Museum Data Exchange: Learning How to Share

Final Report to The Andrew W. Mellon Foundation

Günter Waibel
Ralph LeVan
Bruce Washburn

OCLC Research

A publication of OCLC Research

Museum Data Exchange: Learning How to Share
Waibel, et. al., for OCLC Research

February 2010

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-424-8 (978-1-55653-424-9)
OCLC (WorldCat): 503338562

Please direct correspondence to:
Günter Waibel
Program Officer
waibelg@oclc.org

Suggested citation:
Waibel, Günter, Ralph LeVan and Bruce Washburn. 2010. *Museum Data Exchange: Learning How to Share*. Report produced by OCLC Research. Published online at: www.oclc.org/research/publications/library/2010/2010-02.pdf.

Contents

Figures

Executive Summary

The Museum Data Exchange, funded by The Andrew W. Mellon Foundation, brought together a group of nine museums and OCLC Research to create tools for data sharing, build a research aggregation and analyze the aggregation. The project established infrastructure for standards-based metadata exchange for the museum community and modeled data sharing behavior among participating institutions.

Tools

The tools created by the project allow museums to share standards-based data using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

- COBOAT allows museums to extract Categories for the Description of Works of Art (CDWA) Lite XML out of collections management systems.
- OAICatMuseum 1.0 makes the data harvestable via OAI-PMH.

COBOAT's default configuration targets Gallery Systems' TMS, but can be adjusted to work with other vendor-based or homegrown database systems.

Both tools are a free download from:
http://www.oclc.org/research/activities/museumdata/.

Configuration files adapting COBOAT to different systems can be shared at:
http://sites.google.com/site/museumdataexchange/.

- ➢ For more detail, see *Phase 1: Creating tools for Data Sharing* on page 12*.*

Data Harvesting and Analysis

Harvesting data from nine museums, the project brought together 887,572 records in a non-public research aggregation, which participants had access to via a simple search interface. The analysis showed the following:

- for CDWA Lite required and highly recommended data elements, 7 out of 17 elements are used in 90% of the contributed records
- the match rate against applicable Getty vocabularies for *objectWorkType*, *nameCreator* and *roleCreator* is approximately 40%
- the top 100 *objectWorkType* and *nameCreator* values represent 99% and 49% of all aggregation records respectively.

Significant improvements in the aggregation could be achieved by revisiting data mappings to allow for a more complete representation of the underlying museum data. Focusing on the top 100 most highly occurring values for key elements will impact a high number of corresponding records, and would be low-hanging fruit for data clean-up activities.

For further analysis, the research aggregation will be available for third party researchers under the terms of the original agreements with participating museums.

- ➢ For more detail, see *Phase 2: Creating a Research Aggregation* on page 17 and
- ➢ *Phase 3: Analysis of the Research Aggregation* on page 21.

Impact

In its relatively short life span to date, the project's suite of tools has catalyzed several data sharing activities among project participants and other museums:

- The Minneapolis Institute of Arts uses the tools in a production environment to contribute data to ArtsConnected, an aggregation for K-12 educators.
- The Yale University Art Museum and the Yale Center for British Art use the tools to share data with a campus-wide cross-search, and contribute to a central digital asset management system.
- The Harvard Art Museum and the Princeton University Art Museum are actively exploring OAI harvesting with ARTstor. (Three additional participants have signaled that this would be a likely use for their OAI infrastructure as well.)

Participating vendors contributed to the museum community's ability to share:

- Gallery Systems extended COBOAT for EmbARK, demonstrating the extensibility of the MDE approach.
- Selago Design created custom CDWA Lite functionality for MIMSY XG, freely available to customers as part of their OAI tools.

An increasing number of projects and systems using CDWA Lite / OAI-PMH as a component (for example *OMEKA, Steve: The museum social tagging project, CONA™*) can be seen as a leading indicator for the future need of data sharing tools like the ones created as part of the Museum Data Exchange. When there are applications for sharing data which directly support the museum mission, more data is shared, and museum policies evolve. Conversely, when more data is shared, more such compelling applications emerge.

- ➢ For more detail, see *Compelling Applications for Data Exchange Capacity* on page 40.

Introduction

Data Sharing in Fits and Starts

Digital systems and the idea of aggregating museum data have a longer history than the availability of integrated access to museum resources in the present would suggest. As early as 1969, a newly formed consortium of 25 US art museums called the Museum Computer Network (MCN) and its commercial partner IBM declared, “We must create a single information system which embraces all museum holdings in the United States” (IBM et al. 1968). In collaboration with New York University, and funded by the New York Council of the Arts and the Old Dominion Foundation, MCN created a “data bank” (Ellin 1968, 79) which eventually held cataloging information for objects from many members of the New York-centric consortium, including the Frick Collection, the Brooklyn Museum, the Solomon R. Guggenheim Museum, the Metropolitan Museum of Art, the Museum of Modern Art, the National Gallery of Art and the New York Historical Society (Parry 2007).

However, using electronic systems with an eye towards data sharing was a tough sell even back in the day: when Everett Ellin, one of the chief visionaries behind the project and then Assistant Director at the Guggenheim, first shared this dream with his Director, he remembers being told: "Everett, we have more important things to do at the Guggenheim" (Kirwin 2004). The end of the tale also sounds eerily familiar to contemporary ears:

> “The original grant funding for the MCN pilot project ended in 1970. Of the original fifteen partners, only the Metropolitan Museum and the Museum of Modern Art continued to catalog their collections using computerized methods and their own operating funds.” (Misunas et al.)

Today, the museum community arguably is not significantly closer to a “single information system” than 40 years ago. As Nicholas Crofts aptly summarizes in the context of universal access to cultural heritage: “We may be nearly there, but we have been “nearly there” for an awfully long time.” (Crofts 2008, 2)

Not for lack of trying, however, as a non-exhaustive selection of strategies and experiments to standardize museum data exchange in the US highlights:

- The AMICO Library of digital resources from museums (conceived in 1997, a full year before eXtensible Markup Language (XML) became a W3C recommendation) created a data format consisting of a field-prefix (such as OTY for Object Type) and the field delimiter “}~” to exchange information (AMICO).
- In 1999, a consortium of California institutions (MOAC) implemented a mark-up standard from the archival community (Encoded Archival Description or EAD) to bring their resources into an existing state-wide aggregation of library special collections and archival content.

- Between 1998 and 2003, the CIMI consortium launched a range of projects exploring data standards and protocols for exchange, including Z39.50, Dublin Core and the UK standard SPECTRUM.

All of these initiatives had merit in their particular historical context as well as a heyday of adoption, yet none of these strategies achieved consensus and wide-spread use over the long term.

The most contemporary entry in the history of museum data sharing is Categories for the Description of Works of Art (CDWA) Lite XML (Getty Trust 2006). In 2005, the Getty and ARTstor created this XML laschema "to describe core records for works of art and material culture" that is "intended for contribution to union catalogs and other repositories using the Open Archives Initiative (OAI) harvesting protocol" (Getty Research Institute n.d.). Arguably, this is the most comprehensive and sophisticated attempt yet to create consensus in the museum community about how to share data.

The complete CDWA Lite data sharing strategy comprises:

- A data structure (CDWA) expressed in a data format (CDWA Lite XML)
- A data content standard (Cataloging Cultural Objects—CCO)
- A data transfer mechanism (Open Archives Initiative Protocol for Metadata Harvesting—OAI-PMH)

What follows is a brief example of how these different specifications work hand in hand to establish standards-based, shareable data:

- CDWA, a data field and structure specification, defines a discrete unit of information such as "Creation Date" with sub-categories for "Earliest Date" and "Latest Date."
- CCO, a data content standard, specifies the rules for formatting a date as "Late 14th century" for display and using an ISO 8601 format "1375/1399" for machine indexing.
- CDWA Lite XML, a data format, allows the encoding of all this information, as shown in the code snippet below:

```
<cdwalite:displayCreationDate>Late 14th century</cdwalite:displayCreationDate>
<cdwalite:indexingDatesWrap>
        <cdwalite:indexingDatesSet> <cdwalite:earliestDate>1375</cdwalite:earliestDate>
        <cdwalite:latestDate>1399</cdwalite:latestDate> </cdwalite:indexingDatesSet>
</cdwalite:indexingDatesWrap>
```

- OAI-PMH, a data exchange standard, allows sharing the resulting record. The protocol supports machine-to-machine communication about collections of records, including retrieval from a content provider's server by an OAI-PMH harvester. It also supports synchronizing local updates with the remote harvester as the museum data evolves (Elings and Waibel 2007).

The Museum Data Exchange (MDE) project outlined in this paper attempts to lower the barrier for adoption of this data sharing strategy by providing free tools to create and share CDWA Lite XML descriptions, and helps model data exchange with nine participating museums. The activities were generously funded by The Andrew W. Mellon Foundation, and supported by OCLC Research in collaboration with museum participants from the RLG Partnership. The project's premise: while technological hurdles are by no means the only obstacle in the way of more ubiquitous data sharing, having a no-cost infrastructure to create standards-based descriptions should free institutions to debate the thorny policy questions which ultimately underlie the 40 year history of fits and starts in museum data sharing.

Early Reception of CDWA Lite XML

The launch of CDWA Lite XML was officially announced at the MCN annual conference in Boston on November 5, 2005. The following two data points help illuminate its reception by the community. A small survey among ten prominent museums from the RLG Partnership (seven from the United States, two from the United Kingdom, one from Canada) conducted by Günter Waibel approximately six months after the initial launch of CDWA Lite XML showed that:

- Capabilities for exporting standards-based data of any kind (including CDWA Lite XML) are non-existent.
- Policy issues are a major obstacle to providing access to high-quality digital images. No museum provides free access to publication-quality digital images of artworks in the public domain without requiring a license (one museum has plans), while nine museums license publication-quality digital images for a fee.
- A limited amount of data sharing already happens, primarily with subscription-based resources. While eight museums provide access to digital images on their Web site, four museums contribute to licensed aggregations such as ARTstor or CAMIO, and two contribute to non-licensed aggregations such as state-wide or national projects.

Approximately 18 months after the launch of CDWA Lite XML, the newly minted CDWA Lite Advisory Committee[1] surveys the cultural heritage community writ large to gauge the impact of CDWA Lite, and finds the following:

- CDWA Lite XML garners great interest: 144 respondents (50.7% from museum community) start the 22 question survey.
- Even among the self-selecting group of those taking the survey, few have the experience to complete it: only the first three questions have responses from a majority of respondents, while the numbers drop precipitously once questions presuppose basic working knowledge of CDWA Lite. Only 22 individuals complete the survey.

Given this backdrop, an RLG Programs/OCLC working group called "Museum Collection Sharing," (OCLC Research n.d.c) inaugurated in May 2006, sought to support increased use of the fledgling CDWA Lite strategy by providing a forum for museum professionals to share information and collaborate on implementation solutions. The group identified the following hurdles for getting museum data into a shareable format:

- The complexities of mapping data in collections management systems to CDWA Lite
- The absence of mechanisms to export data out of collections management systems and transform it into CDWA Lite XML
- The complexities of configuring and running an OAI-PMH data content provider

Circumstances made the creation of an OAI-PMH data content provider which "speaks" CDWA Lite XML the lowest-hanging fruit on the list. In their proof-of-concept project with ARTstor, The Getty had implemented a modified version of OAICat, an open source OAI data provider originally written by

Jeff Young (OCLC Research). In collaboration with the working group and supported by Jeff, OCLC Research released a CDWA Lite enabled version of OAICat (OAICatMuseumBETA) in the fall of 2007.

Unfortunately, parallel investigations into widely applicable mechanisms to create CDWA Lite XML records did not immediately bear fruit. For example, the working group discussed the possible application of OCLC's Schema Transformation technology (OCLC Research n.d.b) with Jean Godby (OCLC Research) and explored Crystal Reports, a report writing program bundled with many collections management systems, to output CDWA Lite XML. However, the release of OAICatMuseumBETA provided the impetus for funding from The Andrew W. Mellon foundation to remedy a situation in which museums on the working group had a tool to serve CDWA Lite XML records, yet had no capacity to create these records to begin with.

Grant Overview

The grant proposal funded by The Andrew W. Mellon Foundation in December 2007 with $145,000[2] contained the following consecutive phases, which will also structure the rest of this paper.

Phase 1: Creation of a Batch Export Capability

The grant proposed to make a collaborative investment into a shared solution for generating CDWA Lite XML, rather than many isolated local investments with little community-wide impact. Grant participants aimed to leverage the experience some institutions on the Museum Collection Sharing working group had gained from exploring local solutions to create a common solution. The Yale University Art Gallery, for example, had started developing a command-line tool using customizable SQL files which create database tables corresponding to CDWA Lite; the Metropolitan Museum of Art was working with ARTstor on a CDWA Lite / OAI data-transfer solution as part of the Images for Academic Publishing (IAP) program. To keep the grant manageable and within budget, we limited our investigation to an export mechanism for Gallery Systems' TMS, the predominant database among the museums in the Collection Sharing cohort.

Museum partners: Harvard Art Museum (originally Museum of Fine Arts, Boston; the grant migrated with staff from the MFA to Harvard early in the project); Metropolitan Museum of Art; National Gallery of Art; Princeton University Art Museum; Yale University Art Gallery.

Phase 2: Model Data Exchange Processes through the Creation of a Research aggregation

The grant proposed to model data exchange processes among museum participants in a low-stakes environment by creating a non-public aggregation with data contributions utilizing the tools created in Phase 1, plus additional participants using alternative mechanisms. The grant purposefully limited data sharing to records only—including digital images would have put an additional strain on the harvesting process, and added little value to the predominant use of the aggregation for data analysis (see Phase 3 on the next page).

Museum partners: all named in Phase 1, plus the Cleveland Museum of Art and the Victoria & Albert Museum (both contributing through a pre-existing export mechanism); in the process of the grant, data sets from the Minneapolis Institute of Arts and the National Gallery of Canada were also added.

Phase 3: Analysis of the Research Aggregation

The grant proposed to surface the characteristics of the research aggregation, both its potential utility and limitations, through a data analysis performed by OCLC Research. The CDWA Lite / OAI strategy had been expressly created to support large-scale aggregation—however, would the museum data transported by these means actually come together in a meaningful way?

A minimal interface to the research aggregation would make cross-collection searching available to museum participants.

Museum partners: all nine institutions named under Phase 1 and Phase 2.
All individuals who had a significant role in the activities surrounding the grant are acknowledged in Appendix A. Project Participants (page 45).

Phase 1: Creating Tools for Data Sharing

The first face-to-face project meeting at the Metropolitan Museum of Art in January 2008 resulted in the following draft system architecture for a data extraction tool, which distilled our far-ranging discussions around the required functionality into a single graphic.

This quote, like Figure 1 taken from the original meeting minutes, explains the envisioned flow of the data:

> “The Data Extraction Tool obtains data from the Source Database through application of SQL based mapping profiles. It will store the resulting output in a new and separate CDWA Lite Work Database that resides on a server behind the institution’s firewall. The Work Database provides an efficient means of representing the data defined by the CDWA Lite standard and the OAI header. A Database Publishing Tool will be configurable to push data across the firewall to the Public OAI CDWA Lite XML Database. In addition, the tool will be capable of publishing CDWA Lite XML records with an OAI wrapper to the Public OAI CDWA Lite XML File System, or CDWA Lite XML records with or without and OAI wrapper to an Internal CDWA Lite XML File System. Either the public File System or XML Database could be accessed by an OAI repository to respond to HTTP queries from the Web.”

While Figure 1 and its description hint at the emerging complexity of the grant’s endeavor, some of the devils are still hiding in the details. For example, even a tool providing a solution solely for TMS needs to support significant variability in the source data model: it needs to adapt to a variety of different product versions of TMS used by different project participants, as well as different implementations of the same product version by different project participants. In addition, the tool needs to adapt to a variety of different practices within an institution: the Metropolitan Museum, for example, is running twenty installations of TMS controlled by different departments, while for other participants, a single instance of TMS within a museum contains considerable variability because different departments use that single instance according to different guidelines.

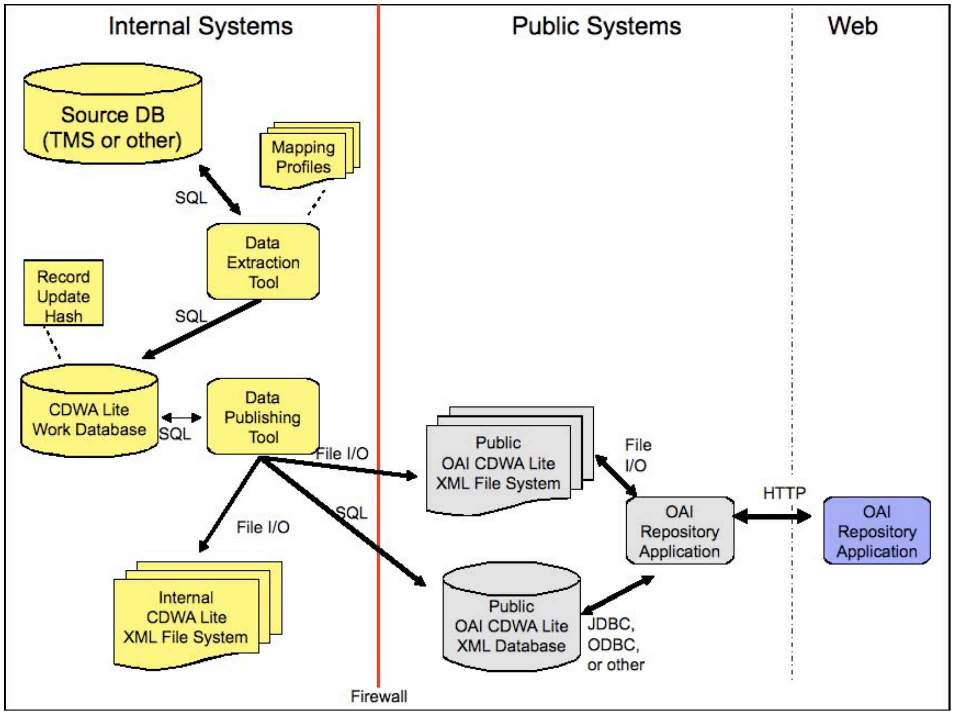


Figure 1. Draft system architecture for a CDWA Lite XML data extraction tool

Supporting crucial OAI-PMH features created additional requirements for the tool: it needs to keep track of updates to the TMS source data so it only regenerates CDWA Lite XML for updated records, and is capable of communicating these updates through OAI-PMH. In addition, the tool needs to be able to mark records as belonging to an OAI-PMH set so museums can create differently scoped packages of metadata for different harvesters.

In short, our first project meeting surfaced a mismatch between required features, timeline and budget for Phase 1 of the grant. In addition, the meeting exposed tension between the open source requirement of the grant, and official policies at the majority of participating museums, which did not have resources for open source development, and supported Microsoft Windows exclusively. While everybody around the table wanted to create an open source solution, lack of support for open source within the group constituted a serious risk factor for successful implementation. Apparently, others shared the concern that overall requirements, timeline and budget for the project were out of sync. The response from open source developers in the museum community who received our RFP was tepid, and only one party wanted to discuss details.

Ben Rubinstein, Technical Director at Cognitive Applications Inc. (Cogapp), a UK consulting firm with a long track-record of compelling museum work, presented us with an intriguing solution to our conundrum. As a by-product of many museum contracts which required accessing and processing data from collections management systems, Cogapp had developed a system called COBOAT (Collections Online Back Office Administration Tool). As part of our project, Ben proposed, Cogapp would make a fee-free, closed-source version of COBOAT available, while creating an open-source, plug-in module which trained the tool to convert data into CDWA Lite XML. Leveraging an existing tool allowed the project to stay within budget limits; creating the open-source plug-in with grant money allowed us to stay within our funding mandate; the overall package would be a good fit for the Microsoft Windows platforms commonly supported at most project participant sites. After review with project participants, Cogapp was awarded the contract to create the batch export capability envisioned by the grant.

COBOAT and OAICATMuseum 1.0: Features and Functionality

The suite of tools which emerged as part of the MDE project includes both COBOAT and an updated version of OAICatMuseum.

COBOAT is a metadata publishing tool developed by Cogapp that transfers information between databases (such as collections management systems) and different formats. As implemented in this project, COBOAT allows museums to extract CDWA Lite XML out of Gallery Systems' TMS. With configuration files, COBOAT can be adjusted for extraction from different vendor-based or homegrown database systems, or locally divergent implementations of the same collections management system. COBOAT software is available under a fee-free license for the purposes of publishing a CDWA Lite repository of collections information at http://www.oclc.org/research/activities/coboat/.

OAICatMuseum 1.0 is an OAI-PMH data content provider supporting CDWA Lite XML which allows museums to publish the data extracted with COBOAT. While COBOAT and OAICatMuseum can be used separately, they do make a handsome pair: COBOAT creates a MySQL database containing the CDWA Lite XML records, which OAICatMuseum makes available to harvesters. The software upgrade from BETA to 1.0 was created by Bruce Washburn in consultation with Jeff Young (both OCLC Research). OAICatMuseum 1.0 is available under an open source license at http://www.oclc.org/research/activities/oaicatmuseum/.

More details: COBOAT's default configuration files make a best-guess effort to run a complete output job, which includes retrieving data from TMS, transforming it into CDWA Lite records, and outputting them to a MySQL database that can become a data source for OAICatMuseum. While the configuration files have evolved through the experience of museum participants, new implementations will likely require modifications to adapt to local practice. The first unpolished export of a handful of sample XML records helps pin-point areas for improvement.

COBOAT can be run across all TMS records, or a predefined subset; in addition, it keeps track of changes to the data source, and outputs updated modification dates for edited records to OAICatMuseum. (In this way, the suite of tools allows a data harvester to request only records which have been updated, instead of re-harvesting a complete set.) Based on an "OAISet" marker in TMS object packages, COBOAT also generates data about OAI sets. (This allows a museum to expose differently scoped sets of data for harvesting.) Beyond CDWA Lite, OAICatMuseum also offers Dublin Core for harvesting, as mandated by the OAI-PMH specification. The application creates Dublin Core from the CDWA Lite source data on the fly via a stylesheet.

The following diagram provides an overview of the modules contained in COBOAT, and the configuration files which instruct different processes.

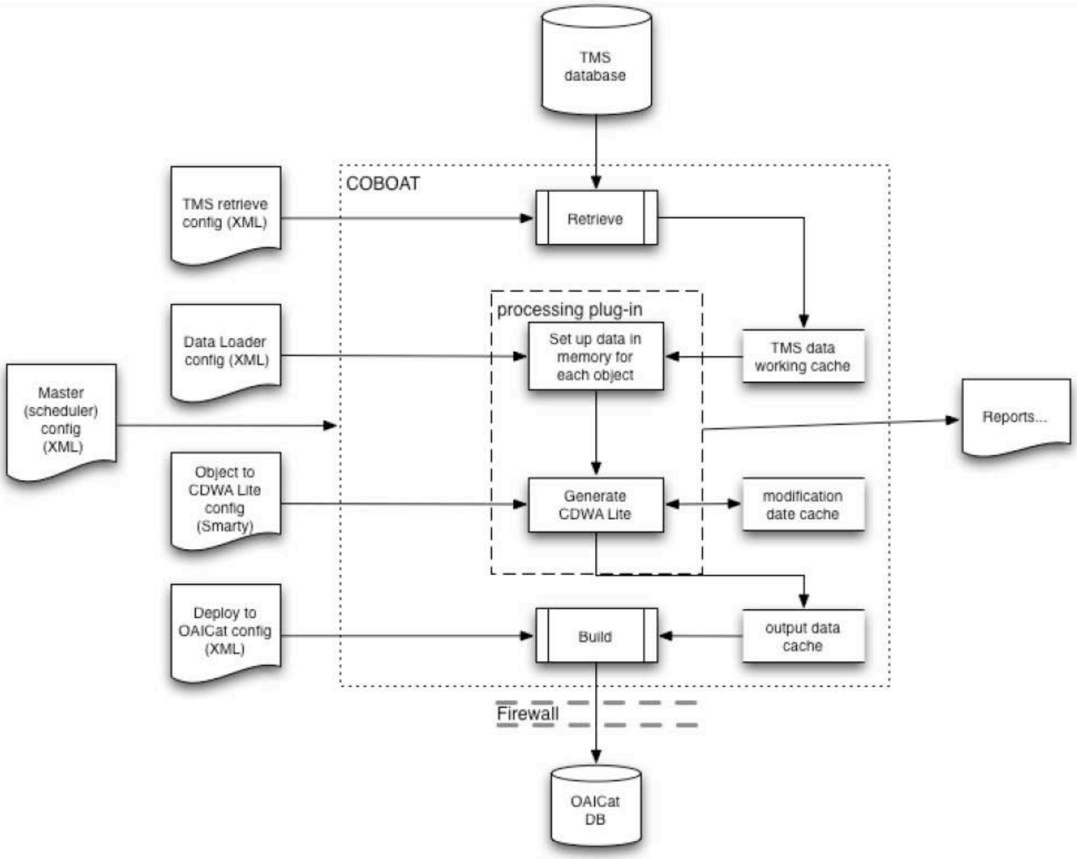


Figure 2. Block diagram of COBOAT, its modules and configuration files

COBOAT consists of a total of five modules. Each of these modules can be customized through configuration files or scripts. Primary to extracting and transforming data to CDWA Lite XML, as well as adapting COBOAT to different databases or database instances, are the following three modules:

- Retrieve module: extracts data out of a database (by default, TMS). The retrieve configuration file (XML) determines which data to grab, and creates a series of text files from the database tables.
- Processing module or plug-in: performs transformation to CDWA Lite XML. Two configuration files are used for this procedure: the 1st pass data loader script (XML) assembles data arrays, while the 2nd pass renderer script (Smarty[3]) turns the data arrays into CDWA Lite XML.
- Build module: the build script (XML) outputs the data to a simple MySQL database (used by OAICatMuseum)

While the MDE project exclusively implemented COBOAT with TMS, it can be extended to other database systems. With the appropriately tailored configuration files, COBOAT can retrieve data from Oracle, MySQL, Microsoft Access, FileMaker, Valentina or PostgreSQL databases, as well as any ODBC data source (such as Microsoft SQL Server).

Gallery Systems has tested the flexibility of COBOAT by running it against its EmbARK product, which is based on the 4th Dimension database system and has a different data structure from TMS. Slight modifications of COBOAT were required to support data extraction from tables and fields with an initial underscore in their names. According to Robb Detlefs (Director of West Coast Operations and Strategic Initiatives at Gallery Systems), the COBOAT configuration files were easily adapted to extract and transform the data, and the renderer script provided sufficient means for applying logic to the data from the original system. Several EmbARK clients are expected to implement COBOAT in the near future.

The MDE project has set up a Web site at http://sites.google.com/site/museumdataexchange/ where configuration files for COBOAT can be discussed and shared. These configuration files could either represent extensions to different database systems, or tweaks of default files to adapt them to a particular instance of an already covered database. The EmbARK configuration files are available at this site.

Implementing and Refining the Suite of Tools

In order to extend the pre-existing COBOAT application to include CDWA Lite XML capability and arrive at a default TMS configuration for the tool, Cogapp built a first instance of the new processing plug-in and tested it with two museum participants. The Metropolitan Museum of Art, with twenty stand-alone instances of TMS and upwards of 300K records one of the project's most complex cases, became the first implementer. In parallel, Cogapp worked with the Princeton University Art Museum—as a smaller institution with fairly limited technical support, this museum represented the other end of the spectrum. Once the entire suite of tools, including OAICatMuseum 1.0, had been implemented at both of these institutions with considerable support from Cogapp and OCLC Research, the remaining museums faced the task of installing the applications as if they had simply downloaded them from the Internet, with no initial support other than the manuals.

In addition to the five museum participants named in the grant, the Minneapolis Institute of Arts added yet another layer of testing for the suite of tools. To support data sharing between the Institute and the Walker Art Museum as part of the ongoing redesign of ArtsConnectEd (Minneapolis Institute of Arts and Walker Art Center. n.d.), the Institute of Arts installed a pre-release version of COBOAT / OAICatMuseum in a production environment (Dowden et al. 2009). The information architecture of ArtsConnectEd revolves around OAI harvesting of CDWA Lite records from each of the contributors, and the MDE tools matched the Institute's needs for a readily implementable CDWA Lite / OAI solution.

When all five Phase 1 museums plus the Minneapolis Institute of Arts museums had tried their hands at implementing the tools, they found COBOAT eminently suitable to the task of transforming their collection data into CDWA Lite XML. Those with slightly higher technical proficiency tended to find the tool easier to use than those with less technical support. The in-depth documentation for COBOAT won universal acclaim, and the additional high-level Quick Start Guide museums wanted to see is now part of the tool's download. Multiple museum representatives commented that matching up a desired effect on the output with the appropriate configuration file seemed like one of the biggest hurdles to jump. The considerable flexibility built into COBOAT clearly has its learning curve, and rewards those who make the time to familiarize themselves with the possibilities. On the other hand, museums also commented on the instant gratification of executing the initial default export, which invariably produced very encouraging, if not perfect, results. An e-mail on the project list read: "After just the initial run of this app I think I might be able to say I'm a 'CogApp Fanboy.' Got any t-shirts?"

Some figures courtesy of the Harvard Art Museum exemplify resource needs and runtime for COBOAT. While there are many variables impacting runtime, the entire process of retrieving, transforming and loading 236,466 records took 85 minutes at Harvard. The raw retrieved data required approximately 80MB of space, while the temporary data created by the processing module occupied 0.9 GB, and the final MySQL production database 2.7GB.

The museums who had implemented OAICatMuseum pronounced it a solid player of the team, with the only caveat being that memory allocations had to be monitored carefully, especially for harvests upwards of 50K records. Increasing the Java Virtual Machine (JVM) memory allocation from 512 (default) to 1GB enabled larger harvests. At over 110K records, the Metropolitan's dataset constituted the largest gathered with OAICatMuseum as part of this project. While there are many variables influencing the duration of a harvest, the Metropolitan OAI data transfer took about 24 hours.

Phase 2: Creating a Research Aggregation

Legal Agreements

A legal agreement governed the data transfers between museum participants and OCLC Research. In the spirit of creating a safe sand-box environment for experimenting with the technological aspects of data sharing, the 1½ page agreement aimed to clarify that access to all data would remain limited to the participating museums; that the data would be purged one year after publication of the project report; that OCLC Research data analysis findings (part of this report) dealing with museum data would be anonymized; and that legitimate third party researchers could petition for access to the aggregation under identical terms to augment the communities knowledge of aggregate museum data.

Observations on the process of executing the agreements reflect how complex data sharing can become in the absence of a community consensus around common policies and behaviors. This, in and of itself, constitutes a finding of the project. With a single exception, museums asked for relatively minor changes in the agreement—nevertheless, the entire process of executing agreements took six months to complete. The single biggest factor in delays seemed to be the different comfort levels of the museum staff working on the project, and legal council and administrators reviewing the agreement. In the process, some institutions which had planned to contribute all of their collection records to the research aggregation had to scale back to a subset. On the other hand, four institutions signed the agreement within six weeks of receipt with practically no changes, highlighting how different the processes, precedents and policies impinging on the decision were at each institution.

Harvesting Records

Not every participant in the grant used COBOAT and OAICatMuseum to encode and transfer their data. For the research aggregation and data analysis portion of the project, three institutions used alternative means to create and share CDWA Lite records.

- The Cleveland Museum of Art used a pre-existing mechanism for creating CDWA Lite records on the fly from their Web online collection database in response to OAI-PMH requests. (Incidentally, this mechanism was built by Cogapp.)

- The Victoria & Albert Museum transformed an XML export from their MUSIMS (SSL) collections management system into CDWA Lite using stylesheets, and ftp'd the data.
- The National Gallery of Canada, like the Minneapolis Institute of Arts, joined the project as a non-funded partner once shared interests emerged. The Gallery's vendor Selago Design crucially enabled their participation by prototyping a CDWA Lite / OAI capacity in their MIMSY XG [4] collections management system, for which the MDE harvest constituted the first test.

A little bit more detail on the MIMSY XG solution: according to James Starrit (Manager of Web Development, Selago Design), a MIMSY OAI-PMH tool set already existed when Gayle Silverman (Director of Community Relations, Selago Design) approached the National Gallery about participating in the MDE project with Selago's support. This OAI provider was developed by Selago using PHP (with OCI8/Oracle extensions), and can be used with unqualified Dublin Core, and extended to other standards via templates. To support CDWA Lite, the appropriate mappings for the National Gallery of Canada's bilingual dataset had to be created. As a result of this work, CDWA Lite is now included in the OAI tool set, and available to any MIMSY XG user.

Of all nine institutions whose records the project acquired, OAI-PMH was the transfer mechanism of choice in six cases, with four institutions using OAICatMuseum, and two an alternate OAI data content provider (Cleveland and the National Gallery of Canada). Two additional institutions wanted to employ OAICatMuseum, yet found their attempts thwarted. Policy reasons disallowed opening a port for the harvest at one museum; at another institution, project participants and OCLC Research ran out of time diagnosing a technical issue, and the museum contributed MySQL dump files from the COBOAT-created database instead. And last but not least, the Victoria & Albert Museum simply ftp'd their records.

Once a set of institutional records was acquired, OCLC Research performed an initial XML schema validation as a first health-check for the data. For two data contributors, all records validated. Among the other contributors, the health-check surfaced a range of issues:

- Element sequencing: valid elements were supplied, but not in the order defined by the CDWA Lite XML schema
- Incorrect paths: for example, missing a "Wrap" or "Set" element in the XML path
- Missing namespaces: for example, type attributes not preceded by "cdwalite:"
- Missing required elements: for example, *recordID* not provided inside *recordWrap*
- Invalid Unicode characters: Unicode characters 0x07 and 0x18 found in some records, preventing validation

The two validating record sets came from institutions using COBOAT who had not tweaked the default configuration files. Many of the element sequencing, incorrect path and missing namespace issues were introduced in those portions of the output which had been edited. While OCLC Research harvested data at least twice from each contributor, mostly to provide an opportunity to rectify schema validation errors, downstream processes and tools (described below) were flexible enough to also handle non-valid records.

Two lessons from harvesting the nine collections stand out:

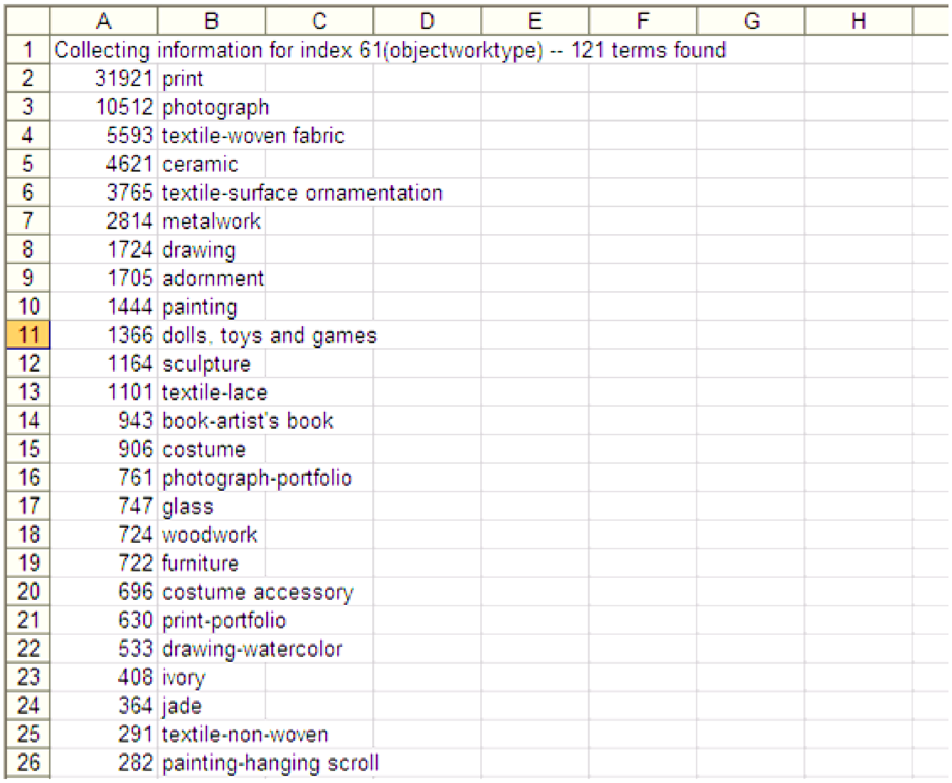
- First, OAI-PMH as a tool is ill-matched to the task of large one-time data transfers, compared to an ftp or rsync transfer of records, or an e-mail of mySQL dumps. Data providers reap the benefit of the protocol predominantly through its long-term use, when additions and updates to a data set can be effectively and automatically communicated to harvesters. Within the confines of the MDE project, however, OAI remained the preferred mode of data transfer, since the grant set an explicit goal of taking institutions through an OAI process.

- Second, the relatively high rate of schema validation errors after harvest leads to the conclusion that validation was not part of the process on the contributor end. Ideally, schema validation would have happened before data contribution. Validation provides the contributor with important evidence about potential mapping problems, as well as other issues in the data; in this way, it becomes one of the safeguards for circulating records which best reflect the museums data.

Preparing for Data Analysis

To prepare the harvested data for analysis, as well as to provide access to the museum data, OCLC Research harnessed the Pears database engine,[5] which Ralph LeVan (OCLC Research) outfitted with new reporting capabilities, and an array of pre-existing and custom-written tools. Pears ingests structured data, in this case XML, and creates a list of all data elements or attributes (referred to as "units of information" from here on out) which contain a data value. The database then builds indexes for the data values of each of these units of information. The values themselves remain unchanged, except that they are shifted to lower case during index building. For data analysis, these indexes provide a basis for grouping values from a specific unit of information for a single contributor, or the aggregate collection; as well as grouping the units of information themselves via tagpaths across the array of contributors.

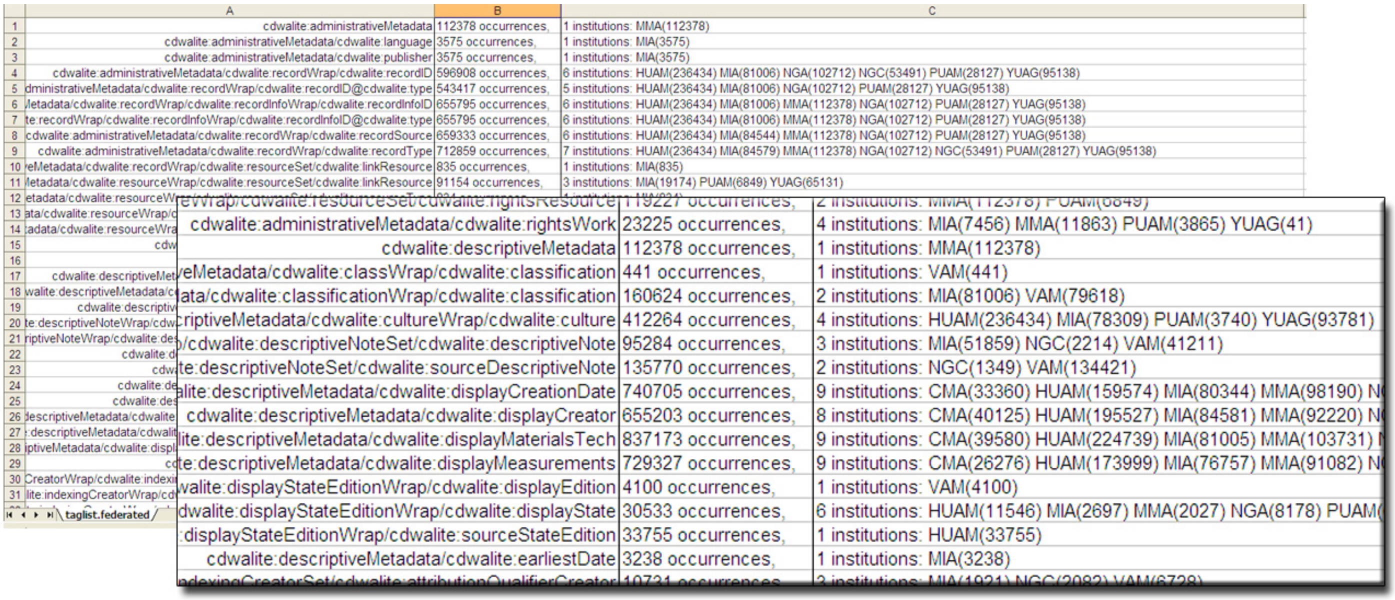
Figure 3 shows screen-shots of sample reports from Pears in spreadsheet format, which should help bring these abstract concepts to life. Column A provides counts for the number of occurrences for each data value. (Note that only 26 of the 121 values for *objectWorkType* are shown.) Just at a simple intuitive level, this report provides some valuable information: first of all, it demonstrates that *objectWorkTypes* for this contributor were limited to a finite number of 121; it shows that a number of terms were concatenated, probably in the process of exporting the data, to create values such as "drawing-watercolor"; at first glance, only these concatenated values seem to duplicate other entries (such as "drawing"). The occurrence data indicates a high concentration of objects sharing the same *objectWorkType*, with numbers quickly falling below a 1K count.



	A	B	C	D	E	F	G	H
1	Collecting information for index 61(objectworktype) -- 121 terms found							
2	31921	print						
3	10512	photograph						
4	5593	textile-woven fabric						
5	4621	ceramic						
6	3765	textile-surface ornamentation						
7	2814	metalwork						
8	1724	drawing						
9	1705	adornment						
10	1444	painting						
11	1366	dolls, toys and games						
12	1164	sculpture						
13	1101	textile-lace						
14	943	book-artist's book						
15	906	costume						
16	761	photograph-portfolio						
17	747	glass						
18	724	woodwork						
19	722	furniture						
20	696	costume accessory						
21	630	print-portfolio						
22	533	drawing-watercolor						
23	408	ivory						
24	364	jade						
25	291	textile-non-woven						
26	282	painting-hanging scroll						

Figure 3. Excerpt from a report detailing all data values for *objectWorkType* from a single contributor

Figure 4 shows an excerpt of a report which groups units of information from all contributors. Even in the impressionistic form of a screenshot, this report provides some valuable first impressions of the data. Data paths (column A) which have long lists of institutions associated with them (column C) represent units of information which are used by many contributors; occurrence counts in column B, when mentally held against the approximately 900K total records of the research aggregation, complement the first observation by providing a first impression of the pervasiveness with which certain units of information are used. Outlier data paths which have only a single institution associated with them often betray an incorrect path, as confirmed by schema validation results. While OCLC Research held fast to a principle of not manipulating the source data provided by museums, the single instance of data clean-up performed prior to analysis consisted in mapping stray data paths to their correct place.



	A	B	C
1	cdwalite:administrativeMetadata	112378 occurrences,	1 institutions: MMA(112378)
2	cdwalite:administrativeMetadata/cdwalite:language	3575 occurrences,	1 institutions: MIA(3575)
3	cdwalite:administrativeMetadata/cdwalite:publisher	3575 occurrences,	1 institutions: MIA(3575)
4	cdwalite:administrativeMetadata/cdwalite:recordWrap/cdwalite:recordID	596908 occurrences,	6 institutions: HUAM(236434) MIA(81006) NGA(102712) NGC(53491) PUAM(28127) YUAG(95138)
5	dministrativeMetadata/cdwalite:recordWrap/cdwalite:recordID@cdwalite:type	543417 occurrences,	5 institutions: HUAM(236434) MIA(81006) NGA(102712) PUAM(28127) YUAG(95138)
6	letadata/cdwalite:recordWrap/cdwalite:recordInfoWrap/cdwalite:recordInfoID	655795 occurrences,	6 institutions: HUAM(236434) MIA(81006) MMA(112378) NGA(102712) PUAM(28127) YUAG(95138)
7	te:recordWrap/cdwalite:recordInfoWrap/cdwalite:recordInfoID@cdwalite:type	655795 occurrences,	6 institutions: HUAM(236434) MIA(81006) MMA(112378) NGA(102712) PUAM(28127) YUAG(95138)
8	cdwalite:administrativeMetadata/cdwalite:recordWrap/cdwalite:recordSource	659333 occurrences,	6 institutions: HUAM(236434) MIA(84544) MMA(112378) NGA(102712) PUAM(28127) YUAG(95138)
9	cdwalite:administrativeMetadata/cdwalite:recordWrap/cdwalite:recordType	712859 occurrences,	7 institutions: HUAM(236434) MIA(84579) MMA(112378) NGA(102712) NGC(53491) PUAM(28127) YUAG(95138)
10	eMetadata/cdwalite:recordWrap/cdwalite:resourceSet/cdwalite:linkResource	835 occurrences,	1 institutions: MIA(835)
11	letadata/cdwalite:resourceWrap/cdwalite:resourceSet/cdwalite:linkResource	91154 occurrences,	3 institutions: MIA(19174) PUAM(6849) YUAG(65131)
12	etadata/cdwalite:resourceW		
13	ata/cdwalite:resourceWrap/c		
14	adata/cdwalite:resourceWra		
15	cdw		
16			
17	cdwalite:descriptiveMet		
18	walite:descriptiveMetadata/c		
19	cdwalite:descriptiv		
20	te:descriptiveNoteWrap/cdw		
21	riptiveNoteWrap/cdwalite:des		
22	cdwalite:d		
23	cdw		
24	cdwalite:de		
25	cdwalite:des		
26	descriptiveMetadata/cdwalite		
27	:descriptiveMetadata/cdwalit		
28	iptiveMetadata/cdwalite:displ		
29	cc		
30	CreatorWrap/cdwalite:indexi		
31	lite:indexingCreatorWrap/cd		

eWrap/cdwalite:resourceSet/cdwalite:rightsResource	119227 occurrences,	2 institutions: MMA(112378) PUAM(6849)
cdwalite:administrativeMetadata/cdwalite:rightsWork	23225 occurrences,	4 institutions: MIA(7456) MMA(11863) PUAM(3865) YUAG(41)
cdwalite:descriptiveMetadata	112378 occurrences,	1 institutions: MMA(112378)
eMetadata/cdwalite:classWrap/cdwalite:classification	441 occurrences,	1 institutions: VAM(441)
ata/cdwalite:classificationWrap/cdwalite:classification	160624 occurrences,	2 institutions: MIA(81006) VAM(79618)
criptiveMetadata/cdwalite:cultureWrap/cdwalite:culture	412264 occurrences,	4 institutions: HUAM(236434) MIA(78309) PUAM(3740) YUAG(93781)
/cdwalite:descriptiveNoteSet/cdwalite:descriptiveNote	95284 occurrences,	3 institutions: MIA(51859) NGC(2214) VAM(41211)
te:descriptiveNoteSet/cdwalite:sourceDescriptiveNote	135770 occurrences,	2 institutions: NGC(1349) VAM(134421)
alite:descriptiveMetadata/cdwalite:displayCreationDate	740705 occurrences,	9 institutions: CMA(33360) HUAM(159574) MIA(80344) MMA(98190) N
cdwalite:descriptiveMetadata/cdwalite:displayCreator	655203 occurrences,	8 institutions: CMA(40125) HUAM(195527) MIA(84581) MMA(92220) N
lite:descriptiveMetadata/cdwalite:displayMaterialsTech	837173 occurrences,	9 institutions: CMA(39580) HUAM(224739) MIA(81005) MMA(103731)
te:descriptiveMetadata/cdwalite:displayMeasurements	729327 occurrences,	9 institutions: CMA(26276) HUAM(173999) MIA(76757) MMA(91082) N
walite:displayStateEditionWrap/cdwalite:displayEdition	4100 occurrences,	1 institutions: VAM(4100)
dwalite:displayStateEditionWrap/cdwalite:displayState	30533 occurrences,	6 institutions: HUAM(11546) MIA(2697) MMA(2027) NGA(8178) PUAM(
:displayStateEditionWrap/cdwalite:sourceStateEdition	33755 occurrences,	1 institutions: HUAM(33755)
cdwalite:descriptiveMetadata/cdwalite:earliestDate	3238 occurrences,	1 institutions: MIA(3238)
ndexingCreatorSet/cdwalite:attributionQualifierCreator	10731 occurrences,	3 institutions: MIA(1921) NGC(2082) VAM(6728)

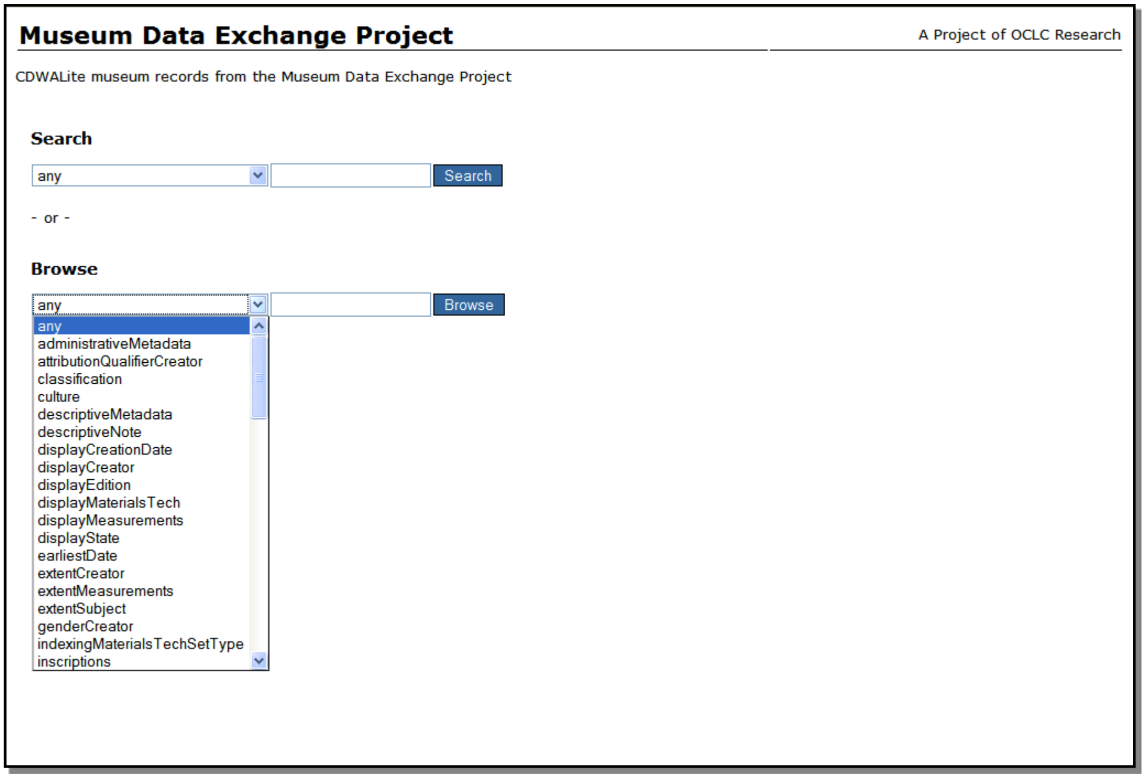
Figure 4. Excerpt of a report detailing all of units of the information containing a data value across the research aggregation

As already noted, for this project Pears and its reporting capabilities were tweaked and extended in many small as well as significant ways. For example, Ralph LeVan added a new application to the OCLC Research array of Pears tools which facilitated the comparison between museum source data values and controlled vocabularies. The entire process for comparing values to vocabularies was the following: initially, each Getty vocabulary (AAT, ULAN, TGN) was transformed into a Pears databases with an exposed SRW/U (Search/Retrieve via the Web or URL) interface. An application walked down the sorted list of data values in an institutional index, and compared them with the preferred and non-preferred terms in the controlled vocabulary, in both instances using the SRW/U interface to Pears as its conduit. The resulting report gave a count of the number of matching terms, as well as how many were found in the preferred and non-preferred indexes of the controlled vocabulary. A separate report listed the 100 most frequently occurring terms in the institutional database index, indicated whether they matched as a preferred or non-preferred term, and provided the term identifiers from the matching Getty vocabulary terms.

Another investigation during which Pears learned a new trick: evaluating the interconnectedness of descriptive terms across the nine contributors. Since the analysis aimed to appraise the utility of aggregating CDWA Lite records, OCLC Research wanted to establish which values used by one contributor could be found in other contributor's datasets. An SRU client walked down an index from one institution and used the words found in that index to search the aggregate database, looking for matches. The resulting report used the 100 most frequently occurring terms of each contributing institution and counted how many times these terms occurred in each of the remaining museum datasets.

Exposing the Research Aggregation to Participants

In addition to supporting data prep for analysis, Pears also provided a low-overhead mechanism for making individual databases as well as the research aggregation of all nine contributors available to project participants. By simply adding a stylesheet, the SRW/U enabled Pears turns into a database with searchable or browseable indexes (see Figure 5).



Museum Data Exchange Project

A Project of OCLC Research

CDWALite museum records from the Museum Data Exchange Project

Search

- or -

Browse

Figure 5. Screenshot of the no-frills search interface to the MDE research aggregation

Project participants had password-protected access to both the aggregate as well as the individual datasets.

Phase 3: Analysis of the Research Aggregation

Before OCLC Research started the data analysis process, museum participants formulated the questions which they would like to ask of their institutional and collective data.

Some of their questions about the institutional data sets were: Are the required CDWA Lite fields present? What is the state of compliance with CCO? Are the same terms used to consistently denote the same concepts? How are controlled vocabularies used? About the aggregate data set, museum participants wanted to know: Do queries across the research aggregation return meaningful results? How is my cataloging different from the other institution's cataloging? Which CDWA Lite fields are used by all institutions? How does the lack of subject data impact the research aggregation? For both the institutional data sets and the aggregate data set, participants evidently assumed that there would be room for improvement, because in either instance, they wanted to hear recommendations for how performance of the data could be enhanced.

These questions were formalized and expanded upon in a methodology[6] to guide the overall analysis efforts. This methodology grouped questions into two sections: a *Metrics* section which contained questions with objective, factual answers, and an *Evaluation* section which contained questions that are by their nature more subjective. The *Metrics* section asked questions about *Conformance* (does the data conform to what the applicable standards—CDWA Lite and CCO—stipulate?), as well as *Connections* (what overt relationships between records does the data support?). *Connections* questions in essence tried to triangulate the elusive question of interoperability. The section on *Evaluation* asked questions about *Suitability* (how well do the records support search, retrieval, aggregation?), as well as *Enhancement* (how can the suitability for search, retrieval, aggregation can be improved?).

The methodology was not intended to be a definitive checklist of all questions the project intended to plumb. It laid out the realm of possibilities, and allowed OCLC Research to discuss which questions it could tackle given expertise, available tools and time constraints. Most questions from the *Metrics* section lent themselves to machine analysis, and have been answered. OCLC Research itself predominantly worked on questions regarding CDWA Lite conformance, while an analysis of CCO compliance was outsourced to Patricia Harpring and Antonio Beecroft (see Patricia Harpring's CCO *Analysis* on page 38)Most questions pertaining to *Evaluation*, however, were beyond the reach of our project. Especially questions about *Suitability* require more foundational research until they are tractable to data analysis. Unless credible and deep data about search behaviors of museum data becomes available, any question about *Suitability* in turn begs the question: suitable in which context for whom to do what? As Jennifer Trant summarized in an introductory blog to a study of search logs at the Guggenheim Museum, "[W]e know almost nothing about what searchers of museum collections really do. [I] couldn't find a single serious [information retrieval] study in the museum domain." (Trant 2007). The *Searching Museum Collections* project, organized by Susan Chun (Consultant), Rob Stein (Indianapolis Museum of Art) and Christine Kuan (ARTstor), may provide some of the lacking datapoints: the project proposes to analyze search logs of museums and data aggregators, including ARTstor logs, to answer questions about user behavior (Searching Museum Collections n.d.).

Getting Familiar with the Data

Overall, a total of 887,572 records were contributed by nine museums in Phase 2 of the grant (as shown in Figure 6). Six out of nine museums contributed all accessioned objects in their database at the time of harvest. Of the remaining three, one chose a subset of all materials published on their Web site, while two made decisions based on the perceived state of cataloging. Among those providing subsets of data, the approximate percentages range from one third to a little over 50% of the data in their collections management system.

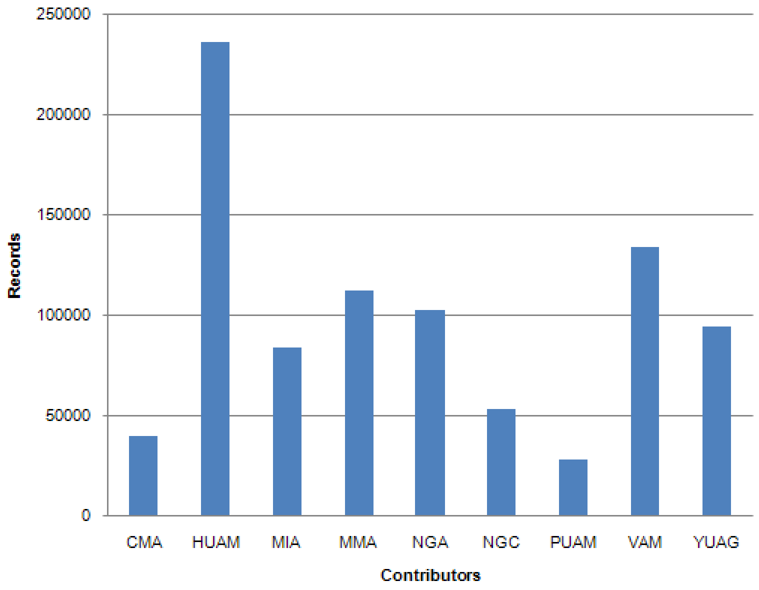


Figure 6. Records contributed by MDE participants

Figure 7 represents the elements and attributes found in the aggregation, placed in the context of all possible 131 CDWA Lite units of information.[7] It shows that few of the available units of information were consistently and widely used, some were used a little, and many were not used at all.

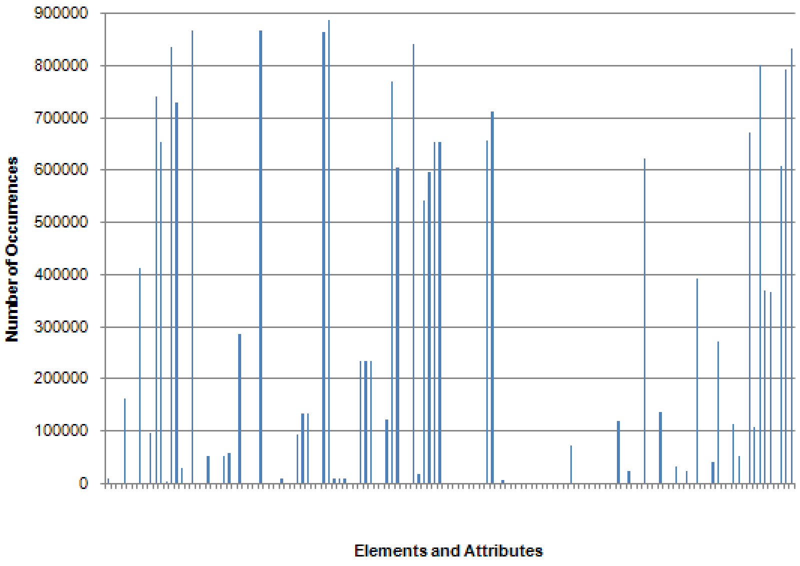


Figure 7. Use of CDWA Lite elements and attributes in the context of all possible units of information

Another take on the distribution of element/attribute use across the aggregation: Figure 8 shows the number of contributors that had made any use of a unit of information. A relatively small number (10 out of 131, or 7.6%) of elements/attributes are used at least once by all nine museums. These units of information are:

- *displayCreationDate* (CDWA Lite Required)
- *displayMaterialsTech* (CDWA Lite Required)
- *displayMeasurements* (CDWA Lite Highly Recommended)
- *earliestDate* (CDWA Lite Required)
- *latestDate* (CDWA Lite Required)
- *locationName* (CDWA Lite Required)
- “type” attribute on *locationName* (Attribute)
- *nameCreator* (CDWA Lite Required)
- *nationalityCreator* (CDWA Lite Highly Recommended)
- *title* (CDWA Lite Required)

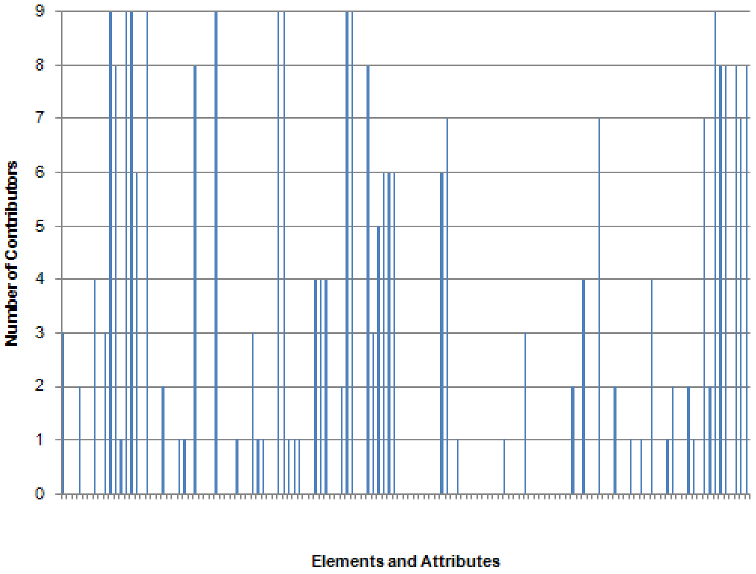


Figure 8. Use of possible CDWA Lite elements and attributes across contributing institutions, take 1

A final look at the distribution of use for the totality of all CDWA Lite elements/attributes (as shown in Figure 9) makes it easy to see how many units of information are not used at all (approximately 54%) and how many are used at least once by all contributors (approximately 8%).

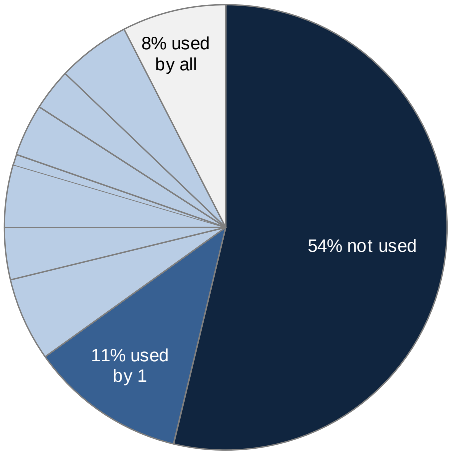


Figure 9. Use of possible CDWA Lite elements and attributes across contributing institutions, take 2

Conformance to CDWA Lite, Part 1: Cardinality

The CDWA Lite specification calls 12 data elements “required.” and 5 elements “highly recommended” (see Figure 10). The specification’s authors deem these elements particularly important for both the description of a piece of artwork as well as its indexing and retrieval. In theory, the required elements are mandatory for schema validation[8]—in practice, the vast majority of records which did not contain a data value in a required element still passed the schema validation test, since declaring the data element itself suffices for validation.

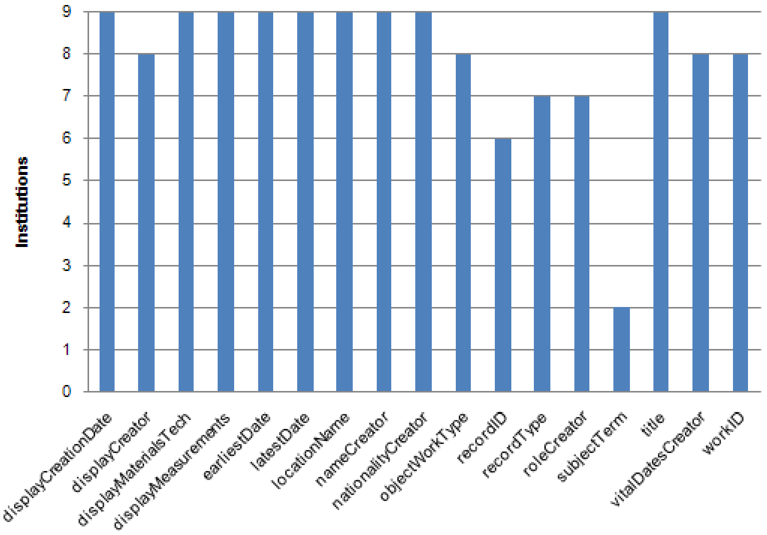


Figure 10. Any use of CDWA Lite required / highly recommended elements

As would be expected, use of CDWA Lite required / highly recommended elements shows a lot more density than use of the schema overall (see Figure 8). The counts shown in Figure 10 reflect the number of contributors who used these elements at least once. 9 out of 17 required / highly recommended elements are used by all contributors. A conspicuous outlier is *subjectTerm* (more on that later).

However, a graph of the institutions which provided values in these elements for *all* of their records gives a different view of how comprehensive these required / highly recommended elements were utilized. *locationName* emerges as the only element consistently present in all records across all nine contributors. A little over 50% (9 of the 17) required / highly recommended elements occur consistently in only three or less museum contributors. Almost 50% (8 of the 17) required / highly recommended elements occur consistently in five or more contributors.

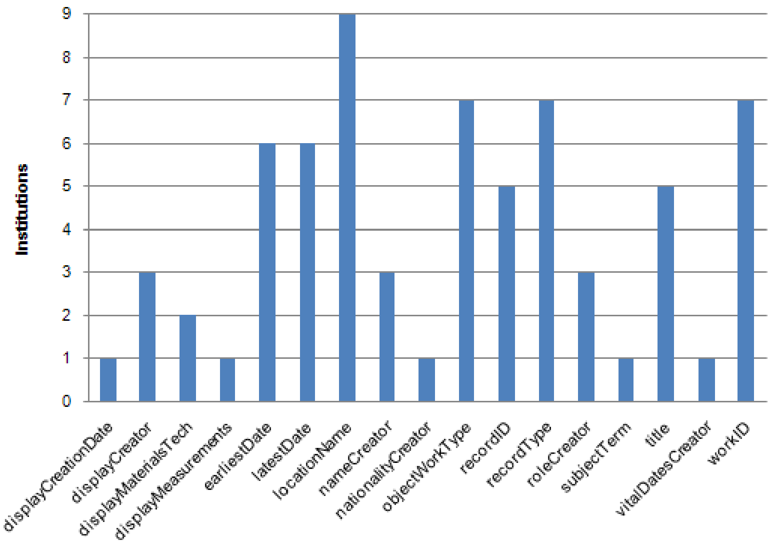


Figure 11. Any use of CDWA Lite required / highly recommended elements

The more realistic middle ground to the overly optimistic Figure 10 and the overly pessimistic Figure 11 is a graph of the percentage of records in the research aggregation that have values in the required / highly recommended elements (Figure 12). Discounting the outlier subjectTerm, the consistency with which these elements occur is greater than 65% overall. For 7 of 17 elements, consistency is above 90%.

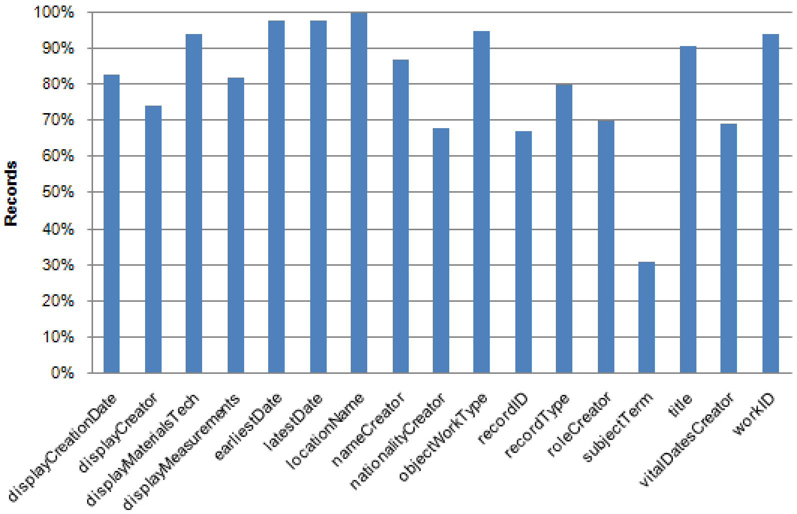


Figure 12. Use of CDWA Lite required / highly recommended elements by percentage

Excursion: the Default COBOAT Mapping

At this point, a short break from figures and a disclaimer about the data is in order. Project participants submitted data to the research aggregation as part of an abstract exercise. The project parameters made no demand of them other than to make CDWA Lite records available. Consequently, fields which may very well be present in source systems remained unpopulated in the submitted CDWA Lite records. At the point where OCLC Research accepted the data contribution because of the time constraints of the grant project, a real life aggregator may very well have gone back to negotiate for further data values deemed crucial to a specific service.

For the 6 of 9 contributors using COBOAT, the default mapping provided with the application heavily influenced their data contribution. However, this default mapping covered all elements which museum participants had provided to Cogapp as part of their TMS to CDWA Lite mapping documents, and in that way, the defaults do represent a consensus of which units of information the museums considered important or unimportant. The default COBOAT mapping uses 32 units of information of the 131 defined in the CDWA Lite schema. All CDWA Lite required/recommended elements and attributes are in the default mapping, except for subjectTerm. (subjectTerm did not appear on any of the mapping documents.)[9]

In hindsight, it would have been beneficial to consciously reflect on those choices as a group, and ponder the impact of the default mapping on the outcome of the data analysis. (This would have likely surfaced the absence of subjectTerm, and perhaps the lack of any termSource attributes to declare controlled vocabularies.) On the other hand, COBOAT participants did have the option of adding further data elements and attributes, which they made limited use of: four out of six COBOAT contributors used, in various combinations, 11 additional elements and attributes.

Conformance to CDWA Lite, Part 2: Controlled Vocabularies

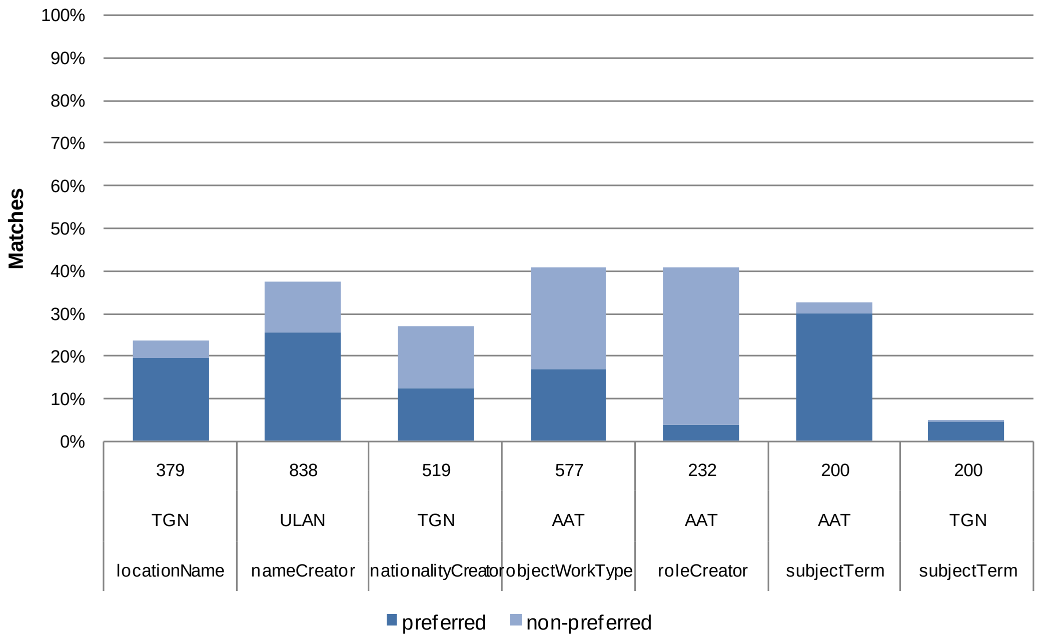
CDWA Lite, as well as its attendant data content standard CCO, recommends the use of controlled vocabularies for 13 data elements, six of which are required / highly recommended:

- objectWorkType
- nameCreator
- nationalityCreator
- roleCreator
- locationName
- subjectTerm

None of the contributing museums had marked the use of controlled vocabularies on any of the six elements in question. (As noted above, the "termSource" attribute was not part of the COBOAT default export). To get a sense of the deliberate or incidental use of values from controlled vocabularies, OCLC Research created a list of the top 100 most frequently used terms for each participating museum, deduplicated the lists (sometimes identical terms were frequently used at multiple museums), and then matched the remaining terms to a controlled vocabulary source recommended by CDWA Lite.

The data matching exploration highlights whether connections with an applicable thesaurus are possible without expertise-intensive and costly processing of data. Matches shown in Figure 13 are exact matches achieved without any manipulation of the source data, and pertain to the top 100 values for any given element from all contributing institutions. (The numbers along the x-axis, above the vocabulary acronym, give the total count of the top 100 values. For example, *objectWorkType* is represented by 577 top 100 deduplicated values from the eight institutions which contributed values for this element.) In some instances, higher match rates might have been achieved by post-processing, for example by splitting concatenated data values museums had contributed. Some values matched on multiple entries in their corresponding controlled vocabulary (more details below). Figure 13 includes the multi-matching values in the percentage counts.

For some of the data elements shown in Figure 13, the results of matching against controlled vocabularies is more indicative than for others. The issues encountered in matching values from museum contributors to controlled vocabularies were semantic mismatches (false hits), matches prevented by concatenated or deviantly structured data, and multiple matches. These caveats make all matches on TGN summarized in Figure 13 (*subjectTerms* in AAT and TGN; *nationalityCreator* in TGN; *locationName* in TGN) somewhat questionable. Reasonably indicative, however, are the matches of *nameCreator* in ULAN, *objectWorkType* in AAT, *roleCreator* in AAT.



(TGN = The Getty Thesaurus of Geographic Names®, ULAN = Union List of Artist Names®,
AAT = Art & Architecture Thesaurus®)

Figure 13. Match rate of required / highly recommended elements to applicable controlled vocabularies

What follows is a more in-depth discussion for each attempt at matching.

objectWorkType and AAT

- 8 out of 9 institutions contributed *objectWorkType* data values. The count for deduplicated top 100 *objectWorkTypes* is 577 across the eight contributing institutions. (As would be expected, for some institutions, the sum total of all their *objectWorkTypes* is less than 100).
- ***41% of objectWorkTypes (236 out of 577) match on an AAT term***, with 98 matching on a preferred term, and 138 matching on a non-preferred term. 8% of these 41% represent terms which match on more than one AAT term.

nameCreator and ULAN

- All nine institutions contributed nameCreator data values. The count for deduplicated top 100 nameCreators is 838 across the nine contributing institutions. ***37% of nameCreators (314 of the 838) match on a ULAN term***, with 213 matching on a preferred term, and 101 matching on a non-preferred term. 1% of these 37% represent terms which match on more than one ULAN term.
- A small disclaimer: for two contributors, large amounts of names did not match because they are inverted and miss a coma, such as Warhol Andy, Adams Ansel, Whistler James Mcneill, Saint Laurent Yves, etc. Had these names, which do exist in ULAN, matched, a higher overall match rate would have been the result.

roleCreator and AA

- 7 out of 9 institutions contributed *roleCreator* data values. The count for deduplicated top 100 *roleCreators* is 232 across the seven contributing institutions. (As would be expected, for most institutions, the sum total of all their *roleCreators* is less than 100).
- ***41% of roleCreators (95 of the 232) match on an AAT term***, with 9 matching on a preferred term, and 86 matching on a non-preferred term. 7% of these 41% represent terms which match on more than one AAT term.

subjectTerm and TGN, AAT

- *subjectTerm* was only used by two institutions, and therefore does not constitute a compelling sample. In addition, matching subject terms on TGN produced many hits which indeed were a letter-for-letter equivalent to the museum data, but where the intended semantic concept was not a place name: the *subjectTerm* "commerce", for example, matched on nine place names in TGN (inhabited places in Alabama, Georgia, Michigan, Mississippi, Missouri, Oklahoma and Tennessee), yet referred predominantly to a collection of photographs taken during the Great Depression with captions such as "Untitled (Sig. Klein Fat Men's Shop, 52 Third Avenue, New York City)".

locationName and TGN

- Ironically, data values which (at least in appearance) mimicked the hierarchical style of the thesaurus (such as "north America, american southwest, united states, new mexico, acoma pueblo") did not match on their entry in TGN. While they would provide a human user with unambiguous information about the place in question, for a machine match, "Acoma Pueblo" unadorned would have made the connection in our test. On the other hand, single values like "florence" often produced multiple hits: Florence, Italy, or which of the 42 inhabited places called "Florence" in the United States?

nationalityCreator and TGN

- In many instances, the data for nationalityCreator contained concatenated strings, such as "belgium, brussels, 18th century" or "american, born england," which could not be matched on TGN without further processing of the source data.

In summary, the vocabulary matching exercise indicates that in order to preserve the possibility of extending museum data with the rich information available in thesauri, even knowing the source thesaurus would have been only marginally helpful. Performing some data processing on the museum data could have created higher match rates. However, the value of using controlled vocabularies for search optimization or data enrichment can only be fully realized if the *termsourceID* is captured alongside *termSource* to establish a firm lock on the appropriate vocabulary term.

Economically Adding Value: Controlling More Terms

The high record count with which many individual data values on the top 100 lists are associated suggests opportunities for adding value to the data by controlling a relatively low number of terms with impact on a relatively high number of records per data set.

Consider the example for *objectWorkType* values in records from the Metropolitan Museum of Art depicted in Figure 14:

- The top 100 *objectWorkTypes* represent 99% of *objectWorkType* values in all 112,000 Metropolitan records.
- The top 100 *objectWorkTypes* matching on either a preferred or a non-preferred AAT term represent 27 matches, which is equal to 73% of all Metropolitan records. (8 of these 27 matches contain terms which match on more than one AAT entry.)
- The top 100 *objectWorkTypes* not matching on any AAT term represent 73, which is equal to 26% of all Metropolitan records.

In other words, **by tending to 73 *objectWorkType* values and disambiguating an additional 8, the Metropolitan could extend *objectWorkType* control to 99% of all 112,000 Metropolitan Museum records.**

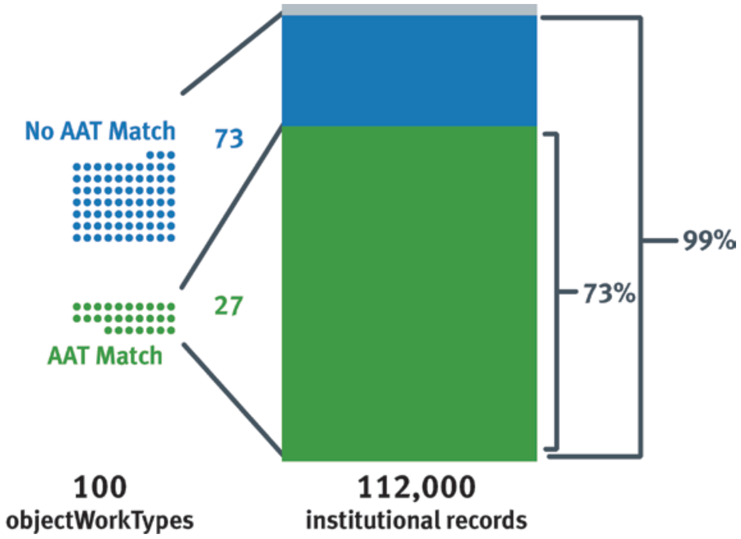


Figure 14. Top 100 *objectWorkTypes* and their corresponding records for the Metropolitan Museum of Art

These numbers by and large hold true for *objectWorkType* values across the aggregation:

- The top 100 *objectWorkTypes* for all 8 contributors combined represent 94% of *objectWorkTypes* in all 847,000 records.

- The top 100 *objectWorkTypes* for all 8 contributors combined matching on either a preferred or a non-preferred AAT term represent 236, which is equal to 64% of all aggregate records. (46 of these 236 matches contain terms which match on more than one AAT entry).
- The top 100 *objectWorkTypes* for all 8 contributors combined not matching on any *objectWorkType* term represent 341, which is equal to 30% of all aggregate records.

In other words, **by tending to 341 *objectWorkTypes* and disambiguating an additional 46, the aggregate collection control for *objectWorkType* could be extended to 94% of all 847,000 records.**

A data element like *objectWorkType* would be expected to produce favorable numbers in this type of analysis: by its very nature, *objectWorkType* contains a relatively low number of values which presumably reappear across many records in a collection. For *nameCreator*, a data element which has a relatively high number of values across aggregation records, one would expect a less impressive result from focusing on top 100 terms.

Consider the example depicted in Figure 15 from the Harvard Art Museum:

- The top 100 *nameCreators* represent 50% of *nameCreators* in all 236,000 Harvard records.
- The top 100 *nameCreators* matching on either a preferred or a non-preferred ULAN term represent 49 matches, which represent 25% of all Harvard records. (2 of the top 49 matches contain terms which match on more than one ULAN entry).
- The top 100 *nameCreators* not matching on any ULAN term represent 51, which represent 26% of all Harvard records.

In other words, **by tending to 51 *nameCreator* values and disambiguating an additional two, Harvard could extend *nameCreator* control to 50% of all 236,000 Harvard records.** While the overall percentages for *nameCreator* are necessarily lower than for *objectWorkType*, doubling the rate of control still constitutes a formidable result.

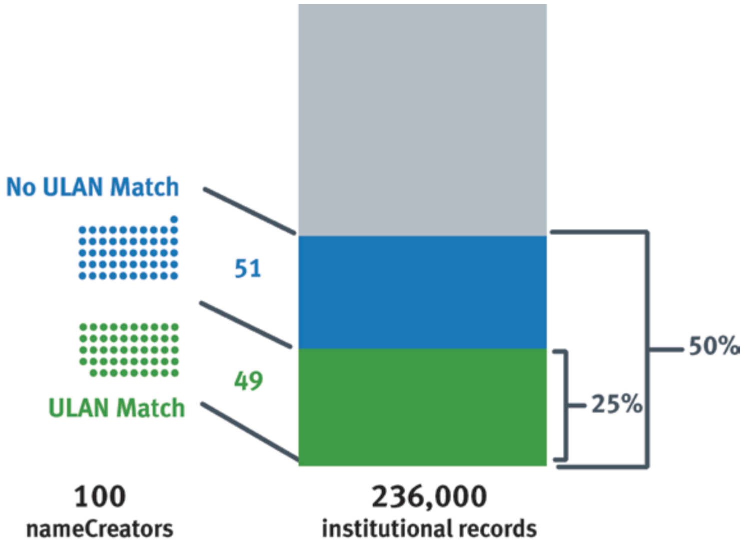


Figure 15. Top 100 *nameCreators* and their corresponding records for the Harvard Art Museum

Again, the comparison statistics across the aggregation:

- The top 100 *nameCreators* for all 9 contributors combined represent 49% of *nameCreators* in all 888,000 aggregation records.
- The top 100 *nameCreators* for all nine contributors combined matching on either a preferred or a non-preferred ULAN term represent 314, which represents 17% of all aggregate records. (Eight of these 314 matches contain terms which match on more than one ULAN entry.)
- The top 100 *nameCreators* for all nine contributors combined not matching on any *nameCreator* term represent 524, which is equal to 31% of all aggregate records.

In other words, **by tending to 524 *nameCreators* and disambiguating an additional eight, the aggregate collection control for *nameCreator* could be extended to 49% of all 888,000 aggregation records.**

Connections: Data Values Used Across the Aggregation

Apart from evaluating conformance to CDWA Lite and vocabulary use, OCLC Research at least dipped its toe into the murky waters of testing for interoperability. By asking questions about how consistently data values appeared across the nine contributors to the aggregation, some first impressions of cohesion can be triangulated. This investigation concentrated on a select number of data elements which are required / highly recommended by CDWA Lite, widely used by contributors and presumably of prominent use for searching and browse lists. For these elements, the top 100 values for each museum were cross-checked against other contributors to establish how many institutions use that same value, and with what frequency (i.e. in how many records).

Figure 16 provides a small sample from the resulting spreadsheets.

nameCreator

Value	Contributors	Records
parmigianino	8	783
raphael	8	1127
abbott, berenice	6	572
albers, josef	6	564
beuys, joseph	6	975
blake, william	6	1054
bonnard, pierre	6	623
bourne, samuel	6	806
brandt, bill	6	611
chagall, marc	6	2207

nationalityCreator

Value	Contributors	Records
american	9	248206
australian	9	1578
austrian	9	2443
belgian	9	1057
brazilian	9	221
british	9	50924
canadian	9	15776
chinese	9	2905
cuban	9	188
danish	9	591

roleCreator

Value	Contributors	Records
artist	6	475747
engraver	6	6465
printer	6	14009
publisher	6	25535
designer	5	19130
editor	5	1704
etcher	5	1822
lithographer	5	1496
painter	5	1376
architect	4	667

objectWorkType

Value	Contributors	Records
sculpture	8	10369
print	6	175588
photograph	6	86017
drawing	6	58140
painting	6	15837
furniture	6	2206
book	5	15644
paintings	5	9612
ceramic	5	5562
textiles	5	3898
textile	5	1775
portfolio	5	861
calligraphy	5	822
glass	5	800
manuscript	5	534
poster	4	3493
metalwork	4	2826
plate	4	1851
costume	4	956
album	4	823
wallpaper	4	500
sketchbook	4	356
frame	4	297
collage	4	132
mosaic	4	116
prints	3	8814
negative	3	7235
photographs	3	7081
jewelry	3	3662
vase	3	2663
dish	3	2440
bowl	3	2328
tile	3	1707
ring	3	1279
medal	3	1277
jug	3	1271

Figure 16. Most widely shared values across the aggregation for *nameCreator, nationalityCreator, roleCreator* and *objectWorkType*

With a small amount of additional processing, the spreadsheets underlying these figures allow statements about how many values in a specific data element are shared across how many institutions, and how many records these elements represent. Figures 17 and 18 explore what can be learned about the Aggregates cohesiveness by looking at *nationalityCreator* and *objectWorkType*.

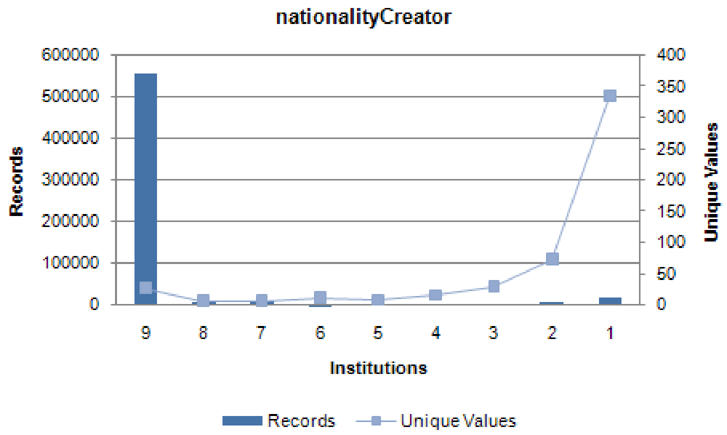


Figure 17. ***nationalityCreator*****: relating records, institutions and unique values**

For *nationalityCreator*, the distribution of unique values and associated records across the nine participating institutions matches what one would expect from browsing the data, given the preponderance of *nationalityCreator* values from a small set of countries. A relatively small number of unique values (28) are present in the data from all 9 participants, and correlate to a large number of records (554K). Additionally, one would expect to see many unique values represented in the data shared by one or two institutions, with correspondingly low numbers of associated records, for those nationalities that are less common. Sure enough, 408 *nationalityCreator* values are held by any two or a single institution, representing 26K records, or 2.9% of the entire aggregate. Given this appraisal, *nationalityCreator* values seem to form a coherent set of data values.

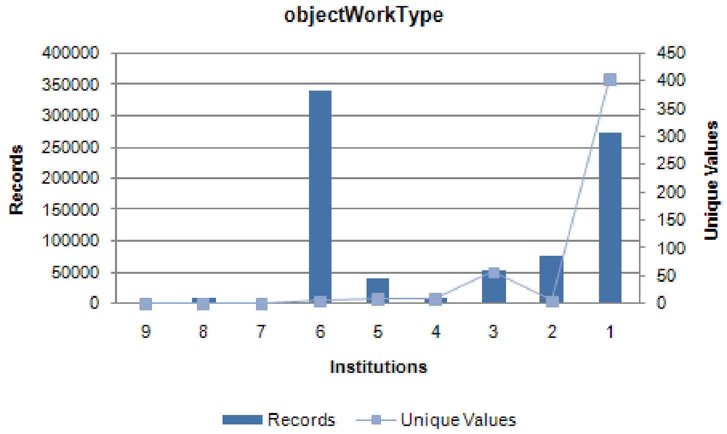


Figure 18. ***objectWorkType*****: relating records, institutions and unique values**

For *objectWorkType*, the distribution of unique values and associated records across the nine participating institutions show a less coherent mix. (In part, this can be attributed to small differences in values preventing a match, such as the singular and plural forms for a term evident for *objectWorkType* in Figure 18: print(s), photograph(s), painting(s), textile(s), etc.) Since only eight institutions contributed *objectWorkType* data, no unique *objectWorkType* values are represented across all contributors. Only one value is found in the data of eight contributors. The first spike in the graph occurs at five values found in the data of six museums. Though those five values account for a significant number of records (338K, or 38% of the Aggregate), one would have expected the number for both widely shared values and corresponding record counts to be higher. Moving on to the spike at values present in a single museum's data, it is hard to conceive that the high number of unique values (404) representing a high number of records (273K, or 30% of the Aggregate) accurately reflects the underlying collections. The large number of unique *objectWorkTypes* suggests an opportunity to reduce noise in the data via more rigorous application of controlled vocabulary (according to Figure 13, the current match rate to AAT is 41%), which would produce a set of unique values that could be more sensible to browse.

Enhancement: Automated Creation of Semantic Metadata Using OpenCalais™

The analysis project made a small foray into exploring automatic ways of enhancing the museum records by exposing a few select records from the MDE aggregation to the OpenCalais Web Service (Thomson Reuters n.d.). The OpenCalais Web Service processes text into semantic metadata, i.e. it locates entities (people, places, products, etc.), facts (John Doe works for Acme Corporation) and events (Jane Doe was appointed as a Board member of Acme Corporation). As the examples in parenthesis, which come from the OpenCalais FAQ, indicate, the Web Service is mainly oriented towards commercial data, but cultural institutions like the PowerHouse Museum (Chan 2008) have also explored its potential.

The results of applying OpenCalais to select MDE records suggest that especially records with unstructured narrative description will benefit, sometimes quite significantly, but even less completely described records can benefit by adding more semantic value to certain elements (e.g., parsing a location name into city, state/province, and country names) and finding additional names for groups, people, and events from within notes and other CDWA Lite elements. OpenCalais also showed surprising skill at transforming certain strings into categories and values. For example, it was able to generate the category and value "Currency:pence" from the string "this chair which cost 16/6 (82p)."

Here are CDWA Lite access points for an MDE record with a moderate level of structured text, and no unstructured or narrative description:

```
objectWorkType: Photograph
title: Claude Monet
nameCreator: Larchman, Harry
roleCreator: Artist
nameCreator: Monet, Claude
roleCreator: Portrait sitter
earliestDate: 1900
latestDate: 1909
locationName: The Metropolitan Museum of Art, New York, NY, USA

Photograph; Claude Monet; Larchman, Harry; Artist; Monet, Claude; Portrait
sitter; 1900; 1909; The Metropolitan Museum of Art, New York, NY, USA
```

. . . it finds the following matching categories and values:

```
City: Art
Country: United States
Facility: The Metropolitan Museum of Art
Person: Claude Monet
Position: Artist
ProvinceOrState: New York, United States
```

OpenCalais also returns a confidence level for these assertions, not shown here, that could help a system demote the “city of Art” it has identified.

But if the MDE terms are plugged into a template that emulates wall label text, for example:

In this *photograph* created during the years *1900-1909*, the *artist Harry Larchman* has portrayed the subject *Claude Monet*. The *photograph* is in the collection of the *Metropolitan Museum of Art, New York, NY, USA*.

. . . then OpenCalais finds more concepts:

```
City: Art
Country: United States
Facility: Metropolitan Museum of Art
Person: Claude Monet
Person: Harry Larchman
Position: artist
Province or State: New York, United States
Generic Relations: portray, Harry Larchman, Claude Monet
Person Career: Harry Larchman, artist, professional, current
```

This relatively simple step of providing a template of narrative structure around specific data element values from the CDWA Lite record source may be an important consideration in any projects that attempt to further enrich or extend the data using tools that look for semantic value within sentence structure, such as OpenCalais.

A Note About Record Identifiers

Persistent and unique record identifiers are essential for supporting linking and retrieval, as well as data management of records across systems. Without a reliable identifier it is difficult or impossible to match incoming records for adds, updates, and deletes, or to link to a specific record. In other words, a reliable identifier unlocks one of the chief benefits of using OAI-PMH for data sharing, the ability to keep a data contribution to a third party current with changes in the local museum system. There are four places in the OAI-PMH and CDWA Lite data where unique record identifiers may be supplied:

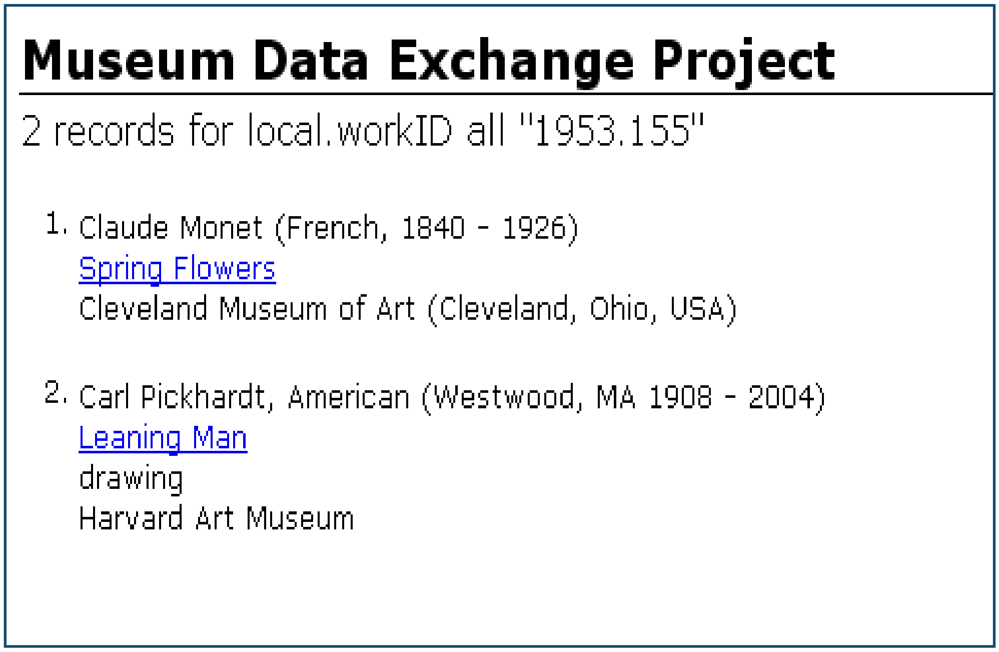
- *recordID*: CDWA Lite Required, CCO Required
  “A unique record identification in the contributor’s (local) system”[10]
- *recordInfoID*: Not required or recommended
  “Unique ID of the metadata. Record Info ID has the same definition as Record ID but out of the context of original local system, such as a persistent identifier or an oai identifier ”

- *workID*: CDWA Lite Highly Recommended, CCO Required
  "Any unique numeric or alphanumeric identifier(s) assigned to a work by a repository"
- OAI-PMH Identifier
  OAI Required. Schema, repository, and item id. E.g., oai:artsmia.org:10507

Among the data contributors, six of nine used *recordID* and *recordInfoID*, eight of nine used *workID*, some used both. (Both were defined in the default COBOAT mapping.) All contributors used at least one of the CDWA Lite identifiers.

The OAI-PMH identifier is required by the protocol, and when available, can help disambiguate record identifiers that would otherwise not be unique across repositories. As noted earlier, six ofnine museums used OAI to contribute data. If relied upon, the OAI identifier needs to be static (changing repository IDs adds volatility) and available along with the CDWA Lite data in whatever system incorporates the data.

When identifiers are supplied that are not unique across an aggregation, disambiguation problems develop. As depicted in Figure 19, the identifier "1953.155" is used by two different contributors for two different works.



Museum Data Exchange Project

2 records for local.workID all "1953.155"

1. Claude Monet (French, 1840 - 1926)
   Spring Flowers
   Cleveland Museum of Art (Cleveland, Ohio, USA)
2. Carl Pickhardt, American (Westwood, MA 1908 - 2004)
   Leaning Man
   drawing
   Harvard Art Museum

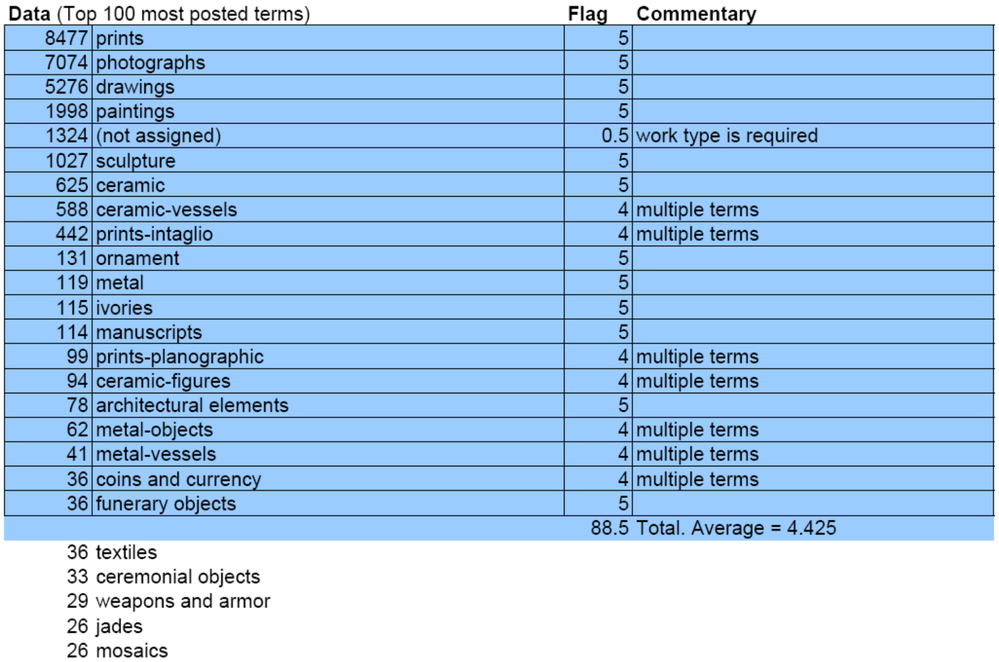
Figure 19. Screenshot of a search result from the research aggregation

From a data aggregator's point of view, the key concerns is not which data element is used to provide an identifier, but that the same element be used consistently by all contributors. Given its required nature, the *recordID* element seems like a good candidate for consensus. Aggregators can supplement it with information about the contributor to make it unique across the collection (e.g. by using the OAI identifier or following its conventions in case not all contributors use OAI), which will help ensure efficient and reliable record processing and retrieval.

Patricia Harpring's CCO Analysis

The OCLC Research data analysis largely focused on testing for conformance against stipulations made by the CDWA Lite specification. The rules outlined in CDWA Lite conform to the much more comprehensive set of guidelines laid out by its companion data content standard CCO, as well as the full CDWA online (Getty Trust, J. Paul 2009). However, more rigorous evaluation of values supplied by contributors against CCO did not lend itself to the kind of machine-processing matching up with OCLC Research's skills, and called for a deep familiarity with the data content standard. OCLC Research contracted with CCO co-author Patricia Harpring, supported by Antonio Beecroft (both from the Getty Research Institute), to spend some of their weekend and vacation time to evaluate the CCO-ness of the data.

As a basis for this additional analysis, OCLC Research provided Patricia with Pears reports for the 20 data elements which either CDWA Lite or CCO mark as required / highly recommended. Each of the elements was represented by its top 100 most frequently occurring values. Patricia drew up scoring principles for the spreadsheets, which she and Antonio used to evaluate the top 20 values for each element from the 9 contributors (see Figure 20).



Data (Top 100 most posted terms)		**Flag**	**Commentary**
8477	prints	5	
7074	photographs	5	
5276	drawings	5	
1998	paintings	5	
1324	(not assigned)	0.5	work type is required
1027	sculpture	5	
625	ceramic	5	
588	ceramic-vessels	4	multiple terms
442	prints-intaglio	4	multiple terms
131	ornament	5	
119	metal	5	
115	ivories	5	
114	manuscripts	5	
99	prints-planographic	4	multiple terms
94	ceramic-figures	4	multiple terms
78	architectural elements	5	
62	metal-objects	4	multiple terms
41	metal-vessels	4	multiple terms
36	coins and currency	4	multiple terms
36	funerary objects	5	
		88.5	Total. Average = 4.425
36	textiles		
33	ceremonial objects		
29	weapons and armor		
26	jades		
26	mosaics		

Figure 20. ***objectWorkType*** **spreadsheet (excerpt) for CCO analysis, including evaluation comments**

Patricia and Roberto subtracted from an institution's score in particular for missing data, multiple terms in a single element, data mismatches (i.e. *roleCreator* containing *attributionQualifier* information), uncontrolled terms, as well as the *title* "Untitled" (CCO requires a descriptive title) and the *displayCreationDate* "Undated" (CCO requires an approximate date).

Figure 21 provides the overall scoring pattern for the core elements under examination, and gives the impression that overall, the MDE contributor's exhibited considerable conformance to CCO. In Patricia's own words:

> "The nine sets of data analyzed for compliance in this study scored quite well. Many of the points deducted in scoring were due to mapping and parsing issues that could be easily corrected. The most frequent issues concerned having multiple terms in one field or missing data that is 1) probably actually available in the institution's local data base (e.g., a missing Work ID) or 2) could be filled using suggested default values (e.g., globally supply "artist" or "maker" for missing Creator Role)." (Harpring 2009)

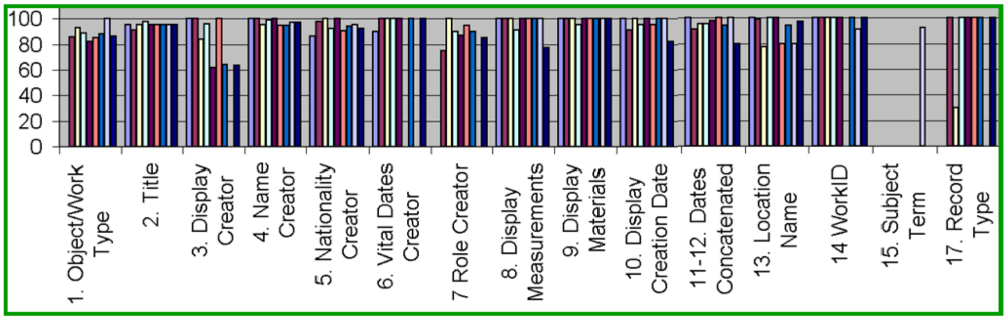


Figure 21. Overall scores from CCO evaluation—each bar represents a museum

More details on the scoring criteria itself, as well as a brief discussion of analysis results, can be found in Patricia's document, "Museum Data Exchange CCO Evaluation—Criteria for Scoring" (Harpring 2009).

Third Party Data Analysis

The agreements with participating museums described in the section "Legal agreements" include a provision which allows a third party researcher to take possession of the data under the terms of the original letters of agreement, analyze the data, and publish findings. From the outset, OCLC Research realized that it could contribute some knowledge about the characteristics of the data, but that other entities with different tools, interests, and (perhaps) contextualizing data could bring additional findings to light. Initially, some of the museums themselves had also expressed an interest in taking a methodical look at the aggregate data. (See OCLC Research n.d.d. for more information on third party analysis.)

Compelling Applications for Data Exchange Capacity

The design of the MDE project allowed participants to experiment and gain experience with sharing data without necessarily having settled the policy issues the museum community at large is still grappling with. While this approach by and large succeeded (witness the creation of tools, the sharing of data, the lessons in the data analysis), the absence of real-life requirements, a real-life audience and real-life applications for the data made it difficult for the museums to calibrate their data for submission, and for OCLC Research to evaluate it for suitability.

To survive and thrive, museum data sharing at participating institutions will need to outgrow the sandbox, become sanctioned by policies, and applied in service of goals supporting the museum mission. Needless to say, sharing data is not a goal unto itself, but an activity which needs to drive a process or application of genuine interest to an individual museum. As a natural by-product of the grant work, participants and project followers surfaced and discussed potential applications for CDWA Lite / OAI-PMH in a museum setting, and the project invited representatives from OMEKA, ARTstor, ArtsConnectEd and CHIN, as well as Gallery Systems and Selago Design, to participate in our final meeting at the Metropolitan Museum of Art on July 27, 2009. In the meantime, some of the museums have already put their new sharing infrastructure to work in production settings; others are actively exploring their options. Below are brief sketches of the institutional goals a CDWA Lite / OAI-PMH capacity could support.

Goal: Create an exhibition Web site, or publish an entire collection online

- OMEKA is a free, open source Web-based publishing platform for collections created by the Center for History and New Media, George Mason University. For a museum to take advantage of this tool, a core set of data has to migrate from the local collections management system into the OMEKA platform. A museum can use COBOAT and OAICatMuseum to create an OAI data content provider, and OMEKA's OAI-PMH Harvester plug-in (George Mason University n.d.) to ingest and update the data. At least one of the MDE project participants supported the test of this plug-in by providing access to their OAI-PMH installations, and others may follow.

Goal: Add a tagging feature to your online collection

- *Steve: The museum social tagging project* is a collaboration of museum professionals exploring the benefits of social tagging for cultural collections. As part of its research, Steve has created software for a hosted tagging solution. For a museum to take advantage of this tool, its data will have to be loaded into the tagging application. The Steve tagger can harvest OAI-PMH repositories of data, and accepts data in both CDWA Lite and Dublin Core. The project team envisions that a future local version of the tagger will contain the same ingest functionality.

Goal: Disseminate authoritative descriptive records of museum objects

- *The Cultural Objects Name Authority* (CONA™), a new Getty vocabulary of brief authoritative records for works of art and architecture, is slated to be available for contributions in 2011. For museums, contribution to CONA ensures that records of their works as represented in visual resources or library collections are authoritative. Although CONA is an authority, not a full-blown database of object information, it complies with the cataloging rules for adequate minimal records described in CDWA and CCO. As Patricia Harpring outlined during our final meeting, the vocabulary editorial team will accept contributions in CDWA Lite XML or in the larger CONA contribution XML format.

Goal: Expose collections for K-12 educators, students and scholars

- *ArtsConnectEd* is an interactive Web site that provides access to works of art and educational resources from the Minneapolis Institute of Arts and the Walker Art Center. The Institute and the Walker pool their resources using a CDWA Lite / OAI-PMH infrastructure, and the Institute of Arts has successfully implemented the MDE tools to contribute to this aggregation. At the final face-to-face MDE meeting, both ArtsConnectEd representatives and grant participants speculated about whether the resource could grow to include additional contributors. Robin Dowden (Walker) demonstrated a private research prototype site including the MDE datasets from the National Gallery of Canada and the Harvard Art Museum, as well as records from the Brooklyn Museum (6K) accessed via an API. These three new datasets had been loaded into ArtsConnectEd within 72 hours by Nate Solas (Walker), and even without any smoothing around the edges, the five-institution version of ArtsConnectEd provided a solid experience: "CDWA Lite format and indexing is very good at first glance," as Robin observed.

Goal: Expose collections to higher education

- As one of the original co-creators of CDWA Lite, *ARTstor* welcomes contributions in CDWA Lite. During our final face-to-face project meeting, Bill Ying and Christine Kuan (both ARTstor) emphasized that ARTstor is eager to create relationships with data contributors in which repeat OAI harvesting to support updating and adding to the data becomes a matter of routine. ARTstor currently counts 80 international museums among its contributor, yet only a very small minority of them have contributed data via OAI. As an outcome of the MDE project, both the Harvard Art Museum and the Princeton University Art Museum are actively exploring OAI harvesting with ARTstor, while three additional participants have signaled that this would be a likely use for their OAI infrastructure as well.

Goal: Effectively expose the collective collection of a university campus (collections from libraries, archives and museums)

- At *Yale University*, both the Yale University Art Gallery (grant participant) and the Yale Center for British Art (an avid follower of the project) are implementing COBOAT / OAICatMuseum with the goal of using this capacity for a variety of university-wide initiatives. The museums contribute data to a cross-collection search effort via OAI-PMH (Princeton has similar ambitions), and the Yale museums will also use the same set-up to sync data with OpenText's Artesia, a university-wide digital assets management system. In addition, the Art Gallery proposes to use CDWA Lite XML records to share data with the recipients of traveling collections.

Goal: Aggregate collection data for national projects

- The *Canadian Heritage Information Network* (CHIN) is currently redeveloping Artefacts Canada (http://www.chin.gc.ca/English/Artefacts_Canada/), a resource of more than 3 million object records and 580,000 images from hundreds of museums across the country. So far, the resource grows largely via contributions of tab-delimited files and spreadsheets representing museum data, and CHIN would like to explore other mechanisms for aggregation. Corina MacDonald (CHIN), who attended the MDE final project meeting, speculated that a test bed of large Canadian institutions using COBOAT / OAICatMuseum might provide lessons for a way forward.
- Another example of a national aggregation project, this time from the UK: A venture of the *Collections Trust*, the Museums, Libraries and Archives Council (MLA), the European Commission and technical partners Knowledge Integration Ltd, Culture Grid pulls together information from UK library, archive and museum databases, and then opens up this content to media partners such as Google and the BBC to ensure that it is available to as wide an audience as possible (Collections Trust n.d.). To drive data into the Culture Grid, The

Collections Trust has created an SDK for collections management system providers, which allows them to easily integrate functionality for offering up structured DC via an OAI-PMH data content provider. While this effort is not built around CDWA Lite XML, the general strategy of opening up collections by providing a low-barrier export mechanism is parallel and complimentary to the MDE work.

Conclusion: Policy Challenges Remain

In his insightful article "Digital Assets and Digital Burdens: Obstacles to the Dream of Universal Access," already cited in the introduction, Nicholas Crofts (2008, 2) provides a list of false premises for data sharing:

i. Adapting to new technology is the major obstacle to achieving universal access
ii. The corpus of existing digital documentation is suitable for wide-scale diffusion
iii. Memory institutions want to make their digital materials freely available

Ironically, these premises can be viewed as structuring the MDE project. The MDE grant posited that a joint investment in shareable tools (cf. i) might help force the policy question of how openly to disseminate data (cf. iii), while also allowing an investigation into how suitable museum descriptions are for aggregation (cf. ii). At the end of the day, however, there is no disagreement with Crofts position: ultimately, policy decisions allow data sharing technology to be harnessed, or create the impetus to upgrade descriptive records.

In the case of the MDE museums, all of them had enough institutional will towards data sharing to participate in this project. As a result of this project, some have already used their new capacity for data exchange to drive mission-critical projects. A quick recap of the most significant developments catalyzed by the MDE tools: the Minneapolis Institute of Arts uses the MDE tools to contribute data to ArtsConnected; Yale University Art Museum and the Yale Center for British Art use the tools to share data with a campus-wide cross-search, and contribute to a central digital asset management system; the Harvard Art Museum and the Princeton University Art Museum are actively exploring OAI harvesting with ARTstor, while three additional participants have signaled that this would be a likely use for their OAI infrastructure as well. Obviously, it is too early to judge the ultimate impact of making the MDE suite of tools available, yet these developments are promising.

While the data analysis efforts detailed in this paper cannot be viewed as a conclusive measure for the fitness of museum descriptions, they ultimately leave a positive impression: the analysis shows good adherence to applicable standards, as well as reasonable cohesion. Where there is room for improvement, some fairly straightforward remedies can be employed. Significant improvements in the aggregation could be achieved by revisiting data mappings to allow for a more complete representation of the underlying museum data. Focusing on the top 100 most highly occurring values for key elements will impact a high number of corresponding records, and would be low-hanging fruit for data clean-up activities. Museums engaging in data exchange will learn new ways to adapt and improve their data output every time they share, and the MDE experiment was just the first step on that journey.

At the end of the day, the willingness of museums to share data more widely is tied to the compelling application for that shared data. When there are applications for sharing data which directly support the museum mission, more data is shared. When more data is shared, more such compelling applications emerge. This chicken-and-egg conundrum provides a challenge to both museum policy makers as well as those wishing to aggregate data. The list of aggregators, platforms, projects and products provided in the previous chapter which support data exchange using CDWA Lite / OAI provides hope that these compelling applications will move museum policy discussions forward.

In the summation of his paper, Nicholas Crofts lays out what is at stake:

> "[O]ther organisations and individuals are actively engaged in producing attractive digital content and making it widely available. Universal access to cultural heritage will likely soon become a reality, but museums may be losing their role as key players." (Crofts 2008, 13)

No matter which museum you represent, a search on Flickr® for your institution's name viscerally confirms the validity of this prediction.

It seems appropriate to close this paper with the words of a man who has fought this same policy battle 40 years ago. While the 1960s were a different time indeed, the arguments sound quite familiar. In an oral history interview from 2004, here is how Everett Ellin remembers making his case for the digital museum and shared data:

> "So museums should know how to reduce all records, all registrar records, records of accessions, to a digital file, and each file is kept in an archive, a digital archive, and we tie all the archives together by a computer network. We take all these archives and we link them up, and then when a technology comes that I know is certain that will let you take reasonably good photos digitally, then we will make digital files of those photos and we will put that in a separate part of the same archive—I have that language from 1966 in print—and we will begin to get into the electronic age. And we or you will be ready for the day when you see what I mean about that you are a medium, and that you have to stand toe to toe with mass media, because it's going to be a battle of images inevitably— inevitably." (Kirwin, 2004)

Appendix A. Project Participants

The following individuals have played a major role in the success of the Museum Data Exchange project by contributing their expertise, perspective and time.

Grant funded museum participants:
Andrea Bour, Doug Hiwiller, Holly Witchey (Cleveland Museum of Art)
Jeff Steward (Harvard Art Museum)
Piotr Adamczyk, Michael Jenkins, Shyam Oberoi (Metropolitan Museum of Art)
Peter Dueker (National Gallery of Art)
Cathryn Goodwin (Princeton University Art Museum)
Alexander Macfie, Alan Seal (Victoria & Albert Museum)
Ariana French, Thomas Raich, Tim Speevack (Yale University Art Gallery)

Additional museum participants:
Andrew David, Michael Dust, Jim Ockuly (Minneapolis Institute of Arts)
Sonya Dumais, Greg Spurgeon (National Gallery of Canada)

Additional contributors:
Christine Kuan, William Ying (ARTstor)
Corina MacDonald, Anne-Marie Millner (Canadian Heritage Information Network)
James Safley, Tom Scheinfeldt (Center for History and New Media, George Mason University)
Nick Poole (Collections Trust)
Patricia Harpring and Antonio Beecroft (Consultants)
Robb Detlefs (Gallery Systems)
Scott Sayre (Sandbox Studios)
Gayle Silverman, James Starrit (Selago Design)
Robin Dowden, Nate Solas (Walker Art Center)

Cogapp
Ben Rubinstein, Stephen Norris, Mat Walker

OCLC Research
Ralph LeVan, Günter Waibel, Bruce Washburn, Jeff Young

Appendix B. Outputs of the Museum Data Exchange Activity

Tools

COBOAT
http://www.oclc.org/research/activities/coboat/

OAICatMuseum 1.0
http://www.oclc.org/research/activities/oaicatmuseum/

Connecting with other users of COBOAT and OAICatMuseum, and exchanging COBOAT application profiles for different databases:
http://sites.google.com/site/museumdataexchange/

Documents

Patricia Harpring - Criteria for Scoring (CCO Evaluation)
A document which outlines general findings from Patricia Harpring's CCO analysis, and her methodology for evaluating CDWA Lite against CCO.
http://www.oclc.org/research/activities/museumdata/scoring-criteria.pdf

MDE Analysis Methodology
A document which outlines the array of possible analysis questions the project surfaced.
http://www.oclc.org/research/activities/museumdata/methodology.pdf

CDWA Light, CCO, COBOAT mapping
A spreadsheet listing all content-bearing data elements and attributes defined by CDWA Lite, plus mappings to CCO. The document also indicates which of these data elements are part of the COBOAT default mapping.
http://www.oclc.org/research/activities/museumdata/mapping.xls

All of these documents are available from
http://www.oclc.org/research/activities/museumdata/default.htm

Bibliography

Chan, Sebastian. 2008. "OPAC2.0 – OpenCalais meets our museum collection / auto-tagging and semantic parsing of collection data." *Fresh + New(er)* (31 March). A Powerhouse Museum blog. http://www.powerhousemuseum.com/dmsblog/index.php/2008/03/31/opac20-opencalais-meets-our-museum-collection-auto-tagging-and-semantic-parsing-of-collection-data/.

Crofts, Nicholas. 2008. *Digital assets and digital burdens: obstacles to the dream of universal access.* Paper presented at the annual conference of the International Documentation Committee of the International Council of Museums (CIDOC), September 15-18, in Athens, Greece. http://cidoc.mediahost.org/content/archive/cidoc2008/Documents/papers/drfile.2008-06-72.pdf. [Conference program available from: http://cidoc.mediahost.org/content/archive/cidoc2008/EN/site/Home/t_section.html.]

Collections Trust. n.d. "Culture Grid." http://www.collectionstrust.org.uk/culturegrid.

Dowden, R., and S. Sayre. 2009. "Tear Down the Walls: The Redesign of ArtsConnectEd." In J. Trant and D. Bearman (eds). *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31. http://www.archimuse.com/mw2009/papers/dowden/dowden.html.

Elings, M.W. and Günter Waibel. 2007. Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums. *First Monday* 12,3. http://firstmonday.org/issues/issue12_3/elings/index.html or http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1628/1543.

Ellin, Everett. 1968. An international survey of museum computer activity. *Computers and the Humanities* 3,2: 65-86.

George Mason University. n.d. "Plugins/OaipmhHarvester." *Omeka.* Center for History and New Media. http://omeka.org/codex/Plugins/OaipmhHarvester.

Getty Trust, J. Paul. 2009. *Categories for the Description of Works of Art*. Ed. Murtha Baca and Patricia Harpring (rev. 9 June by Patricia Harpring). http://www.getty.edu/research/conducting_research/standards/cdwa/.

———. n.d. *Categories for the Description of Works of Art Lite*. Getty Research Institute. http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html.

Getty Trust, J. Paul, and College Art Association. 2006. *CDWA Lite: Specification for an XML Schema for Contributing Records via the OAI Harvesting Protocol*. http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.pdf.

Harpring, Patricia. 2009. Museum Data Exchange: CCO Evaluation Criteria for Scoring. OCLC. http://www.oclc.org/research/activities/museumdata/scoring-criteria.pdf.

IBM Federal Systems Division and Everett Ellin. 1969. *An Information System for American Museums: a report prepared for the Museum Computer Network*. Gaithersburg MD: International Business Machines Corporation. Smithsonian Institution Archives, RU 7432 / Box 19

Kirwin, Liza. Oral history interview with Everett Ellin, 2004 Apr. 27-28, Archives of American Art. Smithsonian Institution. http://aaa.si.edu/collections/oralhistories/transcripts/ellin04.htm.

Minneapolis Institute of Arts and Walker Art Center. n.d. ArtsConnectEd. http://www.artsconnected.org/.

Misunas, Marla and Richard Urban. A Brief History of the Museum Computer Network. Museum Computer Network. http://www.mcn.edu/about/index.asp?subkey=1942.

New Digital Group, Inc. n.d. Smarty: Template Engine. http://www.smarty.net/copyright.php.

OCLC Research. n.d.a. CDWA Lite, CCO, COBOAT Mapping. An output of the Museum Data Exchange activity. http://www.oclc.org/research/activities/museumdata/mapping.xls.

———. n.d.b. Metadata Schema Transformation Services. http://www.oclc.org/research/activities/schematrans/default.htm.

———. n.d.c. Museum Collections Sharing Group. http://www.oclc.org/research/activities/museumdata/museumcollwg.htm.

———. n.d.d. Museum Data Exchange. http://www.oclc.org/research/activities/museumdata/default.htm.

Parry, Ross. 2007. *Recoding the museum: digital heritage and the technologies of change*. Museum meanings. London: Routledge.

Searching Museum Collections. n.d. Searching Museum Collections: A Research Project. https://museumsearch.pbworks.com/.

OpenSiteSearch Community. n.d. OpenSiteSearch. http://opensitesearch.sourceforge.net/.

Thomson Reuters. n.d. OpenCalais. http://www.opencalais.com/.

Trant, Jennifer. 2007. “Searching Museum Collections on-line – what do people really do?” jtrant’s blog. *Archives & Museum Informatics* (January 1). http://conference.archimuse.com/node/7424.

Notes

[1] The original members of the committee: Nancy Allen (ARTstor), Erin Coburn (J. Paul Getty Museum), Ken Hamma (Getty Trust), Michael Jenkins (Metropolitan Museum of Art), Nick Pool (MDA), Jenn Riley (Indiana University), Günter Waibel (OCLC Research)

[2] Grant funds were used exclusively to off-set museum costs; to pay an external contractor for the creation of the data extraction tool; to pay an external contractor for a piece of data analysis; and to pay for travel to face-to-face project meetings. OCLC contributions for project management and data analysis were in-kind.

[3] Smarty is a templating language; see New Digital Group, Inc. n.d.

[4] MIMSY XG was previously owned by Willoughby, and has been acquired by Selago Design in2009.

[5] Available as Open Source through the OpenSiteSearch project at SourceForge (OpenSiteSearch Community n.d.).

[6] Available from OCLC Research n.d.d.

[7] Of these 131 units of information bearing data content, 67 are data elements, and 64 are attributes. For a detailed view of CDWA Lite information units of information bearing data content, as well as a mapping to CCO, please see OCLC Research n.d.a.

[8] There is a slight discrepancy between the schema and the documentation: in the schema, *recordID* and ‹recordType› are only required if their wrapper element (*recordWrap*) is present; the documentation, however, calls both data elements “required.”

[9] A detailed mapping of CDWA Lite to the COBOAT default mapping can be found in OCLC Research n.d.a.

[10] All quotes in this block refer to Getty Trust, J. Paul 2006.