AAC Road Map

Introduction

The American Art Collaborative (AAC), comprised of thirteen 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. 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 18 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 13 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; University of Southern California, Information Sciences Institute (ISI), with principal support from Pedro Szekely; Duane Degler and Neal Johnson, Design for Context; and Diane Zorich.

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 the Smithsonian American Art Museum, 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 44,000 art records from the Smithsonian's American Art Museum.

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 1-2 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 does 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 also 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 18 months: a kick off meeting, a mid-term 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 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. A second 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 13 years at Smithsonian and then at the J. Paul Getty Trust, initially as founder of the Getty vocabulary program, as 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 in reviewing the applications of the CRM to data mapping and to provide a hands on workshop and data expertise as needed throughout the 18 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.

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 over 30 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 on 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 Doerr: 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, ISO21127:2006, a core ontology for the purpose of schema integration across institutions.

VII: Expected Project Results

By the end of the project, at least 100,000 records pertaining to American art from 14 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.