American Art Collaborative (AAC) Linked Open Data (LOD) Initiative

Overview and Recommendations for Good Practices

Eleanor E. Fink

2018

Table of Contents

About the Author

ELEANOR E. FINK is the founder and manager of American Art Collaborative. She has held senior positions at the Smithsonian Institution, J. Paul Getty Trust, and World Bank. She is one of the founding directors of the Getty Center in Los Angeles, for which she initially formed and headed the Getty Vocabulary Program and later became director of the Getty Information Institute (GII). As director of GII, she oversaw the Getty's flagship scholarly art history research databases existing at the time, including the Census of Antique Art and Architecture Known in the Renaissance, and Bibliography of the History of Art and the Getty Provenance Index. She positioned GII around the concept of universal access to images and art information and promoted national and international collaboration among institutions. The National Initiative for a Networked Cultural Heritage, Getty vocabularies, Categories for the Description of Works of Art, and Object ID are some of the products of her leadership. Eleanor serves on the advisory committees of the Department of Art and Archæology at Princeton University, New Jersey; the EU's Virtual Multimodal Museum project, and Marie Curie Research Program on the Initial Training Network for Digital Cultural Heritage, Limassol, Cyprus. She is a former director of the Museum Computer Network and a former president of the Visual Resources Association.

Acknowledgments

I have many people to thank for their assistance in shaping the American Art Collaborative (AAC), explaining its work, and summarizing its recommendations for museums embarking on Linked Open Data (LOD). First and foremost, I wish to thank Rachel Allen, Deputy Director of the Smithsonian American Art Museum (SAAM), Washington, DC, who early on saw with me the possibilities that LOD could deliver to museums, the public, and research communities.

I am deeply indebted to the individuals from the fourteen participating institutions of the AAC, its advisory council, and consulting partners. From the onset, the enthusiasm for AAC and the teamwork among all the partners were sparks that set AAC's initiative in motion and shaped it. Special thanks go to Duane Degler at Design for Context, Washington, DC, for his work on the project as well as his additions and corrections to this guide; David Newbury at the J. Paul Getty Trust, Los Angeles, for his work in designing the browse function and together with AAC adviser Rob Sanderson at the J. Paul Getty Trust for producing AAC's target model; Craig Knoblock and his team, especially the students at the Information Sciences Institute, University of Southern California, Los Angeles for mapping and hosting the data and producing several tools; Kate Blanch at Design for Context and Emmanuelle Delmas-Glass at Yale Center for British Art, New Haven, Connecticut, for reaching out to AAC members who needed help in preparing, submitting, and proofing their data; Stephen Stead, at Paverprime Ltd, Croydon, United Kingdom and Vladimir Alexiev, Ontotext, Sofia, Bulgaria for independent reviews of the mapping and consistency with the AAC target model. Equal thanks are due to the AAC's advisory partners for their guidance and review of this guide: Craig Knoblock, Information Sciences Institute; Rob Sanderson, J. Paul Getty Trust; Martin Dœrr, The Foundation for Research and Technology–Hellas (FORTH), Greece; Vladimir Alexiev, Ontotext; and Thorny Staples, retired from the Smithsonian. I also thank Diane Zorich, Director, Digitization Program Office,

Smithsonian Institution, who was to be a coauthor but for scheduling reasons was not available to do so; she nevertheless contributed ideas and text to these documents.

Most important, the AAC would not have been able to publish two hundred thousand LOD object records, produce open-source tools, develop a browse demo, and write this guide about its experience were it not for generous funding from the Andrew W. Mellon Foundation, New York, and the Institute for Museum and Library Services, Washington, DC. I also thank Rachel Allen and her colleagues at SAAM and Shane Richey and his colleagues at Crystal Bridges Museum of American Art, Bentonville, Arkansas, for their administration of the grants. Writing the grants and reports and administrating the funding were handled smoothly and never interrupted AAC's process.

I express deep gratitude to freelance editor Jane McAllister, who patiently reviewed several versions of this guide, on which she brought to bear her years of experience and adherence to good editorial practices. I thank SAAM's director of publications, Theresa Slowik, and graphic designer Robert Killian for taking this guide through design and production.

Although AAC has reached the end of its current grant cycles, the enthusiasm to build upon what we have learned—as documented in this guide—is just beginning.

Introduction

Since the dawn of antiquity, the world's cultures have collected their artifacts and produced written records to shed light on the connections that give meaning and bind civilizations together. The artifacts are signposts for understanding the story of humankind: where civilization originated, what civilization achieved, and what it can teach us.

Unfortunately, today's documents about artifacts are being chronicled in a variety of machine-readable formats that do not allow us to easily make connections across multiple databases and museum websites at once. Data silos restrict our ability to grasp and make associations about the lifework of a certain artist, the scope of a particular style, or the details of a historic period. To begin research today, you must know where to look or which institution has works by a given artist or school.

Fortunately, as technology advances it provides new opportunities for users to make connections across data and deepen research. One example is Linked Open Data (hereafter LOD), a method for publishing structured data on the World Wide Web that allows information to be interconnected and thus rendered broadly useful. Although LOD has been applied in publishing, commercial enterprises, and the scientific and government sectors for some years, it remains a nascent concept among museums and other cultural-heritage repositories. Now that museums are increasingly deploying technology to reach new audiences, strengthen research, and help people of all ages experience, learn about, appreciate, and enjoy art, interest in LOD is gaining traction. In fact, the notion of expanding audiences, gaining visibility, improving research, building collaborative platforms, facilitating searches across multiple collections simultaneously, and enhancing educational resources are some of the features that inspired the formation of the American Art Collaborative (AAC) Linked Open Data (LOD) Initiative. The AAC is a consortium of fourteen art institutions (thirteen museums and one archive) in the United States established in 2014 to investigate and begin building a critical mass of LOD on American art (see Appendix 1: AAC Partners, Advisers, and Consultants).

This guide, one of the key products of the AAC, is not a starter kit or step-by-step technical manual. Rather, our purpose here is to share with the museum community what AAC set out to accomplish, how AAC approached LOD, the tools we used, the trials we encountered, the lessons we learned, and our recommendations for good practices for museums interested in joining the LOD community (see part 2 of this guide, "Recommendations for Good Practices," which is also a separate document on the AAC website, https://americanart.si.edu/about/american-art-collaborative). For help with acronyms, readers are advised to refer to Frequently Used Acronyms (see boxed inset).

Additional resources are available for those seeking in-depth technical details. One source is a paper titled "Lessons Learned in Building Linked Data for the American Art Collaborative" by Craig Knoblock et al. in the proceedings of the International Semantic Web Conference in Vienna, Austria, October 21–25, 2017 (https://iswc2017.ai.wu.ac.at/wp-content/uploads/papers/MainProceedings/382.pdf). Another is Linked Art (https://linked.art), which contains an overview of the target model that AAC developed as a profile of the Conceptual Reference Model (CRM) ontology created by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM), Paris. Also helpful for explaining technical details is the review tool that the consulting firm Design for Context, based in Washington, DC, created as a testbed for the AAC browse function. Last but not least, in the individual directories of each institution participating in AAC, GitHub contains threaded discussions about mapping. GitHub can be accessed via https://github.com/american-art or the AAC website, https://americanart.si.edu/about/american-art-collaborative.

One of AAC's conclusions and an urgent recommendation to mention up front is the need for the cultural-heritage community to address the problems of legacy data that have accumulated over decades—such as the fact that standards for expressing dates, dimensions, materials and techniques, unknowns, and the like are currently lacking. The status quo, which adds to the challenge of creating LOD, ultimately diminishes the ability to connect cultural-heritage information to the multitude of documents around the globe that tell the story of humankind.

Part 1. Overview

FUNDAMENTALS OF LOD FOR USE IN MUSEUMS

Before presenting how AAC began, what it set out to accomplish, and how we approached LOD (see below, "AAC's Inception and Adaptation of LOD), the reader may benefit from a brief overview of fundamentals of LOD for use in museums (see page 13).

In computing, "Linked Data" describes a method of publishing structured data so that it can be interlinked and therefore useful in web implementations. Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term. Linked Data represents a milestone in realizing what is called the Semantic Web, or web of Linked Data. To achieve the "semantic glue" that provides context and meaning in documents, information needs to be tagged, much like the mark-up language HTML serves for publishing on the web. Unlike web hyperlinks that broadly connect thousands of bits of information that seem similar based on matching words, however, LOD produces highly precise search results. A search for facts about a lost painting by the American artist Winslow Homer that he titled *Boy Reading*, for example, can produce hundreds of outcomes based on keywords such as "boy," "reading," "boy reading," "Winslow Homer," lost or stolen art in general, and so forth. The researcher must read the full list of results to determine which links, if any, are about the lost painting by Homer called Boy Reading. With LOD, on the other hand, the results would be more focused on links about that exact painting. The "noise," or unessential information associated with matching keywords, can be reduced.

Such exactness is achieved by using a data format called the Resource Description Framework (RDF)[1] coupled with a domain ontology (a schema tied to concepts that give meaning to data) or multiple ontologies. RDF breaks down knowledge into discrete parts, according to rules about the semantics, or meaning, of the parts. Information is expressed as a list of statements in the form subject/predicate/

object (e.g., "Rebecca paints portraits"), known together as triples. Each subject, predicate, and object can be identified by a Uniform Resource Identifier (URI), a string of characters used to uniquely designate the subject, predicate, and object in a way that can be read by computers.

The staff that produces LOD must select one or more ontologies that will play the key role of defining the meaning of the relationships between the elements in the subject/predicate/object statements. In essence, RDF and a discipline-specific ontology such as the CIDOC CRM provide pertinent meanings to the terms in the statements based on their structure and the defined predicate relationships. Using URIs with the RDF statements enables the pieces of the triples published as LOD to be computer accessible so they can be searched and connected.

Once triples are tagged and mapped, they are stored in a database called an RDF triplestore. To allow others to query RDF data, many institutions will choose to publish LOD on a Semantic Protocol and RDF Query Language (SPARQL) endpoint, which allows users to query a knowledge base via the SPARQL language. SPARQL is a semantic query language that permits databases to retrieve and manipulate data stored in RDF. In addition, institutions will usually make available an RDF dump, which is simply a file containing all the RDF statements; others can easily download the file to load into their respective triplestores. The RDF dumps are often used to process data efficiently across multiple sites.

Ontologies and the CIDOC CRM

As stated above, an ontology is a schema, or conceptual framework, that gives meaning to data. In the words of Wikipedia, in information science as well as philosophy an ontology is the "representation of entities, ideas, and events, together with their properties and relations, according to a system of categories."[2] Without application of an ontology, data sets would be only loosely connected pieces of information, inconsistent from one provider to another. An ontology essentially creates the "semantic glue" that transforms discrete fragments of data into precise concepts. Typically, an ontology exists for each discipline or domain of knowledge. As described by Wikipedia:

A domain ontology (or domain-specific ontology) represents concepts which belong to part of the world. Particular meanings of terms applied to that domain

are provided by domain ontology. For example, the word card *has many different meanings. An ontology about the domain of poker would model the "playing card" meaning of the word, while an ontology about the domain of computer hardware would model the "punched card" and "video card" meanings.*[3]

In the domain of cultural heritage, ontologies created for bibliographic data include Functional Requirements for Bibliographic Records (FRBR) and Bibliographic Framework (BIBFRAME); Schema.org with some extensions is also available. The most relevant ontologies for museums are the Europeana Data Model (EDM), created for Europe's large-scale Europeana Linked Open Data Initiative of the Europeana Foundation, The Netherlands, which is inter-connecting data across libraries, archives, and museums (https://ec.europa.eu/digital-single-market/en/europeana-european-digital-library-all), and the CIDOC CRM, created by the International Committee on Documentation of the International Council of Museums (http://www.cidoc-crm.org). Both cover cultural-heritage information for museums, libraries, archives, and archæology collections, with the primary objective being the interchange of cultural-heritage data. By integrating cultural information from a variety of sources and bringing it together into an integrated environment, these ontologies and LOD allow the researcher to ask for and obtain more and deeper information than can be derived from an individual resource.[4] To varying degrees, Europeana's EDM and the CIDOC CRM supply definitions and formal structures for describing the implicit and explicit concepts and relationships used in cultural-heritage documentation. The Digital Public Library of America (DPLA; see https://dp.la) uses EDM, which currently makes more than twenty-four million items accessible.

How the CIDOC CRM Works

The CIDOC Conceptual Reference Model (CRM), also recognized as International Organization for Standardization (ISO) 21127:2006, is the most extensive cultural-heritage ontology. In simple terms, the CIDOC CRM consists of eighty-two entity types (classes), or concepts (such as E2 Temporal Entities, E18 Physical Thing, E28 Conceptual Objects, E39 Actors), that can be connected through 263 relationships (properties). Through these classes and properties,

Frequently Used Acronyms

AAC: American Art Collaborative
AAT: The Art & Architecture Thesaurus (Getty)
API: application programming interface
CC: Creative Commons
CIDOC: International Committee for Documentation of the International Council of Museums
CIDOC CRM: Conceptual Reference Model (CRM) of the International Committee on Documentation (CIDOC) of the International Council of Museums (ICOM)
CIS: collection information system
CRM: Conceptual Reference Model
CSV: computerized system validation
DNS: Domain Name System
DPLA: The Digital Public Library of America
EDM: Europeana Data Model
GII: Getty Information Institute, Los Angeles
HTML: Hypertext Markup Language
HTTP: Hypertext Transfer Protocol
ICOM: International Council of Museums, Paris
IIIF: International Image Interoperability Framework
IMLS: The Institute of Museum and Library Services, Washington, DC
IP: intellectual property
ISI: Information Sciences Institute at University of Southern California (USC), Los Angeles
ISO: International Organization for Standardization
JSON: JavaScript Object Notation
JSON-LD: JavaScript Object Notation-Linked Data
KML: Keyhole Markup Language
LOD: Linked Open Data
RDF: Resource Description Framework
SAAM: Smithsonian American Art Museum
SPARQL: Semantic Protocol and RDF Query Language
SQL: Structured Query Language
TGN: The Getty Thesaurus of Geographic Names
TMS: The Museum System
ULAN: The Union List of Artist Names (Getty)
URI: Uniform Resource Identifier
URL: Uniform Resource Locator
USC: University of Southern California, Los Angeles
W3C: World Wide Web Consortium
XML: Extensible Markup Language

the CRM provides a semantic framework to articulate how the classes are related to the wide variety of events that can be connected to the history of objects in collections, such as creation, modification, acquisition, exhibition (see fig. 1). While the CRM is event based and dynamic, museum data is object based and static—consisting of discrete fields of information, such as artist, title, date.

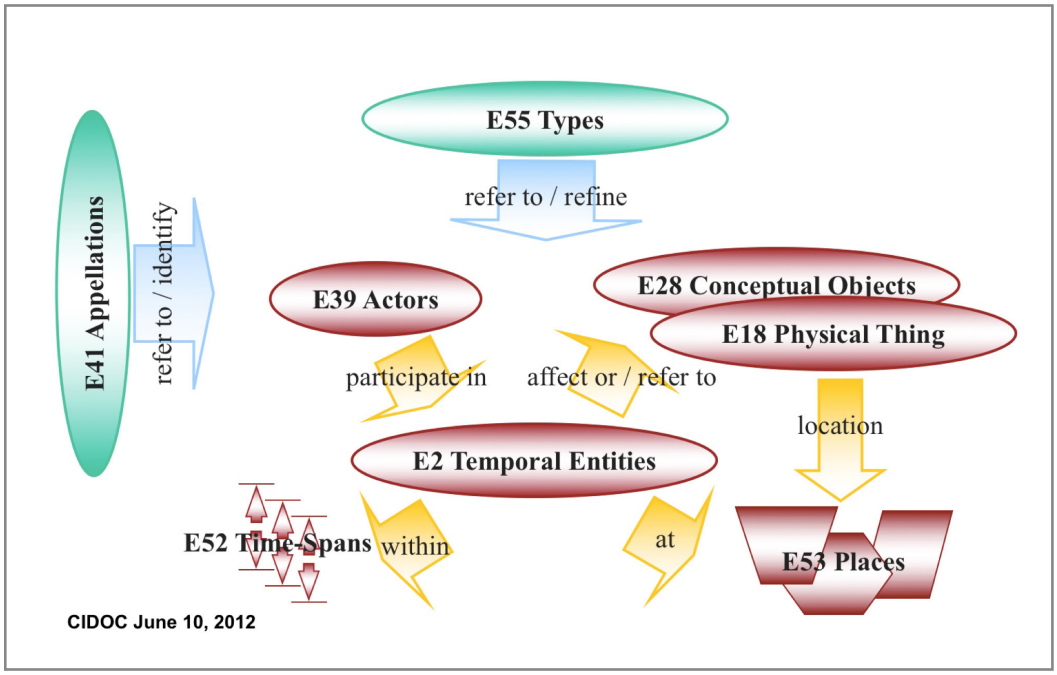

Fig. 1: Illustration of CIDOC CRM high-level entities showing the central role of Events (Temporal Entities). Slide from Martin Doerr and Stephen Stead, "Harmonizing Models for the Digital World: CIDOC CRM, FRBROO, CRMDig, EDM." Presentation for CIDOC CRM Special Interest Group, Helsinki, Finland, June10, 2012. Slide 17.

CRM is a living ontology, and new releases and extensions are added periodically. The event-driven characteristic of the CIDOC CRM, combined with LOD, opens up new prospects for conducting research. AAC chose to use the CIDOC CRM to capture the richness of cultural-heritage material and offer researchers the potential to find relevant and unanticipated facts.

In addition to the participating institutions of the AAC, some other cultural-heritage institutions and/or projects using the CIDOC CRM are: the Arachne Project, Cologne; Arches, a collaboration between the Getty Conservation Institute, Los Angeles, and the World Monuments Fund,

New York; Art Tracks, Carnegie Museum of Art, Pittsburgh, Pennsylvania; Canadian Heritage Information Network (CHIN) of the Department of Canadian Heritage; CLAROS (CLassical Art Research Online Research Services), a federation led by the University of Oxford; Finnish National Gallery, Helsinki; Germanische Nationalmuseum, Nuremberg; the Getty Museum, Getty Provenance Index (GPI) and Getty Vocabulary Program, Getty Research Institute, Los Angeles; Pharos, the International Consortium of Photo Archives; ResearchSpace located in the British Museum, London; and Yale Center for British Art, Yale University, New Haven, Connecticut. For further information on projects applying the CIDOC CRM, visit http://www.cidoc-crm.org/useCasesPage.

A useful resource for learning about the CRM is the CIDOC CRM website, http://www.cidoc-crm.org, from which information in this section is drawn. A CRM primer that explains how to apply the CRM includes a tutorial by Stephen Stead, one of the members of the working group that developed the CRM (http://www.cidoc-crm.org/Resources/the-cidoc-conceptual-reference-model-cidoc-crm-primer). For a fuller explanation of the benefits of the CRM and LOD, see Richard Light's "Toward Linked History," the paper for a talk given at the annual conference of CIDOC, September 6–11, 2014, in Dresden, Germany (http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2014/LIGHT__Richard_CIDOC2014.pdf).

Target Models

A target model is a subset of all the mapping possibilities represented in your data. It works as a road map of your decisions about how your data will be mapped to the ontology or ontologies you select (i.e., which classes and properties pertain). Whether applied to data from one museum or a collaborative project involving multiple museums, a target model will help you simplify the mapping process, avoid guesswork, and provide consistency across the data.

Triplestore and SPARQL Endpoint

The result of using RDF and URIs to tag data is the production of triples: object, subject, predicate. The triples are stored in a purpose-built database called a triple- or RDF store. Triplestores support SPARQL language.

SPARQL is the query language used to search and retrieve RDF data; the endpoint is an interface created to facilitate the queries. An endpoint is used to query a triplestore and deliver results in different formats—such as Extensible Markup Language (XML), computerized system validation (CSV), or JavaScript Object Notation (JSON)—to accommodate the output needs of various users.

Reconciliation

Since LOD is about connecting or linking data, it is essential for data managers to reconcile references to entities across sources. For LOD to be useful across data sets or institutions, it is important to know when two data sets are referring to the same conceptual "thing"—such as a specific work of art or artist. One way to express that two entities are the same is to refer to a common, authoritative external resource such as one of the databases of the Getty Vocabulary Program of the Getty Research Institute, Los Angeles. "Reconciliation" is the process of identifying the common identity for multiple data items. Moreover, as libraries, museums, archives, consortia, government, and commerce generate hundreds of thousands URIs, many will inevitably refer to the same entity or thing. But how does the computer know they are the same? As Rob Sanderson expressed in his talk "The Linked Data Snowball and Why We Need Reconciliation"—presented at the Andrew W. Mellon Foundation Workshop on Reconciliation of Linked Open Data, New York, December 1, 2014—unless we explore reconciliation, the chances of interconnecting knowledge across our own and other domains will be greatly diminished. Anyone interested in LOD should look at Sanderson's presentation (see https://mellon.org/media/filer_public/73/a3/73a341cf-9045-4cf6-b883-acf97f42f3ac/sanderson_linked-data_12-1-2014.pdf), which outlines several approaches to solving this challenge.[5] A report on Sanderson's talk can be downloaded from https://mellon.org/resources/news/articles/linked-open-data.

AAC'S INCEPTION AND ADAPTATION OF LOD

AAC started informally in 2012/13 with discussions initiated by Eleanor Fink among several museums interested in LOD. The museums wanted to learn more about how LOD can connect information across collections and explore the relationships of artworks among those collections. Museum staff were also curious about cross-domain connections, that is, among archives and history or science museums, particularly on a local level.

Fink who would become the founder and manager of AAC, had long championed the idea of being able to connect data across institutions. While director of the Getty Information Institute (GII), Los Angeles (no longer extant), she had positioned GII around the concept of universal and seamless access to images and art information as if it were all one "virtual database."[6] It appeared timely to engage several museums in an LOD initiative.

Two years before those discussions took place, the Smithsonian American Art Museum (SAAM), which has pioneered the use of information technology to increase the understanding of American art,[7] had contracted the Information Sciences Institute (ISI) at University of Southern California (USC), Los Angeles, to convert forty-four thousand SAAM digital records to LOD, which SAAM then made available on its website. When SAAM's data was being converted under the direction of Deputy Director Rachel Allen, it became apparent that few additional art-related resources existed in LOD format.[8] The discovery was disappointing, since one of the exciting benefits of LOD is its ability to interconnect information across institutions.

Recognizing the need to produce more cultural-heritage information as LOD and demonstrate the value of erasing data silos, Fink proposed to SAAM and ISI the idea of enlisting several museums in an LOD project. Since SAAM's content is about American art, the museum partners would work to create a critical mass of LOD on American art. SAAM and ISI agreed the project would be worthwhile to pursue.

As part of her inquiry with museums about forming a LOD collaborative, Fink organized, with the help of ISI, educational briefings and online discussions about LOD and the Semantic Web. During a communication in February 2013, the group of museums agreed to use the name American Art Collaborative, although the commitments were not yet formal.[9] In April 2015, SAAM hosted a meeting of the interested museums at which several guest speakers gave presentations.[10] At the conclusion, the participants agreed that if grant funding could be obtained, they would like to collaborate and create a critical mass of LOD.

A Planning Grant and AAC's Interest in LOD

By 2014 a total of fourteen institutions—a good mix of large and small art museums, both federal and private, including some at academic institutions and one archive—were interested in collaborating to form AAC.[11] Through email exchanges

with Don Waters at the Andrew W. Mellon Foundation, New York, Fink was invited to submit a brief proposal about AAC and what it hoped to achieve. After reviewing her overview, Waters suggested that one of the AAC museums should submit a formal request to the Mellon Foundation for a planning grant. She approached SAAM for that task. The Mellon Foundation awarded a grant to SAAM on behalf of the AAC (September 2014–April 2015). Rachel Allen served as the grant's principal investigator, and Fink continued to lead the initiative as its manager. Each museum provided a signed letter from its director committing the museum to the goals of the grant.

Under SAAM's umbrella, funding covered the costs of an advisory committee, a project manager, consultants, educational sessions with guest experts, and an in-person meeting. Before scheduling the educational briefings, the project manager developed a survey to determine the level of knowledge about LOD among participants and explore their initial interest in it. It was clear from the onset that ACC participants possessed a variety of skills and that their familiarity with information about LOD varied considerably. Most of the planning grant focused, therefore, on sessions to educate AAC members on: LOD and how to create LOD; CIDOC CRM and how to apply CIDOC CRM; International Image Interoperability Framework (IIIF); and reports from LOD projects such as ResearchSpace and Europeana. The AAC website lists the briefings (for which it provides the associated MS PowerPoints and, for most of them, recordings) and posts presentations (see https://americanart.si.edu/about/american-art-collaborative).

An in-person meeting was held toward the end of the grant period. Based on what AAC participants had learned and discussed, all agreed that AAC should move forward to seek additional funding to convert data on American art from each museum's collection to LOD. Group members were particularly motivated by the fact that the current LOD cloud lacks a critical mass of data about art. They also felt that as the hosts and caretakers of art objects, the museums should be the primary entities to produce LOD about the objects in their collections as opposed to having the data produced by a third party without their involvement. Participants' interest in pursuing LOD and working collaboratively as opposed to approaching LOD independently is outlined in Appendix 2: AAC Participants' Comments and Questions about LOD, together with their questions that identify the initial challenges and uncertainties, many of which this guide attempts to answer.

AAC Mission Statement

Mission

The American Art Collaborative (AAC) is a consortium of fourteen art institutions in the United States committed to establishing a critical mass of Linked Open Data (abbreviated as LOD) on the semantic web.

AAC believes that LOD offers rich potential to increase the understanding of art by expanding access to cultural holdings, deepening research connections for scholars and curators, and creating public interfaces for students, teachers, and museum visitors. AAC members are committed to learning together about LOD, identifying good practices for publishing museum data as LOD, and exploring applications that will help scholars, educators, and the public. They are also committed to sharing best practices, guidelines, and lessons learned with the broad museum, archives, and library community and building a network of practitioners to contribute quality information about works of art in their collections to the LOD cloud.

AAC Road Map

With the help of advisers and consulting partner Design for Context based in Washington, DC, Fink devised a road map (see Appendix 3: Road Map) to outline the activities necessary for ACC partners to achieve the following project goals and objectives: create a robust critical mass of LOD on the subject of American art and a browse demonstration thus illustrating the usefulness of LOD; produce open-source tools; publish the data openly on the web; and establish a practice for engaging in LOD across the museum community by promoting the knowledge, tools, and lessons learned from the AAC's experiences.

The road map was later funded by two grants. The Institute of Museum and Library Services (IMLS), Washington, DC, awarded a one-year (2015–17; extended to 2017) leadership grant to AAC partner Crystal Bridges Museum of American Art, Bentonville, Arkansas, with the focus on converting AAC data to LOD, training, and producing open-source tools. The Andrew W. Mellon Foundation, New York, awarded an eighteen-month grant to SAAM (2015–17)[12] for developing a browse demo; training; producing additional open-source tools; implementing the IIIF; and reconciling data (e.g., across AAC submitted data, and between data

sets and the Getty vocabularies, etc.). The two grants split the implementation of a common road map, shared an advisory committee, and engaged Fink as manager.

AAC'S APPROACH TO THE PROJECT

From the onset, the partners knew that involvement in AAC was a tremendous learning experience. In a sense, the AAC process as well as the tools it was using and creating were a testbed for the whole museum community. Additionally, AAC recognized that if the project raised the bar too high, museums would be unlikely to explore LOD, and the effort would become a short-lived technology trend. Therefore, throughout the grant cycles, the AAC adopted a perspective of making practical decisions and choices (forget about 100 percent precision, 100 percent coverage, and 100 percent completeness: start somewhere, learn, correct). Limits had to be set on how many times AAC partners could rethink and revise what data they wanted to submit. As issues arose over legacy data, it became apparent that the project could not realistically resolve decades of data inconsistencies during the short time scale. Producing the AAC target model for mapping to the CRM resulted in a profile of the CRM that would be easier for museums to adopt than deciphering the complete CRM. Likewise, although the open-source mapping tool called Karma (http://karma.isi.edu) works well and can learn from user entry over time (self-learning), some of the museums found it difficult to download and apply without a high degree of technical capability on staff. ISI therefore refined instructions for downloading Karma. While there were no plans for the partners to apply Karma under the grants, three AAC institutions did so and updated their own data: Indianapolis Museum of Art at Newfields, Colby College Museum of Art, and the Archives of American Art. In summary, producing a user-friendly target model, testing new open-source tools, and deciding on recommendations for good practices were some of the positive outcomes of our viewing the project as a learning experience and having a practical mindset.

Data Selection: American Art and IP Considerations

To fulfill the project's collaborative purpose, American art was defined broadly to include artists who were born in, lived in, or spent considerable time in the United States. Works by English artist David Hockney, who has resided in Los Angeles, California, for many years, are included, for example, in the AAC data. Some of the

artists in the collection of the Yale Center for British Art are considered American. The museums participating in AAC were encouraged to submit as much data as possible, starting with the data on American art currently present on their websites. That information was already in the public domain and would therefore reduce the possibility of intellectual property (IP) limitations, which could pose a stumbling block, especially for the museums that had not yet developed an IP policy and needed to do so. Additionally, the rationale for starting with works of art referenced on each institution's website was based on the decision to send only the data (not the images) to the Information Sciences Institute (ISI) for conversion to LOD. ISI proposed that if the data included a link to the images on the website, those images (in the public domain) would be viewable when searching the data. To create a robust set of data and test the capabilities of LOD, partners were asked to augment data from their respective websites by providing, if available, some of the following: curatorial notes, exhibition and publication records, and videos and recordings.

"Hybrid Federation" versus Aggregated Model

AAC's perspective in approaching LOD differed in intent from that of existing projects such as ResearchSpace and Europeana. While the latter are aggregation models that collect, process, and provide access to LOD, AAC wanted instead to explore a more distributed and sustainable model. For support into the future, aggregation models depend on centralized resources and funding, which may be unrealistic conditions for museums in the long run. AAC decided that, in the spirit of the World Wide Web, each institution should be responsible for maintaining and updating its own data, and the data should ultimately reside on each museum's website and triplestore.

Centralized aggregation models do have some benefits, however, particularly enhanced performance and reduced complexity when creating applications that use the data. Since the data is collected in a centralized resource, consistency is one benefit, and access tools are likely to produce good results. Likewise, it is easier to establish relationships across data that exists in one place. Searching across federated sites poses technical challenges, so performance is generally poor.

In the case of AAC, each institution submitted data to ISI. The data was mapped using the CIDOC CRM as the primary ontology, reviewed by a contracted CRM expert, the advisers, and the partners, converted to RDF, and stored in a triplestore. The data would remain at ISI until each institution was capable of either implementing

a SPARQL endpoint, engaging a hosting service, or forming a hub with some of the AAC partners to share a triplestore and SPARQL endpoint. A Domain Name System (DNS) redirect from each AAC museum to the temporary ISI server would make the data viewable using each museum's chosen URL. Although a single institution (ISI) handled the mapping of the data, museum participants were trained, with the intent that each museum would be responsible for managing the updating and maintenance of its own data when the grant ended.

Contracting the Mapping and Conversion to LOD

To build a critical mass of LOD within the time frame of a twelve-month IMLS grant, it would have been unrealistic to expect the AAC partner museums to learn the CIDOC CRM and how to use the open-source mapping tool Karma. Museums typically plan work objectives one to three years in advance. While the grant was welcomed, museum staff could not immediately set aside workloads and concentrate on learning how to map LOD. The road map, therefore, was based on ISI mapping the data. The grant allowed us to have tutorials and hands-on workshops with the aim that some of the AAC museums, by the end of the grant, would be able to update and manage their own LOD. ISI has produced a step-by-step video on how to use Karma; see https://github.com/usc-isi-i2/karma-step-by-step/wiki.[13]

Work Practices

Shared work platforms, the creation of a website, working groups, and frequent webinars were key for keeping the representatives from the fourteen institutions engaged in the process and work flow. The AAC adopted the following tools and practices:

Confluence collaboration space: To provide a common space to share documents, Crystal Bridges Museum of American Art secured Confluence team-collaboration software and helped each museum, consultant, and adviser register to access the space. Instructions on exporting data and creating an ID, a list of options for licensing, and other documents were posted on Confluence, where the AAC participants, advisers, consultants, and USC students working on the project at ISI could access the information.

GitHub repository: ISI set up the web-based repository hosting service for the AAC's data-production work space. Each museum had its own section or directory that contained the data it exported, and each could subscribe to notices on issues and

actions. Thus, if one of the USC students at ISI had a question while mapping the data for an institution, the representative from that institution would receive a notice and could respond. Comments from all parties were posted in GitHub. The AAC implementation plan as outlined in the road map called for a CRM expert to review the mappings the USC students conducted; those comments were also posted to GitHub. If an adviser disagreed or elaborated on a comment, a discussion thread developed (see "Conflicting Views on How to Implement the CIDOC CRM" under "Summary of Lessons Learned," for a discussion of the multiple perspectives on how the CRM could be applied). When an issue emerged, it was given a number (e.g., "29 dimensions"), and when resolved, it was marked closed. Although using GitHub represented a learning curve for the museums, it became a central production work space for the AAC.

AAC website: The project's public website (https://americanart.si.edu/about/american-art-collaborative) was created to include: all materials from the educational briefings and meetings; talks, papers, and presentations given at professional associations; tweets; AAC's high-level work plan (see Appendix 4: High-Level Time Line); and a link to GitHub (https://github.com/american-art).

Working Groups

Four working groups were established to have ongoing discussions on issues and practices as well as share techniques. Topics of the working groups included data, images, communications, and browse functions. The groups often used GoToMeeting for their deliberations and posted decisions on Confluence.

Teamwork was strongly evident across AAC. Those with technical knowledge reached out to help others, some of whom were not as familiar with The Museum System (TMS), the most common collection information system in use within AAC. Kate Blanch from the Walters Art Museum, Baltimore, took charge of AAC's Data Working Group. She gave support to those unfamiliar with how to extract data and often held meetings to advise partners on how to prepare data. Sharing scripts made extracting some types of data easier. Cathryn Goodwin from Princeton University Art Museum, New Jersey, for example, distributed the Structured Query Language (SQL) script used at her museum to pull individual dimension elements from TMS object records (see https://github.com/american-art/PUAM/commit/f694c37de9e4b38903e6e0be3f550e52db811e6d).

Creative Commons Licensing

At the first meeting under the IMLS grant, members discussed the selection of a Creative Commons license (CC). It was surprising to learn that out of the fourteen institutions, only two had developed and published a policy on use of images and data. The CC license gives those who access data the right to share, use, and build upon a work. Several types of licenses, each with its own pros and cons, are offered; see https://creativecommons.org/licenses. The more common versions are: CC0, which allows full use and no restrictions; CC BY, which requires that attribution always be cited; CC BY-NC, which is the same as CC BY but for non-commercial use. For a better understanding of the choices, see recommendation 2, "Choose Image and Data Licenses That Are Easily Understood," in part 2 of this guide, "Recommendations for Good Practices When Initiating Linked Open Data (LOD) in Museums and Other Cultural-Heritage Institutions" (which is also posted on the AAC website, https://americanart.si.edu/about/american-art-collaborative).

Several of the AAC partners chose CC0; an equal split chose CC BY; and a minority chose CC BY-NC. After realizing that CC BY-NC would have to be applied to all the AAC partner museums if we were interested in being able to browse across the combined contents, the minority switched to CC BY. Thus, CC BY was applied across AAC data.

Establishing URIs

During the conversion process, the museums had to determine domain identifications that are expressed through URIs.[14] To comply with this good practice, each AAC museum partner had to choose a "root" domain (the top-level hierarchy of a URI address) for the URI that would be embedded in its RDF data. The selection required forethought, because URIs must be persistent and stable. Changing a URI root domain results in broken links among the data (akin to broken links in web pages), which undermines the value of Linked Data.

Including institutional identity in the design of a root domain was also a key consideration. Every element of Linked Data that an institution publishes on the web will contain the root domain. For this reason, each AAC partner museum chose an internal URI root domain that incorporated an indicator of its institutional identity. The Walters Art Museum, for example, chose http://data.thewalters.org as its root domain, so all the museum's data elements are added to the root domain as a string. The Walters Art

Museum's bust of William Thompson Walters (ID# 32192), for example, is constructed as http://data.thewalters.org/object/id/32192. This URI is used to identify ("name") this work of art in the RDF for the Walters.

Finally, each AAC partner museum had to make sure it owned its respective domain so it could retain control over it. In the case of creating a completely new domain, ownership was accomplished by registering the domain with a registration service. The procedure is the same as registering a website domain. In the case of creating a sub-domain of an existing website, the process required adding the subdomain name (e.g., "data") to the partner's website Canonical Name (CNAME) registration.

AAC'S PROJECT TOOLS AND OUTPUTS

In keeping with the guidelines of the grant from the Andrew W. Mellon Foundation, the data management tools developed for the project are open source. The tools used are:

- Karma, designed and developed by ISI, was used to map the data to the CRM domain ontology. Karma supports a variety of source types, has a machine-learning capability to provide recommendations on the mappings to an ontology, and has an intuitive graphical interface that allows a user to visualize and refine the mappings (fig. 2). Visit https://github.com/usc-isi-i2/karma-step-by-step/wiki to see ISI's step-by-step video on how to use Karma.

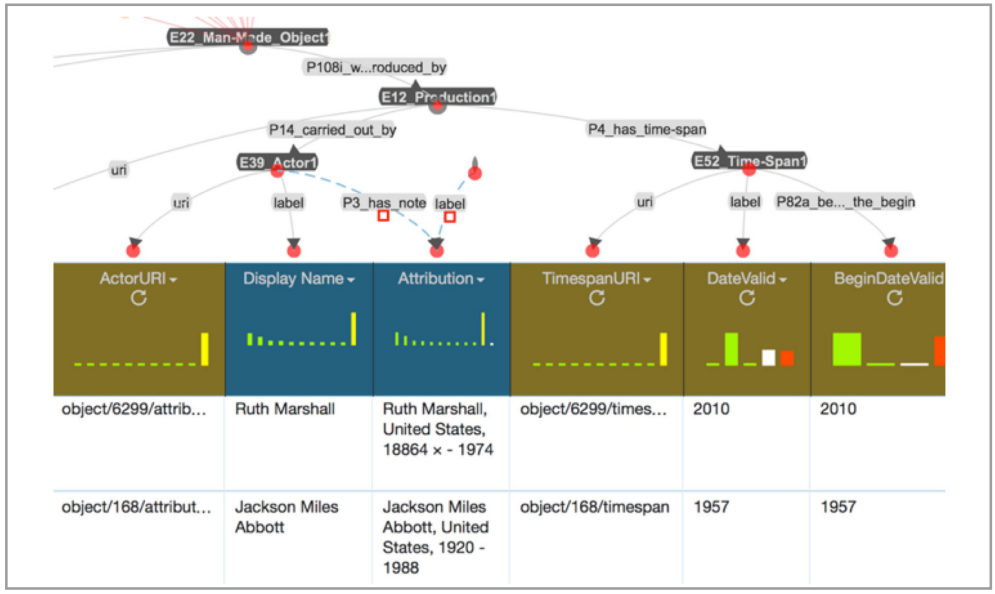

Fig. 2: Screenshot of Karma ontology mapping tool showing CIDOC CRM entities and predicates mapped to an AAC partner museum data table.

- A data generation tool within Karma was used to apply the Karma mappings to the data sets to create the RDF data and load it directly into a triplestore.
- A mapping validation tool, produced by Design for Context, provides a specification of the precise ontology mapping and a corresponding query that returns the data only if it has been correctly mapped. AAC partners used the tool to proof the accuracy of their converted data. See https://americanart.si.edu/about/

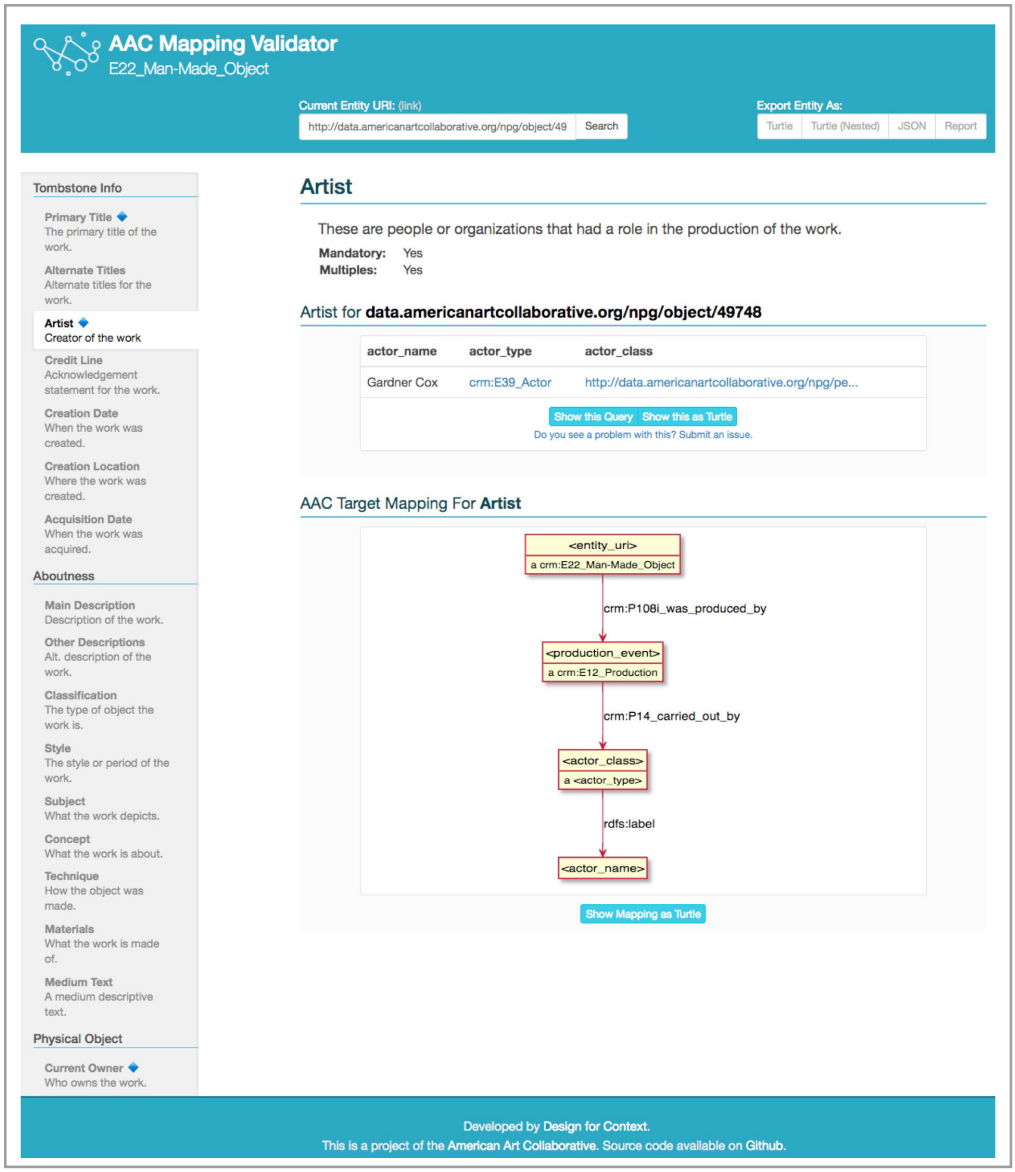

Fig. 3: Screenshot of mapping validation tool showing list of data fields on the left, with triples and human-readable values, together with a visualization of the query used to retrieve the data.

american-art-collaborative (fig. 3).

- A link curation tool designed by ISI allows nontechnical users to quickly and easily review the links to the Getty Union List of Artist Names (ULAN). AAC partners had expressed their desire to link to the Getty vocabularies to create a common reconciliation of artist names across the institutions. Each museum had to verify potential matches. The tool also produced a list of names missing from ULAN that was given to each museum to contribute to Getty at each

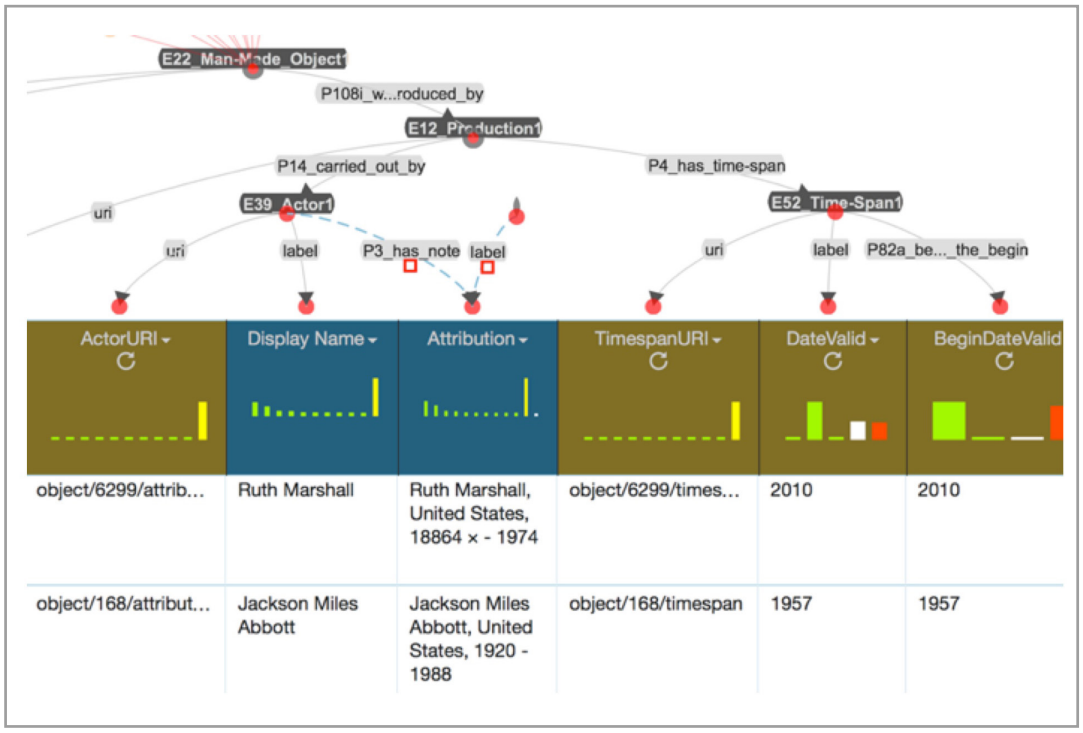

Fig. 4: Screenshot of linking tool, comparing an artist from a partner museum's data set with artist records from ULAN in the Getty vocabularies.

museum's leisure. See https://github.com/american-art/linking (fig. 4).

- IIIF Tool, produced by ISI uses LOD to create IIIF manifests (the format required to run IIIF viewers and applications) from the museum data that the AAC project has mapped to the CIDOC CRM ontology. In addition to creating the required metadata, the IIIF Tool follows the links for every image provided by each museum, determines the size of the image, and adds that data to the manifest, then creates a thumbnail of the image as part of the process of creating the IIIF manifests. See https://github.com/american-art/iiif.

Data Mapping and Conversion to RDF

To convert data to LOD, the AAC used the Resource Description Framework (RDF), a model for encoding relationships between data so that they can be rendered machine-readable via the web. After the AAC museum partners selected and prepared their data, the work flow for converting that data to RDF began with an upload of the data sets to a GitHub repository. From there the data was mapped to the CIDOC CRM ontology using Karma (http://Karma.isi.edu). As stated earlier, the data was reviewed by a CRM expert who made comments in the GitHub so that the owners of the data could learn about the CRM and follow the conversion progress. The triples were then stored in a triplestore that ISI agreed to manage for the ACC for one year after the project, after which each institution is responsible for finding its own hosting solution.

For a detailed description of the conversion process, see the paper "Lessons Learned in Building Linked Data for the American Art Collaborative" (2017) by Craig Knoblock et al. (https://iswc2017.ai.wu.ac.at/wp-content/uploads/papers/MainProceedings/382.pdf).

Browse Demonstration Application

Among AAC's goals for producing a critical mass of LOD and demonstrating LOD's usefulness is the concept of exploration across the collaboration—being able to interconnect data across institutions to support discovery and scholarship. Also important was the need to establish methods that allow federated, linked information to be useful as it grows over time.

To design and develop a usable application for exploring the data across the fourteen partner institutions, the project established a Browse Working Group comprising six of the institutions; the consultant firm Design for Context led the group, which began by identifying mutual institutional goals and taking inventory of the types of explorations that were currently difficult to do on the web. The group developed a survey to gather qualitative input via interviews with curators, registrars, educators, and outside researchers. One conclusion was, as expected, that people find it easy to identify barriers in their existing work processes but struggle to imagine alternative approaches. This finding validated the project's

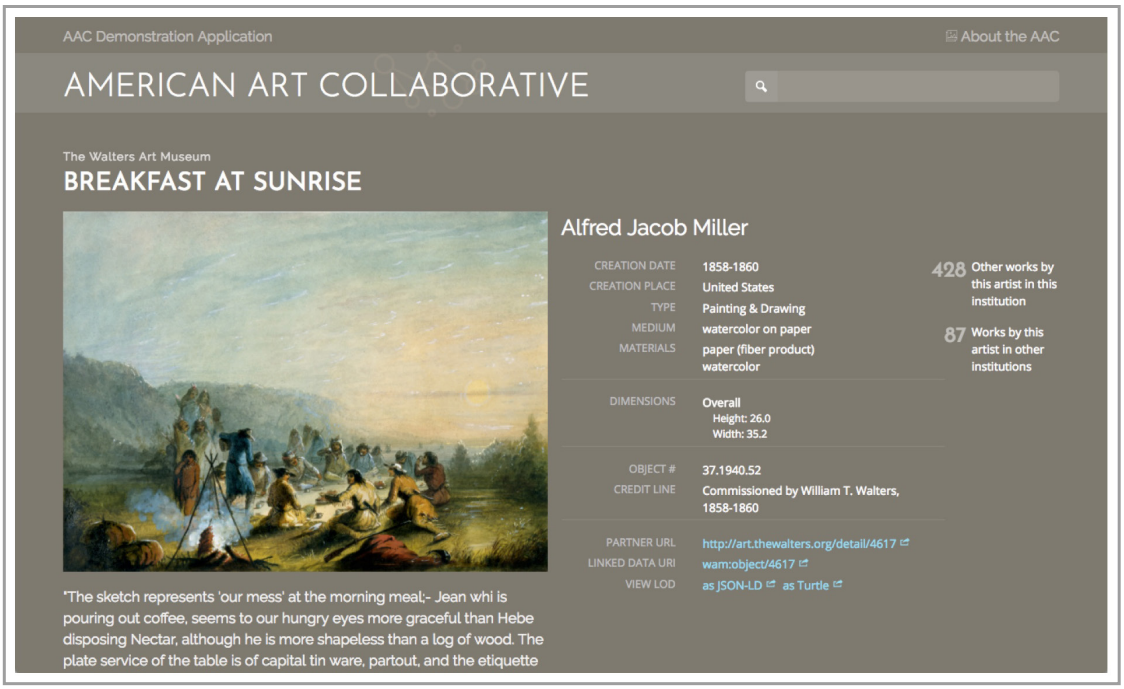

Fig. 5: Screenshot of an art object display page in the AAC browse demonstration application.

commitment to using a browse application that would present information in new ways and to assessing the effectiveness of the application among museum staff at various stages of the project.

In addition to a browse demonstration application, Design for Context produced a tool called the Mapping Validator, described above in the tools section, to check for inconsistencies in mappings and ensure that information in the triplestore can be retrieved consistently. The browse application and Mapping Validator were designed and developed in parallel, to guarantee that the data to be displayed was valuable and available. Iterative design focused on 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 (see fig. 5). The Browse Working Group determined that it would be best to use short phrases that describe the nature of the relationships, to help familiarize users with the available data. The intention was to be clear to users who might have specific research interests that match the described relationships by making it easy for them to follow Linked Data paths. Technically, to help with searches and improve performance across the growing

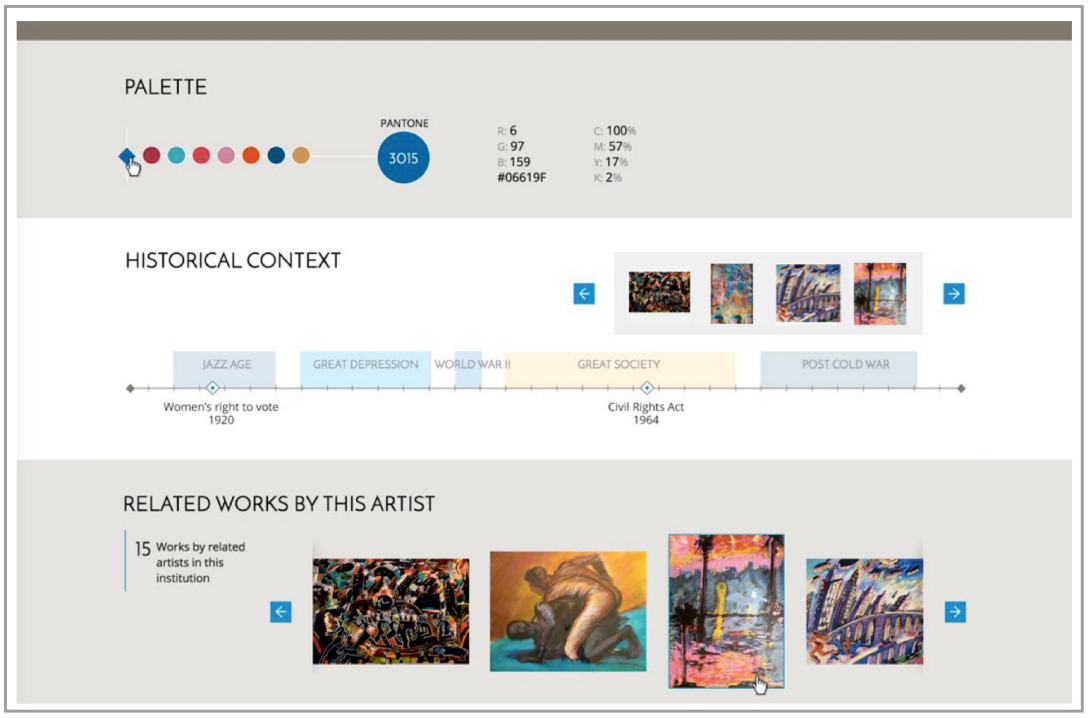

Fig. 6: An example of possible embedded "toy box" tools for the browse demonstration application.

collection of data, the browse application is populated by querying the triplestore and generating JavaScript Object Notation-Linked Data (JSON-LD) documents for each primary entity. The JSON-LD documents include all the associated link references for the entity, to allow rich cross-referencing within the browse application and populate dynamic JavaScript features that need rich data. The documents are stored in Elasticsearch so that they are easily searchable by the application, which is useful for generating lists of related artworks based on specific parameters. The triplestore remains available for more complex and ad hoc queries from the browse application.

Each page was designed to incorporate small visualizations, aggregations, and tools that help expose interesting aspects of the data about the page's artwork or artist. Nicknamed "toys in the toybox," the tools are presented below the artwork's data on the page (see fig. 6). The "toybox," for example, might allow you to select an artist and then see which institutions hold works by that artist and how many, or it might include a time line of the types of works created by an artist in the context of historical periods. Each "toy" has its own horizontal row and 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 section is shown or hidden accordingly. This allows developers from different institutions to create "toys" over time and for each institution to decide what types of display are useful to present with its entity data. The "toybox" approach extends the capabilities that are available as the data contributed by partner museums grows over time.

International Image Interoperability Framework (IIIF)

The International Image Interoperability Framework (IIIF) is a protocol for standardized image retrieval that allows a viewer to compare or contrast images. It was created by a community of the world's leading research libraries, major national libraries, and not-for-profit image repositories in a collaborative effort to produce an interoperable technology and community. IIIF defines two application programming interfaces (APIs) for interacting with museum and other image resources: an Image API that enables client applications to request images or regions of images at different sizes, and a Presentation API that includes just enough contextual information for a viewing application to present to the user, such that user knows what he/she is looking at. The nonpixel content of these APIs is Linked Data, designed as JSON-LD to be easy for developers to use without any knowledge of RDF. The approach has been highly successful, as it has resulted in broad and deep international adoption among national libraries, top-tier research organizations, museums, and for-profit companies (see http://iiif.io). The AAC browse application uses these APIs when available to provide a rich experience for the user.

Since comparing and contrasting works of art is the backbone of art history and a key function in conducting research, AAC expressed interest in IIIF. At the time AAC was established, three of its fourteen institutions were already in the process of implementing IIIF, nine were interested in learning how to do so, and two opted to consider implementation at a future date.

Under the Mellon grant, AAC partner SAAM worked with the International Image Interoperability Framework Consortium (IIIF-C) and offered interested AAC partners two workshops on how to install IIIF. One was held in Dallas, Texas, and hosted by AAC partner the Dallas Museum of American Art. The second was held in Washington, DC, and hosted by SAAM.

SUMMARY OF LESSONS LEARNED

It was clear from the outset of the project that the AAC would face many challenges, given that few AAC partners had direct knowledge or experience with LOD or semantic web technologies. Even familiarity with how to manipulate data from the existing collection information systems varied among the partner museum representatives. To some, extracting and exporting data was new. In addition, several of the museums had not thought in advance about licensing options.

AAC's Pipeline and Its Challenges

Figure 7 illustrates the process and technologies used by the AAC data pipeline to create, reconcile, and publish LOD including the development of a prototype application (browse app).

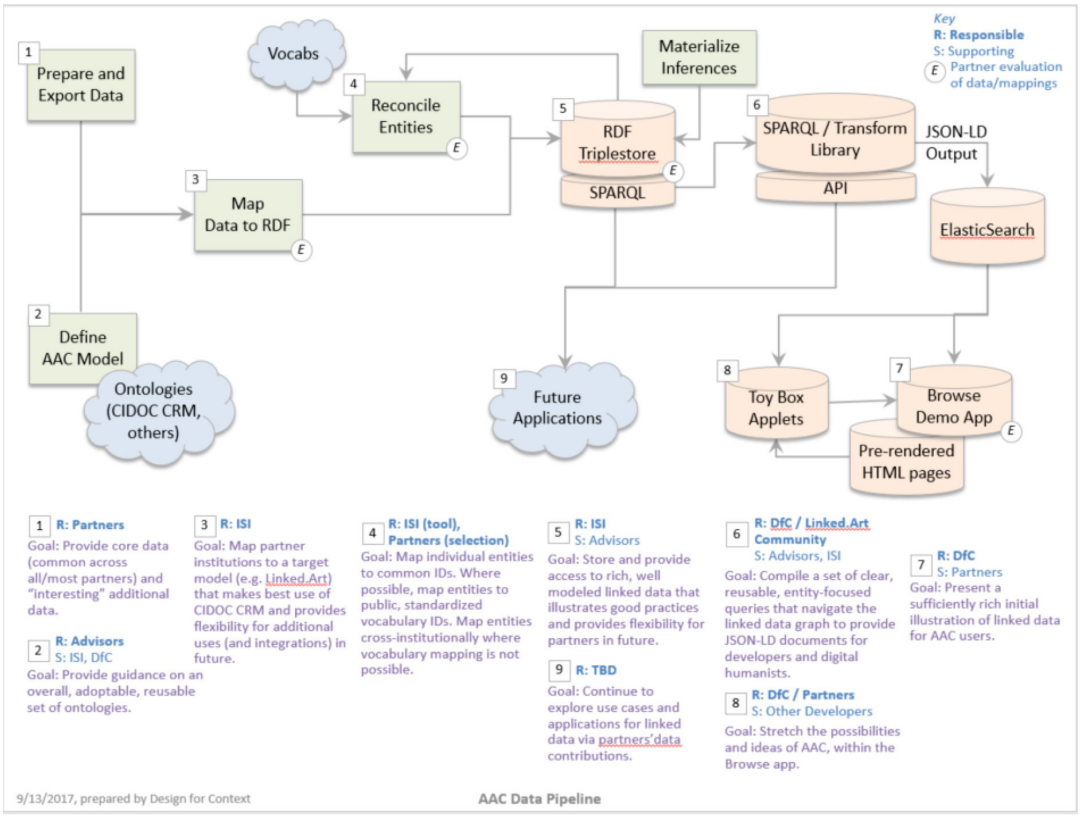

Fig. 7. Illustration of the AAC Pipeline showing functions that are performed, the repositories that have been created, and the various responsibilities of participants in the collaborative (see numbered descriptions relating to the numbered boxes).

In brief, partners exported raw data from their source systems and uploaded
In brief, partners exported raw data from their source systems and uploaded
it into a designated directory for each partner in a shared GitHub repository. Using the target model and the Karma data integration tool, each partner's data was converted to RDF, the Linked Data format. Data was then published to a triplestore, where it was accessible for review, reconciliation, and inferencing (identifying implied relationships between data entities, not just explicit links that exist). To assist developers in building applications that use the AAC Linked Data, a transform library of SPARQL queries was developed to generate JSON-LD documents as a target serialization. This additional representation of the data enhanced its usability for developers and compatibility with tools and services (e.g., Elasticsearch, HTML renderers, applets). A demonstration browse application illustrated the potential of Linked Data for exploring information about the subject of American art.

Working with Legacy Data Systems

AAC observed that exporting the data for mapping to the CRM was sometimes problematic. Collection information systems (CISs) were not produced with LOD in mind. Therefore, some do not allow data to be exported in ways that easily relate to the schema of the CRM. The CISs that do not easily support export of data have text fields, for example, but do not allow you to index or structure those fields to capture the type of information the CRM can provide. Since the CRM is highly event-driven and focuses on capturing relationships (e.g., between objects and people), it becomes frustrating not to be able to easily include this cataloguing information within the CIS so that it can then be exported for mapping. Some AAC partners discovered that to export detailed data, they needed a variety of skills including SQL. Partners also lamented the fact that in preparing their data for this project, they had difficulty inserting identifier codes into their CIS to structure and link fields to various thesauri, classifications, periods, and other references.

Conflicting Views on How to Implement the CIDOC CRM

The application of the CRM is challenging. While we cannot dispute the benefits of using a sophisticated ontology like the CIDOC CRM to provide deep research potential, the AAC experience did cause concern about its practicality for some

institutions. An attraction of the CRM is that it is strong in exploring relationships and modelling events. Museum data, however, is structured by isolated fields: artist, title, date, etc., that may be incomplete or imprecise. In some respects, CRM is a knowledge base that looks to ways museums could catalogue in the future rather than to how they currently catalogue.

In addition, a challenge for museums is the question of who produces the mapping. The CRM depends on capturing information and relationships that are not always contained in current CISs, nor would a collection manager necessarily have access to more detailed information. The CRM was created with the notion that it is best applied by curators who know the most about the objects in the museum. Is this notion realistic for most museums?

In some instances, the CRM did not seem ideally suited for RDF and LOD overall. Since the CRM was not created for LOD, some aspects of the CRM do not follow common protocols. For example, the CRM defines its own data types (string, number, date) in terms that do not map cleanly to regular data types. A number in CRM can be a range (e.g., 1.0–1.8), a format that is not used in other systems. The CRM dœs not define inverse relationships, and mapping dœs not follow best practices, such as reusing existing ontologies.

During the review process by CRM experts and advisers, it became apparent that there is more than one school of thought on how to apply and interpret the CRM. Idealists theoretically view the CRM as an ontology that could change how museums catalogue. The pragmatists see the CRM as a tool to apply to current museum practice and seek ways to make it a more malleable, flexible structure. AAC advisers have advocated for a balanced and practical approach that dœs not set the bar so high in the ideal realm that it precludes most museums from participating in future LOD initiatives.

The Need for a Target Model

Disagreements on how to apply the CIDOC CRM to AAC data, and inconsistencies in the mapping done by USC students at ISI, led AAC to recognize that it needed to develop a target RDF data model. For a deeper understanding of what issues motivated the approach to, and content of, AAC's target model, see threaded discussions within the AAC partners' respective directories in the AAC

project's GitHub (https://github.com/american-art), Linked Art (https://linked.art), and mapping validation tool (https://americanart.si.edu/about/american-art-collaborative). Although funding to produce the model had not been budgeted, AAC's adviser, Rob Sanderson, and application developer, David Newbury, began devising one. The model was then applied to the mapping being conducted at ISI, with the result that the data was mapped more consistently.

Overall, the AAC target model is a profile of the CRM based on the pragmatism that considers its application by multiple museums; accommodations for interoperability with other uses of RDF; can align with other Linked Data projects; and supports the existing online environment. The AAC target model uses rdf:value, schema:url, and other standard predicates for the interactions between AAC data and such things as strings, numbers, and websites, which are not clearly represented in CRM. In addition, the AAC target model provides flexibility for museums that may not always be able to deliver more structured data. With the CRM, the explicit way to model the size of an object, for example, is to describe each dimension as a number, a unit, and a type. Many museums, however, do not currently record dimensions that precisely. Rather than forcing each institution to parse its data manually to meet the stringent requirements of the CRM, the AAC target model offers a parallel model that uses the CRM's Linguistic Object type to record a "block of text that describes the dimensions of a work" and a term from the Art & Architecture Thesaurus (AAT) to associate that paragraph with the formal concept of "dimensionality." Both the unparsed, descriptive text and the formal, numeric model can cœxist together; having two versions means that the AAC target model can support museums with differing levels of precision in their source data. The AAC target model is thus a balance between knowledge representation and ease of use, while it has the flexibility to accommodate concepts and mappings beyond the target model. The Linked Art resource (https://linked.art) describes the philosophy that shaped the AAC target model. It indicates the model will be updated by applying it to other data sets, such as the Getty Museum and Pharos, The International Consortium of Photo Archives, with the intention of having it serve as a resource for the broad museum community.

Once completed, the AAC target model served as a road map, eliminating the discrepancies that arose in mapping owing to differing interpretations of how to implement the CRM. It was particularly effective in helping USC students, who

were not art experts, map the data correctly. Unfortunately, once the target model was produced, AAC data that had already been mapped, had to be remapped to ensure consistency across all data sets, which used up most of the grant resources targeted for mapping. To deliver a solid and consistent mapped version of the data from the museums by the end of the grant period, only essential data could be completed in the time remaining. Exhibition, bibliographic, and provenance data provided by some of the institutions had to be skipped until additional grant funding can be obtained.

Balancing a Desire for "Completeness" with Pragmatism

When first compiling their object data sets, each partner selected data from their in-house CIS based on what was already available on their websites. Once the partners examined the contributed data sets together at a team meeting, however, they noted that some institutions were contributing richer or more varied data sets than others. As the partners learned more about LOD and understood that substantial data sets yield greater potential for discovery and use, they wanted to expand their original submissions to include more object data as well as align data sets so that similar types of data would be contributed by all.

While the desire to resubmit data sets yielded a richer aggregation of data for LOD conversion, it did so at the expense of time especially considering that grants have end dates. Each partner had to extract the newly agreed-upon data from their information systems and resubmit that data set to ISI. Project tasks dependent on having data sets in hand—such as mapping the data to the CRM, or reconciling the data—were put on hold until the new data sets were submitted. The project time line was initially extended to accommodate extra time needed for the partners to contribute their new data sets. Limitations for changes were defined, however, for the reminder of the project.

Images (Quality and Rights Statements) and IIIF API

AAC partners did not need to select and submit their images to ISI for mapping. Instead, the plan was to link from data to images, using images already contained on the partner's websites. To avoid the need to manually store, size/crop, and manage each institution's images, or require the partners to do so,

IIIF was used to "wrap" existing images on the web in a common data interface to benefit the browse demo application. Contributions were made in a separate "media data file" that included image URLs and limited metadata needed for the IIIF manifests, such as label/caption, display order, rights statements, and image credits.

The process prompted renewed discussions among the partners about image use and the role of IIIF, with some participants wondering whether the use of the IIIF platform would create a need for further administrative clearances for rights. Following discussions, it appeared that the main concern was whether the image size and quality would be enhanced by using a IIIF-compatible viewer. Some of the museums had allowed open access to lower resolution (mostly thumbnail) web images, which would not be of the same high quality as images typically made available via an IIIF viewer, since the IIIF community encourages open access. Ultimately, it was decided IIIF would be used to provide a common API for the application, returning whatever sizes are provided by the source images, with no enhancement or enlargement of any kind. Each institutions' rights requirements are respected in that process, balancing the availability of images from each partner with the constraints of what resolution was possible to access.

The AAC project included many discussions about the efficacy of the IIIF platform for displaying images within the project's browse application and its potential long-term uses by the institutions themselves. Because the partners' familiarity with IIIF varied, AAC hosted workshops to bring everyone to a common understanding of the platform.

Difficulties in Understanding Museum Data

ISI hired USC students to handle the work of mapping the AAC data. Some problems arose. The students did not necessarily understand differences between data fields, such as medium and technique or subject and depicted. They sometimes omitted accession numbers, for example, because they did not realize the importance of the accession number or the need for it if a partner data set contained a primary database key. In addition, when mapping credit lines, the students initially split words such as "gift of" from the names of the donors, thus mapping the credit lines separately. The museums had to point out

that parsing the information that way could lead to legal issues if the museums had agreed, when acquiring the artwork, that the credit must always be cited in its entirety. Although the false assumptions were corrected, they could perhaps have been avoided if AAC participants had shared their business rules for their data, such as specifying which artist is primary, or what to do with variations in date syntax, and spent time with the students to review the data rules and point out variables and specific examples.

Next Steps (A Way Forward)

Members of the AAC envision that LOD will become easier for museums to implement and manage, with the result that more of our cultural-heritage data will be open and available online worldwide. A critical mass of LOD about art should invite applications that can connect the dots, particularly across data from other domains, and thereby offer new prospects for discovery and demonstrations of the value of LOD. Addressing legacy data issues and developing tools and procedures that minimize the expense of producing and hosting LOD would help simplify mapping, updating, and maintaining LOD and thereby expedite the formation of increased cultural-heritage data within the LOD cloud. Also, vendors of CISs need to modernize their systems to support more LOD needs.

AAC has played a leadership role in taking the first steps by publishing more than two hundred thousand LOD records on American art. We have provided open-source tools, lessons learned, recommendations for good practices, and a prototype browse application (demo). We intend to seek additional funding to expand application of LOD beyond the subject of American Art; to identify how best to integrate museum and archival data; to help launch LOD mapping services within the cultural heritage community; and to produce additional tools especially those that will help smaller museums engage in LOD. As more museums produce LOD, we hope they will contact us, dialogue will ensue, and opportunities will increase to link or interconnect data and further demonstrate the value of LOD.

Notes

1 For a primer on RDF, see "What is RDF and what is it good for?" by Joshua Tauberer at https://github.com/JoshData/rdfabout/blob/gh-pages/intro-to-rdf.md#Quick%20Intro.

2 Wikipedia, s. v., "Ontology (information science)," last modified August 23, 2017, 20:13, https://en.wikipedia.org/wiki/Ontology_(information_science).

3 Wikipedia, s. v., "Ontology (information science)."

4 Summarized from Stephen Stead's online tutorial for the CIDOC CRM, http://www.cidoc-crm.org/Resources/the-cidoc-conceptual-reference-model-cidoc-crm-primer, which is discussed later in my text.

5 Published by Stanford University Libraries, Stanford, CA, 2014.

6 Under Fink's leadership, the GII produced the video *Virtual Database: Art Information on the Networks*, which illustrates the value of being able to seamlessly access all the works of art and architecture associated with the creation of Saint Peters Basilica, Rome, no matter in which country or institution the information resides. See also Eleanor Fink, "Art Clouds: Reminiscences and Prospects for the Future" (paper presented at Digital World of Art History conference, Princeton University, New Jersey, July 12, 2012), https://ica.princeton.edu/pubs/2012Fink.pdf.

7 The Inventories of American Painting and Sculpture at SAAM document more than four hundred thousand artworks in public and private collections worldwide. The Inventory of American Paintings includes works by artists who were active in America by 1914. The Inventory of American Sculpture, which has no cutoff date, includes works from the colonial era through contemporary times. SAAM has also indexed and published records drawn from 1,057 exhibition catalogues dating before 1876 and including artworks exhibited in the United States and Canada through that Centennial year. The Pre-1877 Art Exhibition Catalogue Index includes artists of all nationalities, including the Old Masters, who exhibited works from 1773 to 1877.

8 Together, Allen and Fink, visited ISI, where they were presented with several new advances in technology. When they learned about LOD, it immediately struck them that art information across multiple institutions could be interconnected. Further, if the Getty vocabularies were also made available as LOD, essential resources for connecting information in the cultural-heritage community would be in place.

9 The initial six museums interested in forming a collaborative to create a critical mass of LOD on American art were Crystal Bridges Museum of American Art, Bentonville, Arkansas; Dallas Museum of Art, Texas; Indianapolis Museum of Art at Newfields, Indianapolis, Indiana; The Metropolitan Museum of Art, New York; National Portrait Gallery, Smithsonian Institution, Washington, DC; and Smithsonian American Art Museum (SAAM), Washington, DC. Later, when the AAC received a planning grant from the Andrew W. Mellon Foundation, the Metropolitan Museum withdrew owing to other commitments.

10 Speakers included Lec Maj from Yale Center for British Art, New Haven, Connecticut; Dominic Oldman from ResearchSpace; Antony Issac from Europeana, The Hague, The Netherlands; and Patricia Harpring from the Getty Research Institute, Los Angeles; Craig Knoblock and Pedro Szekely from the Information Sciences Institute, USC, Los Angeles

11 In addition to the initial museums named above in note 9, the following institutions are AAC partners: Amon Carter Museum of American Art, Fort Worth, Texas; Archives of American Art, Smithsonian Institution, Washington, DC; Autry Museum of the American West, Los Angeles; Colby College Museum of Art, Waterville, Maine; The Thomas Gilcrease Institute of American History and Art, Tulsa, Oklahoma; National Museum of Wildlife Art, Jackson, Wyoming; Princeton University Art Museum, New Jersey; Walters Museum of Art, Baltimore, Maryland;

Yale Center for British Art, New Haven, Connecticut. Colby College and the National Museum of American Wildlife Art became part of the AAC through SAAM's Commission members, who had learned of the initiative and wanted museums with which they were affiliated to participate.

12 As a federally funded agency, SAAM was not eligible to apply for IMLS funding. Instead, Fink worked with AAC partner Crystal Bridges Museum of American Art to create and submit a grant proposal for the purposes stated. The IMLS grant initially ran from September 2015 to September 2016. A no-cost extension was granted to September 20, 2017. The Mellon grant ran from December 2015 to June 2017. A no- cost extension was granted until October 30, 2017.

13 Karma was later tailored under a contract with SAAM to link SAAM data to RDF; under AAC, Karma was updated with new features to simplify accessing and downloading the tool. Karma works as a semiautomated, efficient tool that learns from routine patterns and thereby reduces the time needed for human interaction and adjustments. Since it is open source, institutions with technical staff could, in theory, download it and conduct their own mapping.

14 The purpose of converting museum data to RDF is to make it machine-readable. For data to be machine-readable, it must adhere to formatting rules that can be interpreted by software protocols. One of the rules is to use Uniform Resource Identifiers (URIs) as names for objects, people, places, and things. RDF triples (subject, predicate, object) are thus composed of URIs, not "plain English."

Part 2. Recommendations for Good Practices When Initiating Linked Open Data (LOD) in Museums and Other Cultural-Heritage Institutions

The following text is meant for museums and other types of cultural-heritage institutions but often uses "museum" to simplify wording. The text is addressed to any staff member in charge of an institution's data.

At a Glance: Recommendations for Good Practices When Initiating LOD

1. Establish Your Digital Image and Data Policies
2. Choose Image and Data Licenses That Are Easily Understood
3. Plan Your Data Selection
4. Recognize That Reconciliation and Standards Are Needed to Make Most Effective Use of LOD
5. Choose Ontologies with Collaboration in Mind
6. Use a Target Model
7. Create an Institutional Identity for URI Root Domains
8. Prepare Your Data and Be Sure to Include Unique Identifiers
9. Be Aware of Challenges When Exporting Data from Your CIS; Develop an Extraction Script, or API
10. If Outsourcing the Mapping and Conversion of Your Data to LOD, Do Not Assume the Contractor Understands How Your Data Functions or What You Intend to Do with It
11. Accept That You Cannot Reach 100 Percent Precision, 100 Percent Coverage, 100 Percent Completeness: Start Somewhere, Learn, Correct
12. Operationalize the LOD within Your Museum

1. Establish Your Digital Image and Data Policies

Before or as you begin implementing Linked Open Data (LOD) in your museum, you should establish an institution-wide agreement on the proper use of your images and data. Instituting a plan for usage may require many layers of sign-off and therefore take time to complete. The process of working on your data and images in an LOD initiative

may help your organization extend its thinking about what is possible and desirable for stakeholders and constituents.

LOD is about being able to interconnect your museum's digital records with those in other collections. Although you may choose to convert only your data to LOD, you will likely be interested in eventually including images. A few good examples of statements on digital image usage include the Policy on Digital Images of Collection Objects Usage formulated by the Walters Art Museum, Baltimore, Maryland (https://thewalters.org/rights-reproductions.aspx), and those devised for the Yale Center for British Art, New Haven, Connecticut (http://britishart.yale.edu/collections/using-collections/using-images), and the National Gallery of Art, Washington, DC (https://images.nga.gov/en/page/openaccess.html). You are also encouraged to consult RightsStatements.org, described in the next section, on licenses.

2. Choose Image and Data Licenses That Are Easily Understood

A valuable resource for rights on image usage is http://RightsStatements.org. It focuses on a range of common international options for image rights that the museum community will likely increasingly consult and support.

For data, you will need to provide a license that clearly states how others may use your museum's data. When you engage in LOD, the "open" part means that you are allowing public use. The most widely adopted licenses recognized worldwide are the Creative Commons (CC) licenses, which have been developed specifically for the distribution of data via the web (and thus internationally). A CC license conveys the right to the public to share and build upon a published work (see https://creativecommons.org/licenses). Several types of licenses, each with pros and cons, are available. CC0 allows full use with no restrictions. The other CC licenses offer a set of permissions that you may select individually or in combination:

- CC BY requires that attribution always be cited (e.g., "created BY this person/institution")
- CC NC allows only noncommercial use (e.g., "you cannot sell this content or derivatives that you make from this content or use it in commercial projects/presentations")
- CC SA requires that you "share alike," meaning that anything produced by others based on your content must carry the same license conditions that you have given in your license, no more and no less.

Depending on your institution's needs, you can variously combine these three permissions to fashion the license you want. CC BY-NC ("provide credit to us for the use of our data and only use this noncommercially") and CC BY-NC-SA ("provide credit to us for the use of our data, make whatever you produce available to others with these same license conditions, and use this only non-commercially") are common combinations seen on websites. Other licenses have a number after the CC (e.g., CC4) to identify the version. Previous licenses remain valid, and updated licenses do not take away license rights from earlier versions.

Any of the more restrictive versions may make sense for your museum. If you are part of a collaborative multi-institutional project, however, that choice may require your partners to accept the more restrictive terms as well, as the data query or page display may expose content from multiple institutions.

3. Plan Your Data Selection

Museums are the purveyors of vast information resources. Object information is the most obvious to make available as LOD, but equally important data comes from bibliographic, archival, exhibition, curatorial, and conservation sources. In addition, within each of those categories are quantitative data (dates, dimensions) and narrative data (object descriptions, curatorial notes, educational content). Selecting which data to contribute to an LOD project requires that you carefully consider project goals and time frames and pragmatically assess what is achievable. At the same time, you want to balance your short-term objectives with the long-term aim for the LOD to serve multiple purposes and align with other institutions in the future.

While LOD can provide rich results with a full set of data, converting all the data related to a theme or collection at one time is rarely feasible. Limitations on resources and legal or administrative constraints can render some relevant data unavailable for LOD projects. Curatorial records, for example, offer a wealth of information but may be more proprietary in nature than object data from a collections database. Since it simply may not be realistic for your collection management and digitization plans to include all your data initially, you may prefer an incremental approach—beginning with the basic label copy, or "tombstone" data, for an object, and later adding more descriptive and educational

data. When making your choices, consider how the data will be used, particularly in combination with partners and other institutions.

Depending on an institution's size, a varied group of professionals might be needed to define and identify the appropriate content for an LOD project. The team may be drawn from the ranks of information technology, collections, curatorial, education, and design departments, among others.

4. Recognize That Reconciliation and Standards Are Needed to Make Most Effective Use of LOD

While you may wish to maintain local standards for use in your institution, remember that LOD is about data that is open. One of the key benefits of LOD is its capacity to link data across collections! Opening the usage of your data is part of increasing your institution's visibility. Scholars, the public, other institutions, and developers may wish to link to your data and/or create applications.

If your LOD consists mainly of sketchy and/or unstructured data (nonstandard vocabularies, text strings without unique identifiers, etc.), it will diminish the potential to interconnect the information for global use and be difficult to reconcile with other linked resources, such as the Union List of Artist Names (ULAN), one of the LOD projects of the Getty Vocabulary Program by the Getty Research Institute, Los Angeles. You will want the name, of an artist, and additional descriptive information to be precise enough to determine a match with an artist listed in ULAN.

Resolving decades-long problems with legacy data—such as the disparate ways information on dates, dimensions, titles of works of art, "unknown" values, and other basic details about objects has been recorded—is challenging but critical. If the cultural-heritage community wants to share information, it needs to identify solutions and seek broad agreement around the problematic issues of legacy data.[1] Consider working with organizations such as American Association of Museums (AAM), Arlington, Virginia; Museum Computer Network (MCN), New York; and International Council of Museums (ICOM), Paris, and/or apply for grants to establish community consensus on recommended solutions and tools for these information-sharing obstacles.

5. Choose Ontologies with Collaboration in Mind

From the outset you will need to decide what ontology or ontologies you will deploy to take full advantage of the precision, or "semantic glue," LOD provides. If you choose to use more than one ontology, which is likely, make sure your mapping tool can handle multiple ontologies.

Not many ontologies are specific to the discipline of cultural heritage. The two that museums most commonly adopt are the Europeana Data Model (EDM) and the Conceptual Reference Model (CRM) produced by ICOM's International Committee for Documentation (CIDOC). Optionally, a museum can adapt an ontology to create a profile that best suits its needs, such as the target model that the AAC formed to simplify the CIDOC CRM, and/or incorporate additional, commonly used ontologies from the web community (see https://linked.art). It is important to emphasize that the AAC target model is not another ontology. As stated, it is a profile of the CRM. Note that new and evolving ontologies continue to be produced. The archive and library communities have their own ontologies, which many other types of institutions will want to incorporate into their LOD. One of the many benefits of using LOD is that it is feasible to model data in ways that incorporate multiple ontologies for specific purposes, if necessary.

If you choose the CIDOC CRM, recognize that it poses challenges. In some cases, the CRM's ability to capture details will depend on the availability of curators or scholars to provide information and express appropriate relationships. Nevertheless, just because the CRM was created to work with cultural information at a highly detailed level dœs not mean it is not helpful if applied more generally. The CRM concepts are structured in hierarchies. One can choose to start on a more general level. Some users of the CRM may argue, however, that a museum should apply the CRM as deeply as possible at the first opportunity, partly because the institutions' focus and funding to prepare additional data may not come around again. Plan accordingly.

6. Use a Target Model

Whether your institution is working alone or on a collaborative LOD initiative with other museums, developing or using an existing target model for mapping your data is a top priority (see https://linked.art for the target model that AAC

developed). The model should be a subset of all the mapping possibilities relevant to your data. The model helps eliminate guesswork, keeps the mapping consistent, and significantly reduces the modeling and design effort required in the project. It also provides a reference that developers can use across multiple projects.

7. Create an Institutional Identity for URI Root Domains

Uniform Resource Identifiers (URIs) are unique identifiers that designate objects, people, places, and things in a way that can be read by computers. They are key components of LOD. Thus, Resource Description Framework (RFD) triples—the underlying data format for LOD—are composed of URIs, not "plain English." To establish authority and persistence for the data you are converting to LOD, you should select an institutional root domain (the top-level hierarchy of a URI address). Selecting the root domain requires forethought. Changing root domains results in broken links among the data (akin to broken links in web pages), which create problems for those who will rely on that data in the applications they develop.

When selecting a URI root domain, consult with your IT staff to ensure your chosen name identity is not being considered elsewhere in the museum for other purposes. To avoid an overlap, it is a good practice to identify a single root domain that will be used throughout the institution for all its current and future LOD. A discrete subdomain (e.g., "data.thewalters.org") or a completely new domain (e.g., "thedigitalwalters.org") distinct from an institution's website URL (e.g., "thewalters.org") is recommended.

A second consideration is ownership of the root domain. You want proprietorship so that you can retain control over it. In the case of creating a completely new domain, ownership can be accomplished by registering the domain with a registration service. The procedure is the same as registering a website domain. Creating a subdomain of your existing website requires adding the subdomain name (e.g., "data") to your website's Canonical Name (CNAME) registration.

8. Prepare Your Data and Be Sure to Include Unique Identifiers

It is important to review and clean up inconsistencies in data structures, formats, and values where possible, as irregularities will cause problems for the mapping and conversion of the data to LOD. Make sure you have completed filling in all your data

categories and that you have addressed outstanding issues. Check for spelling errors and content inconsistencies. (Also see recommendation 9 below for additional steps that may be needed to prepare your data.)

You should always make sure your data includes unique identifiers that do not change. For art objects, the identifier may be an accession number or other unique identifiers generated by your collection information system (CIS); use whichever is the most stable and unchanging. Look at examples from existing LOD data sets at other institutions and consult https://linked.art for guidance.

9. Be Aware of Challenges When Preparing and Exporting Data from Your CIS; Develop an Extraction Script, or API

When preparing museum data for its conversion to LOD, you must often transform or alter the format and structure of that data before, during, or after extracting it from the museum's in-house collection information system (CIS), so that you can readily convert it or map it to a semantic model. You might need to reformat dates, for example, or parse measurements so that you can place each element (height/width/depth) in a separate field. To export the data, you may need to use special formats—such as JavaScript Object Notation for Linked Data (JSON-LD), computerized system validation (CSV), or Extensible Markup Language (XML)—which can be ingested by the specific software tools you are using in the LOD project.

Regrettably, some CISs are implemented in ways that do not structure data in formats that facilitate mapping to ontologies such as the CIDOC CRM. Some CISs rely heavily on text fields that are narrative, or human-readable. While provisions for adding a structured linked layer to complement the text fields might be available, the layers can be cumbersome to apply and access.

It is possible that your museum's CIS has embedded Thesaurus tools, for example, that may not be implemented in ways that support LOD cataloging. Some CIS platforms contain the Getty vocabularies, but not always the latest versions, such as LOD. The difficulty of obtaining access to current thesauri prevents information managers from capturing and applying terms that reflect recent research and geographical boundaries.

It is important to encourage your CIS vendor to simplify the processes of entering structured data and getting data out of the system in ways that can be used for LOD. CIS vendors have been responsive in the past in respect to upgrading systems to reflect new standards and directions that have broad adoption and "staying power." As more museums select LOD for sharing their data, CISs also need to be updated. The community's insistence will help move the tools forward for everyone.

Given the complexities of extracting and exporting data, once your museum identifies and addresses the issues that arise with in-house data preparation and extraction, you should aim to construct a work flow for the process and automate as much of it as possible. A scripted extraction method, or application programming interface (API), will minimize the effort it takes for a museum to incorporate updates into its LOD at routine intervals.

10. If Outsourcing the Mapping and Conversion of Your Data to LOD, Do Not Assume the Contractor Understands How Your Data Functions or What You Intend to Do with It

Companies are starting to offer outsourcing of mapping and data conversion. Always check if the vendor has had experience working with museum data. Be prepared to invest time up-front orienting and providing your data rules to the people doing the work so they understand the nuances of the data they are handling. Some museum data can be hard to comprehend for those not educated in its norms, particularly when it may contain ambiguities, such as in date ranges, attributions, and even identifiers. Mapping and conversion mistakes can result, wasting the investment you make in LOD.

11. Accept That You Cannot Reach 100 Percent Precision, 100 Percent Coverage, 100 Percent Completeness: Start Somewhere, Learn, Correct

Museums are traditionally reserved and cautious about releasing their data to the public. Most museums do not want to publish data until it is "complete," however that is defined. Concerns can arise about correctness—especially when dates for an artist or work are ambiguous, with "authorities" providing different information. To some extent, the ubiquity of the World Wide Web and social media is changing restrained attitudes toward publishing information online, one reason being that data can now be easily updated, which is the case with LOD.

Remember that your LOD initiatives can be incremental. Particularly when using a target model, you can, over time, add data, which can include deeper detail as well as new types of information.

Consider that the data you are managing for LOD dœs not sit in isolation of other data and content in your museum. Other initiatives in your museum, such as implementation of the International Image Interoperability Framework (IIIF) and educational and scholarly publishing on your website, may create opportunities for integration with your LOD. All museum applications and websites require museum data in some form, so it makes sense to plan for evolution as new requirements emerge.

12. Operationalize the LOD within Your Museum

Once you have converted your museum's data to LOD, make effective use of it and operationalize it across the museum. LOD can serve as a master resource for many of the digital applications you use to reach your audiences. As a starting point, you could update existing online collections websites and digital interactives so they can draw from your institution's LOD.

You could also set the stage for instituting new cataloging practices. Consider cataloging LOD identifiers in your CIS and building reconciliation into early cataloging work by capturing IDs from, for example, the Getty vocabularies—The Union List of Artist Names (ULAN), the Art & Architecture Thesaurus (AAT), and the Getty Thesaurus of Geographic Names (TGN)—alongside the terms and descriptions you use in cataloging. Make sure that narrative and descriptive fields are complemented by structured data (artist, title, date, etc.). Finally, set up an automated system for refreshing your LOD as new records are added, much as many museums automatically update their website data on a nightly basis.

Note

1 Examples of legacy data issues include:

Dates: The inconsistency in date form has been an issue when extracting a useful and correct value or assigning an accurate relationship. Data for an artist's birth and/or death are sometimes exact, but data sometimes includes only the year or words such as "born," "active," or "circa" embedded in the data field, which makes extracting the numeric value complex.

Dimensions: Some museums have dimensions of artworks entirely separated into length, width, and depth, with the measurements and the units of measure. Most museums, however, only have a paragraph containing all the measurement information, which is difficult to parse.

IDs: Unique identifiers for important data are essential for constructing URIs for things like objects and people. Most museums include IDs for artworks and/or artists, but many forget to include them. Moreover, objects may have two IDs: the accession number (which can change) and a system ID (which can change as systems are replaced).

Part 3: FAQs

"FAQs" and "Recommendations for Good Practices" (parts 2 and 3 of this guide) are each separately available on the AAC website, https://americanart.si.edu/about/american-art-collaborative.

What is Linked Open Data (LOD)?

In computing, Linked Data describes a method of publishing structured data so that it can be interlinked and therefore useful in web implementations. Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term. Linked Open Data (LOD) refers to data that is made available for public use via Linked Data.

What is the Semantic Web?

"Semantic Web" refers to the vision and set of technologies that would enable identities to be linked semantically via the web so that accurate web searches are possible. To achieve the "semantic glue" that provides context and meaning in documents, W3C's Resource Description Framework (RDF) is used to tag information, much like Hypertext Markup Language (HTML) is used for publishing on the web. RDF is a W3C standard model (i.e., web based and web friendly) for interchanging data on the web. RDF breaks down knowledge into discrete pieces, according to rules about the semantics, or meaning, of those pieces, which it represents as a list of statements with three terms: subject, predicate, and object, known as triples. Each subject, predicate, and object is a Uniform Resource Identifier (URI) or, for the object, a literal value such as a number or American Standard Code for Information Interchange ASCII string. An organization that produces LOD must select one or more ontologies to play the key role of defining the meaning of the terms used in the subject/predicate/object statements. In essence, RDF and the ontology give context and meaning to a statement, and the URIs provide the identifiers for all the entities described and published as LOD, which allows them to be discovered and

connected. Approximately 149 billion triples are currently in the LOD cloud.

What are RDF and URI?

As stated above, Resource Description Framework (RDF) is a W3C standard model (i.e., web based and web friendly) for interchanging data on the web. RDF is a way for computers to work with facts and express statements about resources. RDF breaks down knowledge into discrete pieces, according to rules about the semantics, or meaning, of those pieces, which it represents as a list of statements in the form of subject, predicate, and object, known as triples. The format mimics English sentence structure, which accommodates two nouns (the subject and object) and a type of relationship expressed as a verb (the predicate).

Each subject, predicate, and object is a Uniform Resource Identifier (URI) or, for the object, a literal value such as a number or ASCII string. A URI is typically expressed as a Uniform Resource Locator (URL), which provides the location of the identified resource on the web.

What is an ontology?

Simply stated, an ontology is a schema or conceptual framework that gives data meaning. As described by Wikipedia, an ontology in the field of information science is the "representation of entities, ideas, and events, along with their properties and relations, according to a system of categories."[1] An ontology provides a set of definitions for the meaning of URIs used in Linked Data. Ontologies, which are defined using specifications standardized by the W3C, allow the definition of classes, relationships, and properties to be used in a data model.

What is the CIDOC CRM?

CIDOC CRM, the Conceptual Reference Model (CRM) created by the International Committee for Documentation (CIDOC) of the International Council of Museums, is an extensive cultural-heritage ontology containing eighty-two classes and 263 properties, including classes to represent a wide variety of events, concepts, and physical properties. It is recognized as International Organization for Standardization (ISO) 21127:2006. For more information about the CIDOC CRM, visit the website http://www.cidoc-crm.org.

Is all LOD linkable, regardless of which ontology is used?

Yes, all LOD is linkable because the links are between the subject and object entities regardless of the ontology and syntax that are used to model the data. You can therefore link artists in your specific ontology format to the Getty's Union List of Artist Names (ULAN) or DBpedia, for example, and link places to GeoNames or the Getty Thesaurus of Geographic Names (TGN), and the like. You can also link your object LOD to information about an object from another museum that is in LOD, even if that institution uses a different ontology.

What are a triplestore and SPARQL endpoint?

A triplestore is a purpose-built database for the storage and retrieval of triples (statements in the form of subject, predicate, and object) through semantic queries. Triplestores support the Semantic Protocol and RDF Query Language (SPARQL).

SPARQL is a query language used to search RDF data, and the endpoint is a networked interface created to facilitate these queries. An endpoint is used to query a triplestore and deliver results, primarily for researchers with specific queries.

What is GitHub, and is it necessary to use it?

GitHub is a web-based data and code-hosting service. AAC's GitHub was the central working space for producing LOD. Each participating museum had a separate directory in GitHub in which to store its exported data. As the students working at the Information Sciences Institute (ISI) at University of Southern California (USC), Los Angeles, mapped the data, the AAC museums, consultants, and advisers could track their progress and post comments within each museum's repository in GitHub. While GitHub was a key component for AAC, it is not necessary to use GitHub to produce LOD.

What is a target model?

A target model is a subset of all your data mapping possibilities. It acts like a road map when mapping data to minimize the guesswork and help provide consistency across your data. You will need to select your ontology or ontologies before creating

or using an existing target model.
If you chose the CIDOC CRM, remember that there are different schools of thought about how it should be applied. Your target model should reflect how you wish to apply the CRM.

What is AAC's target model?

AAC's target model is a profile of the CIDOC CRM Linked Data model designed to work across many museums and enable functional applications to be built using the model. More specifically, because the AAC target model is standardized, many institutions can use it to publish and share their data with little reworking of data. The AAC target model supports varying levels of completeness, or detail, in the data, as it uses the CIDOC CRM alongside other RDF ontologies available outside the museum community, where needed. It is closely aligned with common controlled vocabularies (currently, the Getty's Art & Architecture Thesaurus). The AAC target model minimizes the learning curve about LOD modeling for institutions' staff. It also supports the important concept that the data should be able to travel "round trip," meaning that the data converted from source systems to Linked Data can be converted back into the format of the original system with no loss of data or change in the level of detail.

The AAC target model currently covers 90 percent of the possibilities the CRM offers, with only 10 percent of the complexity of the full CRM ontology.

What decisions shaped AAC's target model?

Several challenges influenced the shape of AAC's target model. Among them were issues of legacy data, which partners could not resolve within the scope of the project; the complexity of the CIDOC CRM, which depends on details that partner data did not include; places where the CRM did not work well with the aim of the web delivery of LOD, so other ontologies need to be incorporated alongside CRM; and a conscious effort to make data mapping basic for the museum community. A full description of AAC's target model can be found at https://linked.art.

The AAC GitHub link captures the dialogue that influenced AAC's target model. These discussions can be found within individual partner directories at https://github.com/american-art.

What is Linked Art?

Linked Art describes itself on its website (https://linked.art): "Linked Art is a Community working together to create a shared Model based on Linked Open Data to describe Art. We then implement that model in Software and use it to provide valuable content. It is under active development and we welcome additional partners and collaborators." Linked Art provides patterns and models that enable cultural-heritage institutions to easily publish their data for event-based digital research projects and non-cultural-heritage developers. It includes the AAC target model and applies it to additional projects, such as those developed by the J. Paul Getty Museum, the Getty Research Institute (Getty Provenance Index), and Pharos, an international consortium of fourteen European and North American art historical photo archives.

What is IIIF?

The International Image Interoperability Framework (IIIF) is a set of agreements for standardized image storage and retrieval. A broad community of cultural-heritage organizations and vendors created IIIF in a collaborative effort to produce an interoperable ecosystem for images. The goals of the project are to:

- Give scholars an unprecedented level of uniform and rich access to image-based resources hosted around the world
- Define a set of common application programming interfaces (APIs) that support interoperability between image repositories and supporting image viewers
- Develop, cultivate, and document shared technologies, such as image servers and web clients, that provide a world-class user experience in viewing, comparing, manipulating, and annotating images.

For more information about the IIIF, visit the website http://iiif.io.

What open-source tools are available to map, produce, review, and reconcile LOD?

The Open Community Registry of LOD for GLAM (Galleries, Libraries, Archives, Museums) Tools is a good resource. For information, see the spreadsheet https://docs.google.com/spreadsheets/d/1HVFz7p-8Rm3kmDK0apMsrwV_Q0BaDSsw59hLViJSnvs/edit#gid=0. Some key mapping tools for museums are:

- 3M: The 3M online open-source data mapping system has been jointly

developed by the Foundation for Research and Technology–Hellas (FORTH) information systems laboratory in Greece and Delving BV in The Netherlands. It is touted as a tool that allows a community of people to view and share mapping files to increase overall understanding and promote collaboration between the different disciplines necessary to produce quality results. It was expressly designed for making good use of the CIDOC CRM. For more information about 3M, visit the website http://www.ics.forth.gr/isl/index_main.php?l=e&c=721.

- Karma: Karma is an open-source integration tool that enables users to quickly and easily integrate data from a variety of data sources, including databases, spreadsheets, delimited text files, Extensible Markup Language (XML), JavaScript Object Notation (JSON), Keyhole Markup Language (KML), and web APIs. It was developed by professors Craig Knoblock and Pedro Szekely of USC's ISI. Users integrate information by modeling it according to an ontology of their choice using a graphical interface that automates much of the process. Karma learns to recognize the mapping of data to ontology classes and then uses the ontology to propose a model that ties together the classes. Users then interact with the system to adjust the automatically generated model. During the process, users can transform the data as needed to normalize data expressed in different formats and restructure it. Once the model is complete, users can publish the integrated data as RDF or store it in a database. A video explaining how Karma works is available at the website http://karma.isi.edu.
- A data generation tool within Karma applies mappings to the data sets to create the RDF data and load it directly into a triplestore.
- A mapping validation tool, produced by Design for Context, a consultant for AAC, provides a specification of the precise ontology mapping and a corresponding query that returns the data only if it has been correctly mapped. For more information about the mapping validation tool, see https://americanart.si.edu/about/american-art-collaborative.
- A link curation tool, designed and produced by ISI, allows users to review links to other LOD resources such as ULAN. See https://github.com/american-art/linking.
- IIIF translator tool, produced by ISI, automatically creates IIIF manifests (the format required to run IIIF viewers and applications) from the museum data that the AAC project has mapped to the CIDOC CRM ontology. In addition to creating the required metadata, the IIIF translator follows the links for every image provided by the museums, determines the size of each image, and creates a thumbnail of each image as part of the process of creating the IIIF manifests. See https://github.com/american-art/iiif.

Which tools did AAC use?

AAC used ISI's Karma to map all the data and produce RDF. ISI's link-curation tool was used to link AAC names to ULAN. Design for Context's validator tool was used to proof the AAC data for consistency.

What is the difference between an ontology and the Karma mapping tool?

As stated earlier, an ontology is a schema or conceptual framework that gives data meaning. Karma is a tool for mapping data to an ontology and using that mapping to produce RDF.

Dœs our institution need to have a usage policy in place for our data and images before engaging in LOD?

Developing an institution policy requires time for all the appropriate signatures and authorizations. It is recommended, therefore, that your institution develop a digital usage policy before or as soon as you engage in LOD. In theory, you could produce LOD without including images, but most museums will want to provide public access to its images, so having a usage policy for images as well as for data is important.

What are my options for selecting a license?

When you engage in LOD, you should choose a Creative Commons (CC) license, which enables the right to share, use, and build upon a work. For more information about CC licenses, visit the website https://creativecommons.org/licenses, which outlines various types of licenses that are available. For a better understanding of the choices, see recommendation 2, "Choose Image and Data Licenses That Are Easily Understood," in part 2 of this guide, "Recommendations for Good Practices" (also posted as separate document at https://americanart.si.edu/about/american-art-collaborative). You may wish to select different licenses for images and data. In some cases, you may even decide to apply different licenses to different images.

Which person in a museum should undertake the mapping of data to LOD?

The staff members involved in mapping your museum's data to LOD will depend

on how your museum is organized and the technical knowledge of the staff. A team effort is ideal. The actual mapping process and use of a tool such as 3M or Karma (http://karma.isi.edu) will require some technical knowledge, which could come from information technology staff. To benefit from the depth of coverage of the CIDOC CRM, curators as well as collection managers and registrars with in-depth knowledge of your museum's data should be involved.

What technical skills are helpful for staff who are publishing and maintaining LOD?

Knowledge of JavaScript Object Notation for Linked Data (JSON-LD), RDF, RDF databases called triplestores (see FAQ "What are a triplestore and SPARQL endpoint?"), ontologies, and web infrastructure is an asset for staff participating in LOD. To make the processes for converting and publishing LOD sustainable over time, the technical team involved should understand scripting pipelines that can manage the conversion and movement of data between repositories.

How can my institution begin long-range planning for LOD?

Developing an institutional policy for publishing data and images and choosing a license would help set the stage for your museum's LOD initiative. A good example can be seen on the website https://thewalters.org/rights-reproductions.aspx for the Walters Museum of Art, Baltimore, Maryland. The museum's Policy on Digital Images of Collection Objects, states: "The Walters Art Museum believes that digital images of its collection extend the reach of the museum. To facilitate access and usability, and to bring art and people together for enjoyment, discovery, and learning, we choose to make digital images of works believed to be in the public domain available for use without limitation, rights- and royalty-free."

Planning what data you would like to convert to LOD and reviewing your institution's data for completeness would be the next logical steps. See recommendation 3, "Plan Your Data Selection," in "Recommendations for Good Practices" (part 2 of this guide and a separate document at https://americanart.si.edu/about/american-art-collaborative), which covers planning, exporting your data, and legacy data issues.

It would also be useful to educate yourself about LOD by reviewing the educational

briefings, papers, and presentations posted on the AAC website.

What minimum data is needed to begin LOD, and can I add more data in phases?

A museum could choose to begin with the basic descriptions of its objects, sometimes called "tombstone data," and add more data over time. Others may wish to take a project-based approach by using LOD to develop a theme in depth, such as focusing on a certain artist or collection.

If I need to update my data, do I have to remap everything?

If you are using Karma (see http://karma.isi.edu) to map your data, many updates to the data can be made without requiring additional work. The three possible scenarios are:

1) The format of your updated data is the same as the data that was published. In other words, you have updates to some of the data values but no additional fields or other changes. In this case, the Karma model that you built will apply directly to the updated data without any additional changes.

2) The format of the data has changed for a data set that was previously mapped. In this case, you will need to load the revised data set into Karma. You can apply the earlier Karma model to this data, but the model will need to be updated to reflect the changes in the data set. Once this is done, you will need to save the model and apply it to future versions of this data set only if the data values change.

3) The format is a new data set that you have not previously modeled in Karma. In this case, you will need to load the new data set into Karma, then construct and save a model of the data set.

Note that you can save time updating if you prepare a script to automate your work flow for extracting your data before mapping. Scripts would also be helpful for taking the converted data and moving it into the triplestore and other repositories. Scripted methods will minimize the effort a museum must make to incorporate updates into its LOD at routine intervals.

Do I need to link to the Getty vocabularies before I convert our data to LOD?

It is not necessary to link to the Getty vocabularies before converting data to LOD. Should you wish to link to the Getty vocabularies before mapping, please note that

the Getty Vocabulary Program's website contains a quick reference guide. See http://www.getty.edu/research/tools/vocabularies/vocab_contributions_flier.pdf.

How do I link to the Getty vocabularies once I am using LOD?

The Getty Vocabulary Program is working with AAC to produce an API that will streamline submission of LOD vocabularies and return the URIs or IDs to the contributing institutions for incorporation into their collection information systems (CISs).

What are some of the LOD resources to consider for linking?

The most common LOD resources today are DBpedia, the Getty vocabularies, GeoNames, and Virtual International Authority File (VIAF).

Once I convert my museum's data to LOD, can I interconnect it with LOD from other museums?

Yes! The "Linked" part of Linked Open Data is a critical component, but it dœs not occur automatically. As web pages link to one another by adding in the links to the HTML, so can LOD link to other LOD resources. For example, you can choose to link your data to related resources as mentioned in the FAQ above.

Will my institution lose its identify or authority over usage of our object information with LOD?

The idea of LOD is to make it widely available so that other institutions, scholars, and the public can connect with it or use it. It is thus important to state clearly your institution's license conditions, if any. The data published as Linked Data from a museum should use a URI that designates the museum's identity. Anyone using the data will give authority to the institution that is publishing the data.

Approximately how much time dœs it take to convert museum data to LOD?

The time involved in converting museum data to LOD will largely depend on the ontology or ontologies you choose for mapping, the tool you select for converting the data, who performs the mapping, their familiarity with your data, and their level of expertise in using mapping tools. The Yale Center for British Art estimates

that it took about two years to map some fifty thousand objects (paintings, sculpture, prints, drawings, watercolors, and frames). The time included the intellectual mapping, writing/code for the script that dœs the transformation, and putting in place the triplestore and various other pieces of their digital infrastructure to share their CRM-based RDF dataset.

In the case of the AAC, the mapping was outsourced and corrected at intervals as needed. The AAC grant ran for eighteen months, during which time the participants attended many educational briefings, workshops, and in-person meetings and dedicated time as needed for preparing and extracting their data, proofing data, and performing similar tasks. The average time spent by the AAC participants was 169 hours over the duration of the grant.

What resources will I need to host LOD?

As an entry-level approach, a museum could choose to produce JSON-LD and put it in a web server to be served statically or via another software application that works with that format (the AAC browse demo application uses ElasticSearch). In the case of AAC, the initial recommendation was for each museum either to have its own server that could support a triplestore/SPARQL endpoint or form a hub and share a SPARQL endpoint (see FAQ, "What are a triplestore and SPARQL endpoint?").

During the production phase of AAC, ISI hosted the data using the SPARQL server Apache Jena Fuseki 2+, which is free; for details, see https://jena.apache.org/documentation/fuseki2/index.html. Note that an institution could also use a cloud-computing solution and would therefore be renting space, not requiring any hardware.

How do options for hosting LOD influence how it is accessed and used?

Once the data is generated, people will want to access the data in several ways. One is via a SPARQL endpoint, through which a developer can run Linked Data queries directly against the LOD graph. This function is useful for semantic developers who want to ask specific research questions or build live applications against the data, but it is unreliable and expensive to maintain.

Another possibility for hosting and accessing the data is via a dump, providing an easy, bulk download for all the generated triples. This would be useful for researchers or developers who want to build an application or run a research project using large amounts of the data. The bulk download is inexpensive and easy to host, but it limits access to developers who can not only write SPARQL but also set up their own triplestores.

Finally, the data can be presented via an API, so that developers can then use a common set of functions that request and receive responses via Hypertext Transfer Protocol (HTTP) such as GET and POST that provides Linked Data documents (typically, one document per entity), which are specific, curated subsets of the data. Creating an API is the easiest way for non-semantic developers to work with the data. But each document is only a subset of data, so if the research question or application you want to describe is different from the one that the data curator assumed, the result can be disappointing. The data presented for an entity can also leave out some data that might be connected to the entity in the graph—there is no guarantee that all the triples are available in that specific view.

What did AAC's browse demo set out to achieve?

The browse application aimed to provide a way for users—including museum staff, scholars, and eventually public art enthusiasts—to engage with the data from multiple institutions and see benefits that can arise only from the way the data is linked.

The AAC partners wanted an application that allows users to find connections between works that come to the surface by the links in the data. The goal is not to search for a specific item of interest. Rather, it is to move from item to item based on the compelling—and sometimes unexpected—relationships between them. Serendipity, discovery, and rich relationships were desired to illustrate the value of using a Linked Data approach.

The AAC produced a browse demo in part to help the partner institutions see what was possible given the unique links that could be established across the works of the partner institutions. Owing to limitations in the amounts of data that could be provided, the initial demonstration version dœs not offer all the richness we envision

with LOD, but it points the way to future capabilities.

What challenges dœs LOD present in designing a browse function?

The main challenges that arise when creating an application to browse LOD data are the same ones that ensue when producing other types of applications. Is the data you want to use available within the source systems, and dœs it have a sufficient quality to be used by computers? Sometimes, computer applications require specific formats that are different from the original, human-readable formats of the data—they need machine-readable dates, for example, or dimensions in which the numbers and units of measure are in separate fields. Another challenge is making sure, when there are multiple values for a data item (such as the title of an artwork), that the computer knows which title is the preferred title, to be used by default.

Data in source systems carry a lot of museum-specialized assumptions, rules, and norms that need to be checked by subject specialists in the museum. LOD itself, as a data format, requires an understanding of the technical syntax on the part of the people who work with it. The browse application helps museum staff see their data in a familiar, human-readable way.

The technical environment of the triplestore may not be sufficiently robust for use in real time. AAC's approach was to use the triplestore for complex, specialized analysis and at the same time export the RDF triples through a conversion process that produced JSON-LD. This allows for simpler processing that supports high-performance, scalable day-to-day use.

As new LOD becomes available, how do I learn about and connect to it?

This "discovery" challenge is a recognized issue within the cultural-heritage community, and standards-based solutions are emerging. In the meantime, the LOD for Libraries, Archives, and Museums (LODLAM) community is an excellent source as new data sets become available.

What do I tell my administration about LOD to convince them to support it?

- The internet is undergoing another revolution, with Linked Data formats increasingly being used as the underlying lingua franca for web applications.

- The internet is changing from an internet of documents to the internet of knowledge!
- Until now, when people have searched the internet, they have been presented with an array of hyperlinks to potentially relevant pages. The researcher must review all the links to determine which are relevant to the search at hand.
- A new way to publish information is called Linked Open Data (LOD), which precisely links and interconnects information so that searches are direct, accurate, and immediate. The links contain expressions about why two things are linked, and the meaning of the relationships between them, stated in ways that computers as well as humans can read.
- LOD uses a markup language called Resource Description Framework (RDF) that, when combined with an ontology, interconnects concepts (people, places, events, and things). The result is that a search connects to the exact concept being sought and avoids the "noise" that sometimes confuses online searching.
- LOD is making headway in the commercial, communications, and publishing worlds. Google, Facebook, the *New York Times*, US government agencies, Defense Advanced Research Projects Agency (DARPA), and many other institutions are implementing LOD. The European Union is building bridges across its libraries, archives, and museums using the digital platform Europeana, the Linked Open Data Initiative of the Europeana Foundation, The Netherlands.
- LOD is here to stay. Its flexible approach to creating meaningful links is the way of the future for data sharing.
- The benefits of LOD for museums are huge: LOD could connect data about one artist, for example, across all museums. Millions of people researching that artist would discover who has what art by that artist and where. LOD will increase museum visibility; LOD will reveal relationships among works of art because it will make connections among hundreds of related works; By linking concepts such as events, dates, people, and places across all domains, LOD will expose new information about a work of art. It will thus boost research that will lead to new discoveries; By its nature, LOD is a collaborative platform that museums can use to deepen audience engagement. Like Wikipedia, LOD provides an opportunity for the public to participate and help supply information (note that unlike Wikipedia, LOD provides no editing capability, but it is possible to build an application on top of Linked Data that would allow people to contribute suggested changes).

- In summary, LOD leverages the power of digitization.

What are some of the unique features that can be achieved with LOD over traditional research?

We are all looking for the concrete performance of LOD that illustrates its value. As more data becomes available as LOD, it will be possible to demonstrate the benefits of being able to search across several collections; have cross-domain access; create new opportunities that focus on its collaborative structure, and the like. One feature of LOD that has already intrigued art historians is the graphical display of LOD that can point to interconnections of time and place, for example, all the artists associated with a certain café in Paris during a specific period. The capability of LOD to produce graphical networks or connections can lead to new observations and conclusions.[2]

What LOD museum projects are under way in the cultural-heritage domain?

In addition to the LOD Initiative of the American Art Collaborative (AAC), some current projects engaged in LOD are the Arachne Project, Cologne; Arches, a collaboration between the Getty Conservation Institute, Los Angeles, and the World Monuments Fund, New York; Art Tracks, Carnegie Museum of Art, Pittsburgh, Pennsylvania; Canadian Heritage Information Network (CHIN) of the Department of Canadian Heritage; CLAROS (CLassical Art Research Online Research Services), a federation led by the University of Oxford; Finnish National Gallery, Helsinki; Germanische Nationalmuseum, Nuremberg; the Getty Provenance Index (GPI) and Getty Vocabulary Program, Getty Research Institute, Los Angeles; Pharos, the International Consortium of Photo Archives; ResearchSpace located in the British Museum, London; and Yale Center for British Art, Yale University, New Haven, Connecticut.

How do I join AAC?

The AAC has an ongoing interest in helping the broad museum community engage in LOD. Owing to fixed grant funding, however, we have had to limit the number of institutions involved thus far. The good news is that we are providing guidelines for the approaches, practices, and many of the tools that

museums can utilize to produce LOD. We plan to seek additional funding to expand application of LOD beyond the subject of American art and produce tools that will particularly help small museums to explore LOD. As additional museums produce LOD, we hope that those who are implementing it will

Notes

1 Wikipedia, s. v., "Ontology (information science)," last modified October 9, 2017, https://en.wikipedia.org/wiki/Ontology_(information_science).

2 Paul B. Jaskot, "Digital Art History: Old Problems, New Debates, and Critical Potentials" (keynote address for symposium Art History in Digital Dimensions, University of Maryland, College Park, and Washington, DC, October 19–21, 2016).

contact the AAC (via Eleanorfink@earthlink.net) to discuss opportunities to link or interconnect data and further demonstrate the value of LOD.

Appendix 1. AAC Partners, Advisers, and Consultants

AAC PARTNERS

Amon Carter Museum of American Art, Fort Worth, Texas
Katherine Moloney, Teaching Resources Coordinator

Archives of American Art, Smithsonian Institution, Washington, DC
Karen Weiss, Supervisory Information Resources Specialist
Michelle Herman, Digital Experience Manager
Toby Reiter, Information Technology Specialist

Autry Museum of the American West, Los Angeles, California
Rebecca Menendez, Director, Information Services and Technology

Colby College Museum of Art, Waterville, Maine
Charles Butcosk, Digital Projects Developer

Crystal Bridges Museum of American Art, Bentonville, Arkansas
Shane Richey, Digital Media Manager

Dallas Museum of Art, Texas
Shyam Oberoi, Director of Technology and Digital Media
Brian MacElhose, Collections Database Administrator

The Thomas Gilcrease Institute of American History and Art, Tulsa, Oklahoma
Diana Folsom, Head of Collection Digitization

Indianapolis Museum of Art at Newfields, Indianapolis, Indiana
Stuart Alter, Director of Technology
Heather Floyd, Software Developer
Samantha Norling, Digital Collections Manager

National Museum of Wildlife Art, Jackson, Wyoming
Emily Winters, Registrar

National Portrait Gallery, Smithsonian Institution, Washington, DC
Sue Garton, Data Administrator

Princeton University Art Museum, New Jersey
Cathryn Goodwin, Manager of Collections Information

Smithsonian American Art Museum, Washington, DC
Rachel Allen, Deputy Director
Sara Snyder, Chief, Media & Technology Office
Richard Brassell, Software Developer

Walters Museum of Art, Baltimore, Maryland
Will Hays, Assistant Registrar, Data and Images

Yale Center for British Art, New Haven, Connecticut
Emmanuelle Delmas-Glass, Collections Data Manager

ADVISORY COUNCIL

Vladimir Alexiev, Lead, Data and Ontology Management Group, Ontotext, Sofia, Bulgaria

Craig Knoblock, Director of Data Integration, Information Sciences Institute, University of Southern California, Los Angeles

Martin Dœrr, Research Director at the Information Systems Laboratory and Head of the Centre for Cultural Informatics of the Institute of Computer Science, FORTH, Crete, Greece

Robert Sanderson, Senior Semantic Architect, J. Paul Getty Trust, Los Angeles

Thorny Staples, Director (retired) of the Office of Research Information Services, Office of the Chief Information Officer, Smithsonian Institution, Washington, DC

CONSULTANTS

Vladimir Alexiev, Lead, Data and Ontology Management Group, Ontotext, Sofia, Bulgaria

Kate Blanch, Data Architect, Design for Context, Washington, DC

Duane Degler, Principal Consultant, Design for Context, Washington, DC

Eleanor Fink, Founder and Manager, AAC

David Newbury, Applications Developer Architect, J. Paul Getty Trust, Los Angeles

Pedro Szekely, Project Leader at the Information Sciences Institute, University of Southern California, Los Angeles

Appendix 2. AAC Participants' Initial Comments and Questions ABOUT LOD

PARTICIPANTS EXPRESSED WHY ENGAGING IN LOD HAS MANY BENEFITS:

Everyone wants more meaningful content and LOD promises to provide more precise searching.

LOD provides an opportunity to find new ways to share our information and keep audiences engaged.

LOD is an opportunity to augment our collection information by connecting to related art information in other museums and institutions.

LOD's precision and interconnecting of art information can improve access.

Being able to search across several collections is richer than searching just one.

The interconnections will make our collections more discoverable.

Because LOD is dynamic, making many connections possible, it will allow us to tell fuller stories about our objects.

The idea of complementarity or interconnecting data from museums, libraries, and archives is enriching.

LOD will allow users to explore cross-domain opportunities (e.g., linking information about something depicted in art objects to related items in a science or natural history museum, etc.).

Since LOD functions in concept like Wikipedia, in which someone can post or add information, it gives museums interested a powerful collaborative platform.

Output in a JSON-LD format is much easier for app development and therefore can make our data more useable to developers.

LOD will help us better adapt to the changing web.

AAC PARTICIPANTS SUPPORTED THE IDEA OF WORKING ON A COLLABORATIVE PROJECT:

AAC is an opportunity to build on diverse skills and tap into each other's expertise.

Through ACC, we will receive more comprehensive education.

AAC makes it easier to engage and convince management (several stated they would not have been able to investigate LOD had it not been for AAC).

AAC is cost effective owing to grant funding and leverage (the grant covered conversion of the data, workshops, and linking to Getty ULAN).

AAC provides for testing a critical mass to better learn about the value of LOD.

AAC is an opportunity to create good practices for the broad museum community.

PARTICIPANTS ALSO IDENTIFIED CHALLENGES AND UNCERTAINTIES, MANY OF WHICH THIS GUIDE WILL ATTEMPT TO ANSWER:

Dœs my institution need a license and what are the options?

I'm concerned about mastering the ontology, which is the CIDOC CRM.

What are the implications of federated or hybrid models versus aggregation models?

What is the learning curve and skill set required?

How accurate are the tools we will be using?

How much work will I have to undertake?

I'm concerned about working with high-end technology experts who are not art specialists.

As new data sets come online, how do I connect to them?

How do I create a URI?

How do I know what to connect to?

How do I make sure I'm connecting to the right entity (person, place, etc.)?

How do I maintain the data?

How do I obtain metrics on the impact of converting my data to LOD and see how they are using our data?

I'm interested in exploring what openness means.

How do I get DBpedia to link back to my data? (DBpedia aims to extract structured content from the information created in Wikipedia and make it available on the World Wide Web.)

How do I encourage people to link to my LOD?

How do I reinforce the authority of my LOD?

Appendix 3. Road Map

October 6, 2015

INTRODUCTION

The American Art Collaborative (AAC), comprising fourteen museums, has spent the past nine months engaged in learning about Linked Open Data (LOD) and planning how to move forward to implement LOD in the museum community. AAC representatives met in person on February 4 and 5, 2015, at the Smithsonian American Art Museum (SAAM). A series of online educational briefings preceded the in-person meeting. Topics covered by the briefings included an introduction to LOD, the CIDOC CRM ontology, and a review of projects such as DPLA, Research Space, Europeana, International Image Interoperability Framework, and Yale Center for British Art.

The Washington meeting was both stimulating and productive. Members, advisors, and consultants were active and fully engaged in working groups and brain storming sessions. There was unanimous agreement to continue as a collaborative and to articulate a road map for next steps. Representatives were in agreement that a collaborative provides a supportive environment to leverage knowledge and skills across museums, as opposed to each institution attempting to produce LOD in isolation. The representatives were keen to work together to create a rich and useful mass of linked data drawn from their collection metadata repositories, identify potential programmatic applications for a broad range of linked data types and sources, and establish a sustainable network of LOD for increasing the understanding and appreciation of art. This consensus is reflected in the Mission Statement appended to this roadmap document.

AAC members are ready to move forward over an eighteen-month period to engage in a robust demonstration project that will consist of publishing a rich and diverse critical mass of data drawn from the thirteen partner institutions; reconcile the data across institutions as well as with key vocabulary resources such as the Getty Union List of Artist Names; make their data available publicly on the web so that it can be discovered and harvested; collaborate with open source resources like the International Image Interoperability Framework (IIIF); and develop demonstration applications that illustrate the value of LOD within a network of connected resources that can provide richer, more diverse information than any single institution could hope to provide on its own.

AAC expects to explore a federated approach for providing access to the critical mass in which each institution (or sets of smaller institutions) can maintain responsibility for managing and refreshing its data, yet work closely as a collaborative network in terms of best practices and tools. AAC believes that a non-aggregation approach is more sustainable and less costly in the long term and offers more buy in and responsibility for each participating museum. AAC also believes a federated approach is more in keeping with the LOD concept envisioned by its creator Tim Berners-Lee in which institutions and individuals publish and share data that resides in a linked open data cloud. Yet AAC equally believes that in order to avoid publishing LOD that cannot be reconciled, it is paramount to explore the need for consistency in terms of URI generation, best practices in applying the CIDOC CRM, and linking to such international terminology resources as the Getty vocabularies. One of the key challenges AAC wants to address in the road map is how its data can be aligned to maximize access for research purposes and to demonstrate the values of LOD.

PROJECT WORK PLAN

I. Administrative Preparation

Letter of Agreement: Before moving forward, each partner institution will recommit to the next phase of work by signing a Letter of Agreement. The Agreement will detail the museum's responsibilities for preparing and providing data, proofing its mapped data, and participating in review meetings, discussions, and hands on workshops. Reciprocally, the Agreement will outline the AAC's commitments, contingent on funding, to assist the museums in converting their data to LOD, to reconcile data, map to the CIDOC CRM, and provide logistical support for training, workshops and meetings.

Hiring of Contract Consultants: The AAC grant administrative museum will hire consultants on contract to support the program, to include: project manager Eleanor Fink; data modeling coordinator Emmanuelle Delmas-Glass; CIDOC CRM expert Stephen Stead;[1] University of Southern California, Information Sciences Institute (ISI), with principal support from Pedro Szekely; Duane Degler and Neal Johnson, Design for Context; and a technical writer (to be determined).

II. Mapping and Conversion to LOD

Data Selection: Each institution will identify the types and quantity of data it is able to commit to the AAC linked data resource. In addition to collection data, some museums have expressed interest in contributing curatorial notes, exhibition records, archival materials, videos, and provenance records. The Archives of American Art, an AAC member, may further enrich the critical mass by contributing research materials, like artist letters, diaries, notebooks, photographs, oral histories, and gallery records.

From a longer-term perspective members expressed interest in eventually expanding beyond a focus on American art to cover their entire art collections and also include a selection of local materials from their communities that would demonstrate the value of LOD within their specific domain of influence. These materials could be drawn from archives, historical societies, or natural history and science museums and would serve to demonstrate cross domain research potential of LOD.

Preparation of Data for CIDOC CRM: The CIDOC CRM, also recognized as ISO 21127:2006, will be used as the principal ontology for mapping AAC data. The CIDOC CRM is robust, containing 82 classes and 263 properties, including classes to represent a wide variety of events, concepts, and physical properties. Furthermore, it includes many properties to represent relationships among entities. The CRM is already being used by the British Museum, Yale Center for British Art, and SAAM, so there is an established group of practitioners.

Data modeling coordinator Emmanuelle Delmas-Glass will help AAC members prepare their data and identify data points for mapping to the CIDOC CRM. This individual assistance will help members grasp the potential of the CIDOC CRM and how it is applied in relation to their own data. This step will minimize misconceptions that can occur due to the size and complexity of the ontology and make it easier for partners to proof their data once it has been mapped.

Data Conversion: Data will be converted by ISI using their data integration tool, Karma. Karma is semi-automated and self-learning. It has successfully mapped hundreds of thousands of records from the Intelligence Advanced Research Projects Activity (IARPA). It also converted forty-four thousand art records from SAAM.

ISI will accept the data in a variety of formats, including but not limited to TMS, XML, or Excel. A number of AAC museums have expressed interest in posting their raw data to GitHub, a web-based repository hosting service. GitHub will be a work platform that will allow ISI to collect the data for conversion to LOD. AAC's smaller museums, with insufficient technical support on staff, may opt to use other formats familiar to them for sending their data directly to ISI where it will be prepared for conversion to LOD. Once the data is obtained by ISI, it will be loaded into Karma and mapped to the CIDOC CRM. The mapping will be reviewed by a CRM expert. The conversion of AAC data will be handled in batches, as museums are ready. ISI will return both the mapping of their data as R2RML as well as the data itself in RDF and JSON-LD to the museums. Museums may elect to post the returned mapping and data to GitHub.

Mapping Review: Stephen Stead from Paveprime Ltd. will be contracted to work with ISI and AAC in reviewing the mapping produced by Karma and verify the CRM was applied accurately. During the grant period Stead will also provide AAC participants with a best practices hands-on workshop that will be drawn from the lessons learned during the modeling and mapping processes.

III. Reconciliation, Linking and Publication

Reconciliation: A major goal for AAC is to demonstrate the reconciliation of data across large and small institutions as well as archives and museums through primary entities such as creators, places, and subjects.

The Getty's Union List of Artists' Names, Thesaurus of Geographic Names, and Art and Architecture Thesaurus are authoritative and recognized vocabulary tools for art history. By linking to these global vocabulary standards as well as the Getty Cultural Name Authority when ready, AAC members can reconcile against a common standard. This will help ensure consistency across AAC data and enhance the development of research and educational applications. In addition, additional research can identify standardized vocabulary resources from other domains that could in future extend the usefulness of the AAC members' LOD.

To ensure that AAC is applying best practices in linking to the Getty vocabularies and mapping to the CRM, a small reconciliation workshop will be held at the beginning of the implementation period to clarify how best to achieve harmonization of the concept based vocabularies with the event model of CRM. The meeting to be facilitated by Design for Context will involve a CRM specialist, data modeling coordinator, Getty Vocabulary technical staff, ISI technical staff, and one or two representatives from AAC partner institutions.

Link Curation: ISI will develop its Karma link curation tool to facilitate linking between member museums and from member data to the Getty vocabularies and other LOD resources. Linking between member museums will be a critical part of the project Making links to recognized hub sites within and outside of the art domain will enrich the context and authority for the materials and deepen research potential. In addition to the Getty resources, the link curation tool will also be applied to common hub sites such as DBPedia and the *New York Times*.

IIIF: AAC is interested in IIIF because it would provide the capability of comparing and contrasting images for all of AAC and additional IIIF clients of research interest. This particular application combined with members' interest in browsing AAC's LOD by such entities as artists, events, and places would provide a means of demonstrating the value of LOD. But IIIF dœs not use the CIDOC CRM as its ontology. Therefore, ISI, with the help of Stanford University, will update Karma to include a tool for mapping to IIIF. ISI believes Karma can be easily adapted and would be willing to work with IIIF and Stanford University to build this added feature. IIIF leaders concur that adding this feature to Karma would be a boost for the broader cultural-heritage community interested in IIIF combined with use of the CRM. If AAC data is consistently mapped to the CRM and to IIIF, the remaining challenge will be for each AAC museum to implement the API and viewer for IIIF.

Training and Workshops: The Collaborative will provide expert training, technical support, and guidance to help the partner museums learn to apply the CIDOC CRM, use the curation linking tool to be designed by ISI, update and manage their own data, and for those interested install a IIIF API and reader. Regular virtual meetings among participating institutions will be held throughout this process to discuss issues, work out problems, and ensure all partners are meeting objectives.

Hosting: As a first step, members will make their data available publicly on the web so that it can be discovered and harvested for use cases. ISI has agreed to host the data for a period of one year in a triple-store database for AAC members who do not have the capacity to host their own data. It is hoped that over time AAC members will have the capability to launch their own SPARQL endpoints or, as has been suggested, agree to create small hub sites that provide a shared SPARQL endpoint. Should ISI need to host the data longer than a year, there would be a reasonably priced service charge. Although the data will reside initially on a server at ISI, the museums can put in a DNS redirect from their web site to the ISI server. Visitors will see the museum URLs and never know that the data is hosted elsewhere. AAC will explore the feasibility of the LOD residing on the museums-sponsored GitHub repositories.

Data Review, Meetings, and Testing: At least three in-person meetings are planned over the course of the eighteen months: a kickoff meeting, a midterm stock-taking meeting, and a meeting to discuss next steps and strategies toward the end of the period.

Data review and quality assurance meetings in which the partners can discuss issues or raise questions about the mapping of their data are a critical part of the road map and will be held as often as needed throughout the grant period via GoToMeeting. No data will be published until they have been vetted by the partners.

Data review will include the CIDOC mappings as processed by ISI via Karma, particularly the indicated event relationships. Although Stead will verify the application of the CRM and relationships expressed for accuracy, each partner institution will also need to review the results. Review and quality assurance will also include the reconciled data to make sure the entities align correctly. Last, but not least, it will also involve editorial review to correct any errors and inconsistencies.

IV. Applications and Demonstrations

A core AAC discussion during the in-person meeting at SAAM was what kind of applications to develop if resources become available. AAC will build on existing open source projects like IIIF, which will be valuable in providing a means to compare and contrast images of related works of art. But AAC believes additional *f*use cases are also needed that demonstrate the power of LOD. An initial list of use cases that identify applications is in the

attached chart. AAC has vetted the list and has prioritized applications based on a set of criteria that included: what kind of research and educational applications clearly demonstrate the value of LOD; what content can AAC partners provide that can successfully address scholarly research and educational and public program use cases; which applications meet both intra- and extra-institutional interests; which applications support short-term publishing, review and quality goals; and which applications are feasible within the context of a demonstration project.

At the top of the list of applications is the ability to browse the sum of data as if it were a "virtual database." AAC wishes to be able to use the "virtual database" for purposes of demonstrating the ease of finding all examples by such entities as person, place, and event and will develop a demonstration proof of concept. Another high priority among the use cases identified by AAC partners involves a flexible way to identify and navigate into the context and deeper scholarly content associated with an object. Scoping the requirements for this application will run in parallel with data mapping and conversion, but development of the application will require funding not yet identified.

Design for Context will shepherd the discussions, scoping, design, and testing of demonstration applications to be developed.

There will be meetings with AAC partners throughout the process, to discuss and provide feedback on design of demonstration applications. It will be an iterative, user-centered process. Input from AAC institutional leads and end users from the partner institutions will be critical.

V. Guidelines and Best Practices

Publication: AAC is committed to sharing its experiences with lessons learned from the member museums. Diane Zorich has agreed to co-author with Eleanor Fink a publication on best practices with case studies.

Conference Presentations: AAC consultants and members will present use cases and achievements at such conferences as Museums and the Web, Museum Computer Network, and American Alliance of Museums. Ideally, as part of the AAC's outreach to other cultural institutions, presentations will also be proposed to major library and archive conferences.

VI. Contract Consultants and Advisory Council

Contract Consultants:

Project manager Eleanor Fink, who founded the AAC, will continue in her management role, serving as the point of contact with AAC members, consultants, and advisors to advance the project and resolve issues as they come up. She will plan and execute the meetings,

track project goals and deliverables, and communicate with practitioners in the field. Fink served for thirteen years at Smithsonian and then at the J. Paul Getty Trust, initially as founder of the Getty Vocabulary Program and its program officer for scholarly resources, and then as director of the Getty Information Institute (GII). She positioned GII around the concept of universal access to art information and promoted collaboration across institutions. The National Initiative for Networked Information (NINCH), Getty Vocabularies, Categories for the Description of Works of Art, and Object ID are some of the products of her leadership.

Data modeling coordinator Emmanuelle Delmas-Glass will help AAC members prepare their data and identify data points for mapping to the CIDOC CRM. Delmas-Glass is the Collections Data Manager in the Collections Information & Access Department at Yale Center for British Art, which has been working with the CIDOC CRM for several years. CIDOC CRM expert Stephen Stead will work with the Information Sciences Institute (ISI) in reviewing the applications of the CRM to data mapping and to provide a hands-on workshop and data expertise as needed throughout the eighteen-month grant period. Stead from Paveprime Ltd. is a highly qualified expert who helped build and develop the CRM. He conducts CRM workshops in Latin America, Europe, and the US.

University of Southern California, Information Sciences Institute (ISI), with principal support from Pedro Szekely, will apply their Karma data integration tool to convert AAC records to LOD, develop a curation linking tool, and provide hands on workshops on how to map, refresh, and maintain data. ISI is a world leader in research and development of cyber security, advanced information processing, and computer and communications technologies. A unit of the University of Southern California's Viterbi School of Engineering, ISI is one of the nation's largest, most successful university-affiliated computer research institutes. ISI has developed an open source data integration tool, Karma, for converting data to LOD.

Duane Degler and Neal Johnson, Design for Context (DfC), will serve as facilitators to coordinate meetings, develop application specifications with the AAC members, and advise on best practices for using LOD. DfC provided invaluable support during the Mellon planning grant, helping to plan and facilitate the in-person meeting. DfC specializes in articulating visual and interaction design requirements for web applications, software, and websites with specific interest in leveraging linked data and semantic technologies.

A technical writer, to be determined, will work with the AAC to document guidelines, best practices, and case studies to be shared with the broader museum community.

Advisors: The advisory committee established during the planning phase will be invited to continue as project advisors during the implementation phase. The advisors represent a balance of expertise in LOD, the Semantic Web, ontologies, and use of technology in museum and academic research settings.

Robert Sanderson: Information Scientist at Stanford University Libraries. His research focuses on digital libraries, archives, and museums and their interaction via LOD and the web. Sanderson brings relevant experience from his participation in the International Image Interoperability Framework and general LOD experience with cultural-heritage institutions.

Thorny Staples: Director of the Office of Research Information Services at the Smithsonian Institution, Office of the Chief Information Officer (CIO). His work has touched almost every area of digital projects, from technical programming to software development, with a focus on research systems in the humanities.

Craig Knoblock: Director of Data Integration, Information Sciences Institute (ISI), USC. Knoblock is an expert in the area of AI and Information Integration. He has worked on a wide range of topics within this area, including information extraction, wrapper learning, source modeling, record linkage, mashup construction, and data integration. Craig and his team specialize in research and development of tools that streamline creation of linked data from existing data repositories (e.g., Karma).

Tim Finin: Professor of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. Finin has more than thirty years of experience in applications of Artificial Intelligence to problems in information systems and language understanding. His current research is focused on the Semantic Web, mobile computing, analyzing and extracting information from text and online social media, and enhancing security and privacy in information systems. He was a member of the W3C Web Ontology Working Group that standardized the OWL Semantic Web language.

Martin Dœrr: Research Director at the Information Systems Laboratory and head of the Centre for Cultural Informatics of the Institute of Computer Science, FORTH. He has been leading the development of systems for knowledge representation and terminology, metadata, and content management. His long-standing interdisciplinary work and collaboration with the International Council of Museums on modeling cultural-historical information has resulted in an ISO Standard, ISO 21127:2006, a core ontology for the purpose of schema integration across institutions.

VII: Expected Project Results

By the end of the project, at least one hundred thousand records pertaining to American art from fourteen museums will be made available as LOD in a demonstration of how to reconcile LOD across multiple institutions. The data will be available through IIIF and potentially other aggregators (e.g., DPLA). The data will also be openly available through each AAC parter institutions' website. The data will be linked to authoritative linked data resources such as the Getty ULAN, TGN, and AAT (and conceivably CONA). There

will be demonstration use cases that underscore the value of LOD. AAC partners will have been trained to maintain and update their own data using the Karma tool, which is open source.

Museums and future collaborators will benefit from the Initiative's leadership role through:

- practical and detailed knowledge of how LOD is created;
- developments in reconciliation of data and identification of best practices;
- a set of open source tools that will facilitate reconciliation;
- establishment of a sustainable and scalable collection LOD;
- technical requirements for publishing and maintaining LOD;
- experience with a federated model of access;
- demonstrations of practical applications of LOD in the context of art historical research and museum programs that include the ability to compare and contrast images;
- open access policies for cultural-heritage information management strategies.

AAC members will be able to promote and facilitate the process of conversion to LOD for other museums. The Collaborative will share what it has learned, including best practices, case studies, technical resources, and means to help other museums avoid costly mistakes or incompatible data representations. The project will strive to ensure that the linked data produced by the partners is accessible, sustainable, and scalable as a collection of resources. It is hoped that the experience will be translatable to, and inform the success of, institutions attempting to replicate or join the AAC effort. AAC partner museums and other institutions will have a collection of data against which they can add, experiment, and develop applications to serve curators, researchers, students, educators, and the public.

Note

1 Vladimir Alexiev, Lead, Data and Ontology Management Group, Ontotext Corp, took over the review of the mapped data in February 2017, replacing Stephen Stead. Alexiev also became an AAC adviser, replacing Tim Finin.