS EXPLORE BY CATEGORIES COLLECTION PROFILE

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 x 55 1/2 x 3 in. (120.7 x 141 x 7.6 cm)
 Framed: 47 1/2 x 55 3/8 x 3 1/4 in. (120.7 x 140.7 x 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, Native American, American Indian

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 institutions

Fig. 6. Screenshot from the Browse application, which allows museums to review their data and access relationships

shown or hidden accordingly. This allows developers from different institutions to create toys over time and for each institution to decide what toys they think are useful to present with their entity data. The toybox approach extends the capabilities that are available as the data contributions from partner museums grows.

Even in its early stages as a prototype, the browse application is proving useful to all the partner institutions when reviewing the data that has been generated. This clear, human-readable presentation helps museum data managers check that the full pipeline from export of their initial data, through mapping and conversion to RDF, through querying and populating the browse application, produces the high quality and accuracy they expect.

Lesson 8 - Data Visualization: An easy to understand visualization is needed for non-technical users to review the linked data. With a complicated ontology, existing Linked Data interfaces, such as Pubby,¹⁵ are not useful for users to view their data.

Lesson 9 - Simple Schema: We needed a simple schema rather than the complicated one provided by the CRM, so we created SPARQL queries to map subjects, persons, and objects into JSON and then used Elasticsearch to analyze the interconnections to build the interface.

¹⁵ <https://github.com/cygri/pubby>

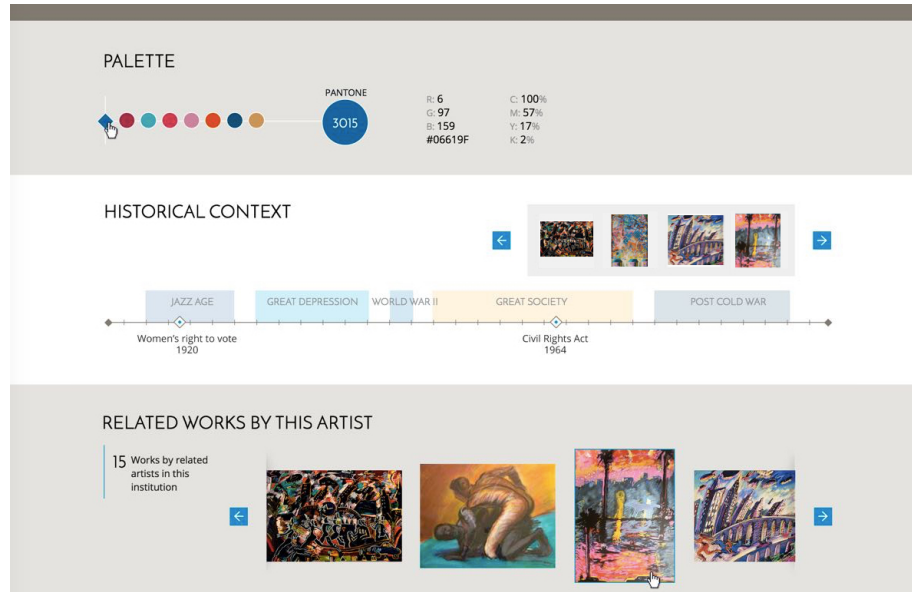


Fig. 7. The artwork and artist pages allow small independent tools to be incorporated, so scholars can discover patterns and have different ways to explore the available data

5 Related Work

There is a great deal of interest in publishing museum data as Linked Data. Europeana[6], one of the most ambitious efforts, published the metadata on 17 million items from 1,500 cultural institutions. The Canadian Heritage Information Network (CHIN), published the data on 85,000 items from 8 Canadian museums.¹⁶ For both Europeana and CHIN, they integrate the data by publishing a fixed schema and requiring all of the participating institutions to transform their data into the required format (in a few cases CHIN mapped the data for the museums). The MuseumFinland published the metadata on 4,000 cultural artifacts from museums and sites in Finland [7] and the Amsterdam Museum [8] published the metadata on 73,000 objects. In both of these efforts the data is first mapped directly from the raw source into RDF and then complex mapping rules transform the RDF into an RDF expressed in terms of their chosen ontology. The LODAC Museum published metadata from 114 museums and research institutes in Japan [9]. They defined a relatively simple ontology that consists of objects, artists, and institutions to simplify the mapping process.

Research Space is a large effort to create the infrastructure for conducting research on cultural heritage data. A number of institutions participate in research space and have mapped their collections to the CRM ontology and published their data as Linked Data. These include the British Museum¹⁷ and the Yale Center for British Art.¹⁸ There are also consortiums that are participating in Research Space, such as the PHAROS project, a consortium of fourteen historical photo archives that are in the process of publishing their data as Linked Data using the CRM ontology.¹⁹ In all these projects, the individual institutions are responsible for publishing their own data to the CRM ontology and these are multi-year projects with experienced technical staff that have a strong working knowledge of both the Semantic Web and the CRM ontology.

In a precursor to this project, we collaborated with the Smithsonian American Art Museum to publish their data as Linked Data [10]. In that project we also mapped the data to both the EDM and CRM ontologies using Karma, linked the artists to other sources (DBPedia), and created an initial link review tool. The AAC project forced us to address the issues of how to do all this work in a consistent fashion across multiple museums and to do so at scale, such as the mapping validation tool, the browse application, and a link review tool that supports crowd sourcing.

In this project we go beyond earlier work in several important ways. First, we developed a workflow that transforms and maps the data using Karma, and then validates the mappings using the Mapping Validation tool. Other approaches first map data directly into RDF [11] and then aligns the RDF with the domain ontology [12]. There is also work on specifically mapping to CRM using X3ML [13], a system that requires mapping data into XML and writing rules to map the

¹⁶ <http://chin-rcip.canadiana.ca/aclod/about>

¹⁷ <http://collection.britishmuseum.org/>

¹⁸ <http://britishart.yale.edu/collections/using-collections/technology>

¹⁹ <http://pharosartresearch.org>

XML to the corresponding CRM terms. For the AAC that would require manually writing a prohibitive number of such rules. These other approaches automate less of the mapping task and all of the data cleaning, mapping validation, and data verification would need to be done by hand.

Second, in order to provide an integrated view of artists across museums, we developed a crowd-sourcing link review tool. There is a great deal of work on linking data, such as the work on Silk [5], but very limited work on how to review the proposed links. Museums want to publish high-quality data, so verifying the links is critical part of the linked data creation process. There are several other tools for solving this problem. Mix'n'match²⁰ is a tool for importing and linking new data into Wikidata.²¹ The tool runs a fuzzy name match to generate a set of candidate entities in Wikidata and then allows the user to confirm or remove the matches. OpenRefine [14] provides a reconciliation capability that allows users to link a dataset to another one using specified fields and then interactively disambiguate the links. Both of these tools provide a link review capability, but they are targeted to highly technical users and are not well suited to experts in other fields.

6 Discussion

In this project we collaborated with 14 American art museums to build high-quality linked data about their artwork. We also created a set of tools that allowed a team of USC students to map the data without being experts in the CRM ontology and allowed the staff of the museums to review links to other resources. These tools, which are all available as open source, include: 1) the Karma data integration tool,²² which cleans and maps the data, and has been extended for this project to store all of the associated mappings and data directly in Github, 2) a Mapping Validation tool²³ that provides both a specification of the precise ontology mapping and corresponding query that returns the data only if it has been correctly mapped, 3) a data generation tool (available as part of Karma), which applies the Karma mappings to the datasets to create the RDF data and load it directly into a triplestore, 4) a general link review tool²⁴ that allows non-technical users to quickly and easily review the links to other resources, and 5) a browse application²⁵ that allows both the museum staff, art historians, and the general public to review and explore the resulting Linked Data. All of these tools are being released as open source. The mapping tools were used extensively by the students that worked on the project (with limited background in the Semantic Web) and the link review and browsing tools were used extensively by the museum staff (who had limited technical background).

²⁰ <https://tools.wmflabs.org/mix-n-match/>

²¹ <http://www.wikidata.org>

²² <http://karma.isi.edu>

²³ <http://review.americanartcollaborative.org>

²⁴ <https://github.com/american-art/linking>

²⁵ <http://browse.americanartcollaborative.org>

In future work we plan to explore techniques to simplify the task of publishing Linked Data so additional museums can easily join the AAC. Since many museums already publish their data on the web, we would like to gather, map, and link their data directly from the content they already make available online. We also want to extend the types of information supported and to link the existing data to other resources, such as VIAF,²⁶ Geonames, and DBpedia.

Acknowledgements

The American Art Collaborative was made possible by grants from the Andrew W. Mellon Foundation and the Institute of Museum and Library Services. We thank Eleanor Fink for leading the AAC project and all of the USC students and museum personnel who helped with all aspects of the project.

References

1. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E. In: *Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets*. Springer International Publishing (2014) 226–230
2. Knoblock, C.A., Szekely, P., Ambite, J.L., , Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyani, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: *Semantic Web*, Springer (2012) 375–390
3. Taheriyani, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: Learning the semantics of structured data sources. *Journal of Web Semantics* **37**(C) (2016)
4. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: *Proceedings of ACM SIGKDD*. (2003) 39–48
5. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk—a link discovery framework for the web of data. In: *Proceedings of the 2nd Linked Data on the Web*. (2009)
6. Haslhofer, B., Isaac, A.: data.europeana.eu: The europeana linked open data pilot. In: *International Conference on Dublin Core and Metadata Applications*. (2011)
7. Hyvonen, E., Makela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland - Finnish museums on the semantic web. *Web Semantics* **3**(2-3) (2005) 224–241
8. Boer, V., Wielemaker, J., Gent, J., Hildebrand, M., Isaac, A., Ossenbruggen, J., Schreiber, G.: Supporting Linked Data Production for Cultural Heritage Institutes. In: *Lecture Notes in Computer Science*. Springer Berlin (2012) 733–747
9. Matsumura, F., Kobayashi, I., Kato, F., Kamura, T., Ohmukai, I., Takeda, H.: Producing and consuming linked open data on art with a local community. In: *Proceedings of the COLD Workshop, Volume 905, CEUR-WS.org* (2012) 51–62
10. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E., Allen, R., Goodlander, G.: Publishing the data of the Smithsonian American Art Museum to the linked data cloud. *International Journal of Humanities and Art Computing* **8** (2014) 152–166
11. Cyganiak, R., Bizer, C.: D2r server: a semantic web front-end to existing relational databases. *XML Tague* **2006** (2006) 171–173
12. Bizer, C., Schultz, A.: The R2R Framework: Publishing and Discovering Mappings on the Web. *1st International Workshop on Consuming Linked Data* (2010)
13. Marketakis, Y., Minadakis, N., et al.: X3ml mapping framework for information integration in cultural heritage and beyond. *Digital Libraries* (2016) 1–19
14. Verborgh, R., De Wilde, M.: *Using OpenRefine*. Packt Publishing Ltd (2013)

²⁶ <https://viaf.org/>