Creating a Digital Cultural Heritage Ecosystem: Global Documentation Standards and Linked Open Data

Eleanor E. Fink

Introduction

The cultural heritage and scholarly community would agree that works of art and artifacts are signposts for understanding the story of humankind: where civilization originated, what civilization achieved, and what it can teach us. They would also agree that while there is no substitute for viewing a work of art or artifact in the flesh, documentation whether fielded data description or digital image capture is essential for collection management and research.

In museums, documentation is a central resource for stewardship or collection management. It is a means for capturing descriptive information about a cultural object (material, size, name, dimensions, etc.) It also can capture contextual information and/or stories that could otherwise be lost within an institution: how the work was acquired, relationships across collections, discoveries upon examining the work, etc. In research, it functions as a means of finding information and items of interest and potentially pointing to associations across domains.

A major challenge for museums is that senior management does not always recognize that documentation is a long-term infrastructure commitment[1]. Rather than financially commit to a digital strategic plan that includes updating technology, improving documentation practices, adhering to international standards, hiring additional staff with technical expertise and improving staff skills, museum leaders tend to prefer budgets for short-term, "shiny," technology projects that draw audiences and please trustees. These executives are missing the big picture – namely that the data points used within an institution when developed according to international standards and good practices can allow data to be shared and enable a museum's objects and/or artifacts to be connected with related data about them as part a larger network of museum and non-museum data throughout the world. In contrast, to begin research today, you must know where to look or which institution has works by a given artist or school. Connecting data would allow us more easily to grasp and make associations about the lifework of an artist, the scope of a particular style, or the details of a historic period. It could drive new kinds of entry points for audiences, new types of exhibition storytelling, and new possibilities for creative inquiry:
https://drive.google.com/file/d/1IuavP0M7HThcYydGAEpzxGKebmGSyKwU/view?usp=sharing

[1] Preference by museum directors for "shiny" technology projects versus investing in long-term infrastructure projects such as global documentation standards appears more widespread in the US than it is in Europe. One reason perhaps due to the fact that there is no ministry of culture in the US for policy.

Documentation Standards

Standards are essential for recording information consistently and fundamental for retrieving information efficiently, particularly when using computers. They enable research, promote data sharing, improve content management, and reduce redundant efforts. When asked about the role of standards in art documentation, I often reply: Which standards? Whose standards? Local standards particular to your institution or are we talking about national and international standards? If we are interested in sharing or connecting data, institutions need to implement global standards that have been vetted by the cultural heritage community. I know a little about art documentation standards and the effort required to create and establish them. Together with my staff, when I was founder of the Getty vocabulary program and later director of the Getty Information Institute, we created many of the essential vocabulary and metadata standards used today around the globe http://www.dlib.org/dlib/march99/fink/03fink.html.

Educated as an art historian, my first position out of graduate school was to establish a photo archive and slide library for the Smithsonian's National Collection of Fine Arts, NCFA (now known as the Smithsonian American Art Museum or SAAM.) Joshua Taylor, the director of the museum at the time, viewed photo archives as vital research resources that provide art historians the opportunity to compare and contrast works of art from museums and private collections around the globe. Taylor additionally appreciated being able to access related research materials and in particular the comments made by scholars about the works they were studying (the Berenson Archive being an example). As early as the 1970's, Taylor also did not hesitant to engage technology for research purposes. Degrees in American art were not as prevalent across schools of higher education as more traditional subjects in art history such as Dutch, Italian, Medieval, Renaissance, Baroque, or even "modern" art. To help the academic community as well as the public better understand American art, the museum was home to a scholars and research center that supported pre and post-doctoral candidates interested in studying and writing about the museum's collection and American art in general. Under Taylor's leadership and efforts to learn more about American art, the museum initiated a nationwide research project The Bicentennial Inventory of American Painting Executed Before 1914 - a multi-year project that collected data across the United States about American paintings executed before 1914. The data was collected from public and private sources including schools, museums, and private collections. Volunteers were engaged to complete forms and subsequently send them to the NCFA/SAAM where project staff typed the data via optical character recognition that was then scanned and stored on a mainframe computer. Use of a computer, albeit relatively new, was a logical choice given the volume of data and the possibility of producing a number of indices by origin, state, owner, date, title, artist, etc.

The fact that the Museum supported a scholars and research center and engaged in computerized research projects that collected information about art beyond its own collection and walls profoundly influenced my career and views on research requirements and how to collect and manage information (http://mcn.edu/mcn50-voices-eleanor-e-fink-dara-lohnes-davies/). It was clear that in order to

serve the fellows, images needed to be accessible from a variety of data points. More prevalent than date, period, or medium, was subject depicted. For example, the fellows were interested in images of woman with white parasols, children playing games, depictions of death, etc. Access had to be diverse as opposed to one-dimensional browsing through a card catalogue. Thus, early on, I wanted to understand how best to make use of data for research purposes.

Following The Bicentennial Inventory of American Painting Executed Before 1914, the museum and I initiated additional computerized research projects. There were five discrete databases numbering over 500,000 records. To help the scholars search the combined contents of the projects, I worked with the Smithsonian Office of Computer Services to create a finding aid called the comprehensive artist index that pointed to works by an artist in each of research projects. The index indicated lack of consistency in how artists' names were recorded across the projects. Computer programs require consistency in cataloguing terminology. Without consistency, the inconsistent data will not appear and research results will be incomplete.

At the time, there were no national or international standards in art history for documenting and recording information. In fact, art historians disliked the idea of "standardizing" the words or terms used to describe and catalogue a work of art. In contrast, Librarians worked collaboratively as a community to create national and international standards to provide consistency in terminology and cataloguing practices. I admired how librarians worked together and thought if the art museum and art history community collaborated, perhaps there could be networks of art information. If networks existed, instead of depending on written forms and sending out volunteers across the United States to record information as NCFA/SAAM had done, it would be possible to obtain a significant amount of data by accessing the networks. I reached out to the library of congress and discussed my needs with the staff in charge of NACO, the Library of Congress Authority File – a nationwide collaboration between libraries and the Library on Congress. Participants provided with training contribute authority records for agents, places, works, and expressions. I raised the question why not include authority records for artists and works of art? The response was "who would serve as the authority" The lack of standards for describing works of art that makes computer access incomplete and less accurate spurred me to create standards and work collaboratively with national associations. Therefore, when visitors from the J. Paul Getty Trust came to see my work, the exchange led to an offer from the Getty Art History Information Program to become the founding director of the Getty vocabulary program. Similar to my work at NCFA/SAAM, the Art History Information Program also had several discrete scholarly research projects that were created using different computer platforms. There was little consistency in the standards used for recording information. Under my leadership at the Getty, my staff and I initiated and launched the Union List of Artist Names and the Getty Thesaurus of Geographic Names.

Later I initiated and launched Categories for the Description of Work of Art (https://www.getty.edu/research/tools/vocabularies/and Object ID for Protection of Cultural Property that is currently managed by the International Council of Museums (ICOM) https://icom.museum/en/activities/standards-guidelines/objectid/.

All of these projects were multiyear consensus building efforts that involved establishing partnerships and working with the global cultural heritage community. For example, Object ID is now widely implemented by organizations as a standard that provides essential data for uniquely identifying cultural property (Interpol, Carabinieri, several police databases, Art Loss Register, US Immigration and Customs, US military, etc.) UNESCO, Council of Europe and ICOM officially endorse it[2]. Convincing international organizations to work together to establish Object ID was a challenge. Success was based on building trust using the following approach implemented over a five-year period:

Establish partnerships

Over a three-year period, I discussed the need for Object ID with key agencies that I invited to form an alliance: ICOM, UNESCO, Council of Europe, USAID, and The Organization for Security and Co-operation in Europe (OSCE).

Engage the broader community

Under the logos of the Alliance participants, surveys were distributed to member countries of these organizations to answer questions about what data is needed for uniquely identifying cultural property.

Analyze and Point to Common Practices

The data from the surveys was analyzed to indicate where there was agreement and a report was prepared. From this step, a core data standard emerged.

Discuss and Communicate Results

Meetings and an International Conference were held with key constituents and associations that had documentation committees.

[2] My visit to the FBI in Washington, DC, inspired me to create the Object ID project. A report came in about a painting stolen in Amsterdam. It was shocking to learn that it would take weeks before the data about the stolen artwork could be communicated internationally to law enforcement. The description of the work of art came from curatorial files and comprised 10 pages. It certainly would not help a customs official quickly determine if something found in a suitcase or the back of a vehicle was a stolen item. The records were in Dutch and would need to be translated. Police databases were not synchronized and the standards varied so it was not possible to quickly share and exchange data. The concept behind Object ID was to keep the metadata simple and provide a set of guidelines to help better communicate items lost or stolen.

By the 1990's the international community was adopting the standards my staff and I created. With the raise of the internet and my new position as Director of the Getty Art History Information Program (AHIP - later renamed the Getty Information Institute or GII), I structured our programs around the concept of universal access to images and art information. The vision centered on partnerships with cultural heritage institutions, harmonization of standards and education and training in documentation. From my perspective, the tools we launched at the Getty were helping the cultural domain collaborate and begin sharing data. Thus promoting universal access

The concept of the Internet as a web of webs held new possibilities. Rather than just connect data points within a single institution, why not identify a means to prepare data so that it can be shared as part of a larger network or cultural heritage ecosystem. In other words, why not be able to search all the photo archives and art data across institutions from a single access point. I termed the concept the virtual database and in the 1990's after becoming Director of AHIP, my staff and I produced a video that explained the concept (https://drive.google.com/file/d/1IuavP0M7HThcYydGAEpzxGKebmGSyKwU/view?usp=sharing).

For a short duration before our institute was shuttered following a change of leadership at the Getty, we developed demonstrations of the virtual database concept through projects such as L.A. Culture Net - a gateway to arts in Los Angeles and American Strategy - a gateway to art collections in Federal Institutions located in Washington, DC.

All of our energy and vision came to an end in and around 1998 when the Getty's new president and CEO eliminated the Education and Information Institutes. Within two years, the directors of the remaining institutes were gone. Only a few of my staff – namely the vocabulary program and some of the scholarly research databases remained and were transferred to the Getty Research Institute. In the years following, one or two Getty successors attempted to shadow GII's vision in respect to the concept of the virtual database and finding digital pathways across institutions.

Today the Getty Vocabularies are known globally and are available in many languages https://www.getty.edu/research/tools/vocabularies/. Object ID has become a defacto standard for art theft databases. Categories for the Description of Works of Art is well known and has helped many projects determine which information would be useful to record.

There remains, however, an urgent 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 ultimately diminishes the ability to connect cultural-heritage information to the multitude of documents around the globe that tell the story of humankind.

Linked Open Data - a Key-Pin for Connecting Data and Building Knowledge Graphs If we value our cultural heritage as signposts for understanding the story of humankind: where civilization originated, what civilization achieved, and what it can teach us, why would we lock documents about our civilization in data silos.

Fortunately, the concept of enabling data to be interconnected and shared as part of a larger network of museum and non-museum data throughout the world is gaining momentum through Linked Open Data (LOD). Examples, include Europeana: the EU's digital platform for cultural heritage consisting of 3,000 institutions across Europe; ResearchSpace: Collaboration with several museums in Europe – privately funded by Andrew W. Mellon Foundation; Canadian Heritage Information Network: based on a pilot project involving eight museums across Canada that inspired CHIN to expand LOD as the way forward for museums and galleries to share their data with each other and the world; Enslaved: People of the Historic Slave Trade, The Center for Digital Humanities & Social Sciences at Michigan State University; and the American Art Collaborative: 14 US museums and one archive that has published over 230,000 LOD records publicly available on the World Wide Web.

LOD is a method for publishing structured data on the World Wide Web so that it can be interlinked and therefore useful in web implementations. It relies on global standards developed by the W3C. To convert data to LOD, pieces of information have to be tagged much like HTML for publishing on the web. Unlike web hyperlinks that broadly connect thousands of bits of information that seem similar based on matching words, LOD produces highly precise 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.

The W3C has developed rules for structuring and tagging the data. A data format called the Resource Description Framework (RDF) is the basis. RDF is a method for describing data by

defining relationships between data objects. There is a good explanation of RDF at https://ontotext.com/knowledgehub/fundamentals/what-is-rdf/. Also, check the extensive FAQ, a section of the AAC Guide: Overview and Recommendations for Good Practices on https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAccessible.pd f?sequence=1&isAllowed=y

When formatting RDF, knowledge needs to be broken down into discrete pieces, with some rules about the semantics, or meaning, of those pieces. Information is expressed as a list of statements in the form Subject/Predicate/Object, known as triples. e.g., “Rebecca paints portraits”. An ontology or knowledge representation appropriate to the cultural heritage field must be selected to play the key role of defining the meaning of the terms used in the statements. 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.

RDF, ontologies, and URIs enable the data to be read by computers and interlinked. They are the basis for building knowledge graphs. They enable very precise relationships across LOD data that can lead to improved search results and opportunity for new discoveries.

Once triples (subject, predicate, and object) 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.

Case Study: The American Art Collaborative Linked Open Data Initiative (AAC)
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 the subject of the visual arts in America.[3] After a planning grant from the Andrew W. Mellon Foundation, AAC engaged in a two-year project funded by a national leadership grant from the Institute of Museum and Library Services and a second grant from the Andrew W. Mellon Foundation. At the end of the grant cycles AAC produced over 230,000 LOD records drawn from the fourteen institutions. A guide with a detailed description of the AAC project, the tools created, lessons learned, good practices recommendations, and an extensive FAQ is available at https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAccessible.pdf?sequence=1&isAllowed=y

Several perspectives and decisions shaped AAC's approach, planning, and results: (1)Partners were motivated by the concept of linking their data across museums, archives, libraries and to resources such as the Getty Vocabularies, and Wiki Data. They were interested in populating the linked open data cloud with art documentation. One of the principal benefits of the linking is the greater visibility it provides to museums especially those museums located in rural areas. For example, a researcher interested in the works of the American sculptor, Paul Manship might know that the Smithsonian American Art Museum has several works by Manship. Through LOD, they discover that Colby Museum of Art in Maine, the National Museum of American Wildlife Art in Wyoming, and Crystal Bridges Museum of American Art in Arkansas have works by Manship and related works in their collections. Documents on the AAC Website https://americanart.si.edu/about/american-art-collaborative and the Guide: Overview and Recommendations for Good Practices contain comments from partners pointing to additional benefits of LOD from their perspective. Some comments include, providing more meaningful content through more precise searching; being able to augment collection information by connecting to related art information in other museums; being able to tell fuller stories about works of art in collections; providing cross-domain linking; utilizing the LOD platform as a means of collaborating across museums; etc. As the founder of the initiative and project catalyst, LOD was also an opportunity to fulfill the vision of the Getty Information Institute as described in a video my staff and I produced in the 1994: The virtual database: (https://drive.google.com/file/d/1IuavP0M7HThcYydGAEpzxGKebmGSyKwU/view?usp=sharing

[3] AAC partners include: Amon Carter Museum of American Art, Archives of American Art, Autry Museum of the American West, Colby College Museum of Art, Crystal Bridges Museum of American Art, Dallas Museum of Art, The Thomas Gilcrease Institute of American History and Art, Indianapolis Museum of Art at Newfields, National Museum of Wildlife Art, National Portrait Gallery, Princeton University Art Museum, Smithsonian American Art Museum, Walters Museum of Art, Yale Center for British Art.

(2)To enable the creation of a critical mass of LOD for testing and learning purposes within the limits of a two-year funding timeframe, it was decided to engage the Information Services Institute (ISI) at the University of Southern California (USC) to convert the data from the fourteen participating institutions to LOD. ISI has experience in producing LOD and they have developed an open-source data integration tool called Karma. Karma can 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.

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. After converting the data, training was to be provided going forward so that each of the partners would be able to update and maintain their own LOD.

(3) As stated earlier, in addition to using RDF, an ontology or multiple ontologies are needed for expressing the relationships between bits of information particularly the triples: subject, predicate, and object. 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. AAC selected the CIDOC CRM, created by the International Committee on Documentation of the International Council of Museums (http://www.cidoc-crm.org). It is currently the most comprehensive ontology in the domain of cultural heritage 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.

(4) 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 end. 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.

Although ISI would initially convert the data from the institutions to LOD to jump-start the project, training was provided to enable each institution to update and maintain its own data going forward. 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.

Key Lessons Learned

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 and 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.

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 curators who know the most about the objects in the museum best apply it. 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 does not define inverse relationships, and mapping does not follow best practices, such as reusing existing ontologies.

The Need for a Target Model

Different schools of thought among advisors on how to interpret and thus apply the CIDOC CRM to AAC data, and inconsistencies in the mapping handled 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 (http://review.americanartcollaborative.org). Although funding to produce the model (later named Linked.Art) 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, Linked.Art, 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 coexist together. Allowing 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. Since producing the AAC target model, Linked. Art (https://linked.art) is being updated by Digital Getty to produce a defacto standard through international working groups and by applying it to other data sets, such as the Getty Museum and Pharos, The International Consortium of Photo Archives.

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.

AAC's Pipeline and Its Challenges

Figure 1 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). 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, and applets). A demonstration browse application illustrated the potential of Linked Data for exploring information about the subject of American art.

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, the International Image Interoperability Framework, IIIF https://iiif.io/ 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 an 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.

Difficulties in Understanding Museum Data

ISI hired USC information technology students to convert the data. 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. 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.

Communication Gaps between Project Staff and Museum Senior Management

Before AAC could move forward, each institution was asked to provide a letter of commitment signed by its director. In many respects, AAC representatives were able to obtain support from senior management by pointing out that the project allowed their museum to accomplish LOD through grant funding and the formation of a cooperative that would include teamwork, leveraging technical expertise, and training. Several of the representatives pointed out that they would never have been given the green light to produce LOD on their own! During the two-year course of the project, it was unclear if project representatives periodically updated senior management on AAC's progress, challenges, or lessons learned. AAC achieved more than it set

out to achieve. At the end of the project cycle, each institution mounted a webpage containing their own LOD and a link to the entire 230,000 LOD records drawn from the collaborative. However, when it became time to seek another grant to update and refine the LOD published, representatives felt their senior management would not deem AAC a priority over other grants their institution wanted to seek.

It was unclear if the representatives had alerted senior management in advance that additional funding would be needed or that their institution might need to become the lead institution. Although the Andrew W. Mellon and Terra Foundations were willing to consider a grant, none of the AAC representatives were willing to take on an administrative role and submit proposals on behalf of the collaborative. In hindsight, AAC advisors and management should have made sure senior museum officials understood the significance of AAC, what is was accomplishing and next steps. It is unclear if AAC's sudden halt will undermine one of its primary goals: inspiring the broader museum community to engage in LOD by example and by providing methodologies, lessons learned, guidelines, and tools.

Good Practices Recommendations

The lessons learned yielded the following good practices recommendations to help the broader museum community plan and engage in LOD initiatives (For a complete description of the Good Practices, see Part. 2. Recommendations for Good Practices When Initiating Linked Open Data (LOD) in Museums and Other Cultural-Heritage Institutions: https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAcce ssible.pdf?sequence=1&isAllowed=y)

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.

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://art.thewalters.org/license/), and those devised for the Yale Center for British Art, New Haven, Connecticut (https://britishart.yale.edu/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. For example, CC BY requires that attribution always be cited (e.g., "created BY this person/institution") and 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"), etc. (for additional possibilities, see Part 2 https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAccessible.pdf?sequence=1&isAllowed=y)

If you are participating in a collaborative multi-institutional project, the choice may require your partners to accept the 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 and linked. 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 does not mean it is not helpful if applied more generally.

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 (for additional advice, see Part 2.
https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAcce ssible.pdf?sequence=1&isAllowed=y)

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)—that can be ingested by the specific software tools you are using in the LOD project.

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.

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

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.

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.

Conclusions and Way Forward

We 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 universally. 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. In addition, 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. It has provided open-source tools, lessons learned, recommendations for good practices, and a prototype browse application. AAC remains hopeful that it will succeed in obtaining 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.

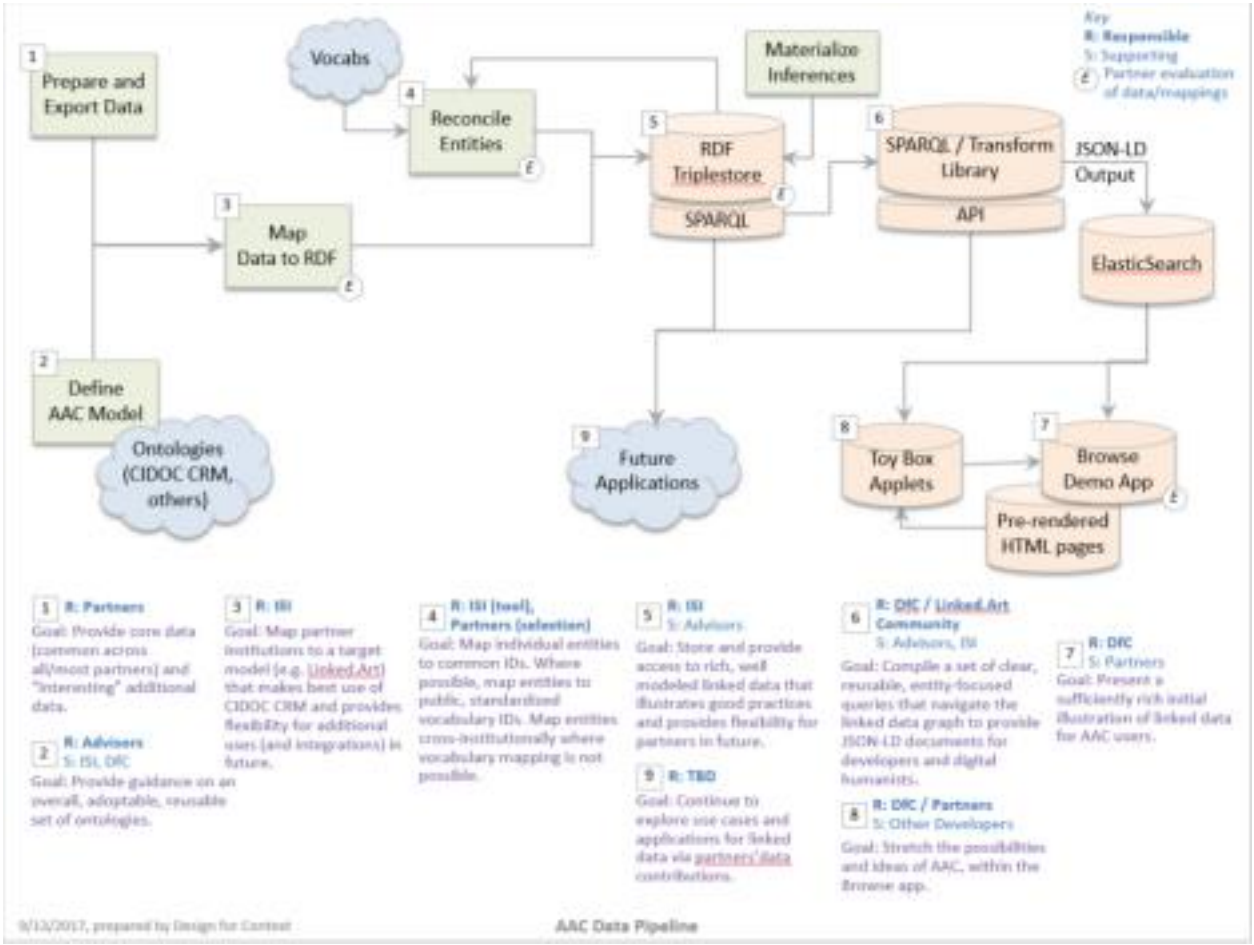

Fig.1 (source American Art Collaborative Linked Open Data Initiative: Overview and Recommendations for Good Practices, by Eleanor E. Fink, 2018).

Illustration of the AAC data 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).