7-27-2011 (revised Jan 2014)
Emily Nedell Tuck
Museum of Fine Arts Houston
elibrarian@hotmail.com

CCO as a Metadata Standard for the Retrieval of Museum Cataloging Records: A Critical Review

CCO was developed by the Getty Research Institute and the ARTstor Digital Library (www.artstor.org) as a set of rules for populating CDWA Lite, an XML schema for contributing records to the ARTstor digital art image library. *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images* was then published as a cataloging standard, referred to as "CCO," where it was deemed ideally suited to relational databases,[1] largely on the assumption by the authors of the standard that relational databases excel at *relating* (what they call "linking") data all different ways, including reciprocally and hierarchically,[2] rather than an understanding that relational databases are optimized for storing and retrieving *structured* data and require requiring a very precise semantics to ensure effective retrieval of data from the database. [3] Relational databases typically use SQL, Structured Query Language, to perform queries based on Boolean logic, where a record is considered either relevant or not according to the presence or absence of a term in a particular indexed field.

When it comes to information retrieval, Boolean systems are characterized by the following limitations:[4]

- Retrieval based on binary decision criteria with no notion of partial matching
- No ranking of the documents is provided (absence of a grading scale)
- Information need has to be translated into a Boolean expression which most users find awkward
- The Boolean queries formulated by the users are most often too simplistic
- As a consequence, the Boolean model frequently returns either too few or too many documents in response to a user query

Data standards for text search or "text-centric XML"[5] (like CDWA Lite records), as opposed to those written for relational databases, are different, because they present different requirements for information retrieval. If CCO is to provide the data content standard for

[1] *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images.* ALA: Chicago, 2006, p. 20.

[2]Hierarchical and relational are two different and mutually exclusive database models.

[3]"Unfortunately, the Boolean model suffers from many drawbacks. First, its retrieval strategy is based on a binary decision criterion (i.e., a document is predicted to be relevant or non-relevant) without any notion of a grading scale, which prevents good retrieval performance. Thus, the Boolean model is much more a data (instead of information) retrieval model." Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, New York: Addison-Wesley, p. 26.

[4] Rada Mihalcea, "The Boolean Model," www.cse.unt.edu/~rada/CSCE5200/Lectures/BooleanModel.ppt, p.6.

[5] XML can be either "data-centric" or "text-centric," depending on its contents. The CDWA Lite XML records and CCO catalog records are more text than data, particularly because of the lengthy descriptive free text fields in catalog records and the lack of precision of the data (for example, some institutions may use Cubism, other Cubist, etc.). The leading expert on Information Retrieval, Dr. Christopher Manning, puts it this way: Data-centric XML "mainly encodes numerical and non-text attribute-value data. When querying data-centric XML, we want to impose exact match conditions in most cases." It corresponds to the kind of data one finds in databases. Relevance ranking is not required. On the other hand, with text-centric XML, "unstructured retrieval methods" are "adapted to handling additional structural constraints." . . . text-centric XML document retrieval is characterized by (i) long text fields, (ii) inexact matching, and (iii) relevance-ranked results." *Introduction to Information Retrieval*, By Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, 2008. http://nlp.stanford.edu/IR-book/html/htmledition/text-centric-vs-data-centric-xml-retrieval-1.html.

cataloging records in relational databases and for populating XML records for online repositories, these different requirements for search need to be reconciled. Text search and database search are quite different, requiring different requirements for data standards.

CCO as a Data Standard for "Enhanced End User Access" in a RDBMS

As a data content standard for a collection management system, CCO does not enforce the kind of precise semantics needed for a traditional relational database management system (RDBMS), and cannot realistically be implemented without extending the database's search capabilities to incorporate text handling techniques which could equate, for example, lexical variants such as "Impressionist" and "Impressionism." Text search engines do this through a process called word stemming, where the root of the word (impression) is indexed after punctuation, accent marks, and common endings are stripped out.

Because CCO was intended to be expressed in XML, a content standard is better suited to less structured search strategies than what is required by traditional relational database applications, the CCO standard and its incumbent authorities do not define, for example, specifically what values go into "Period" as opposed to what values go into "Style" or "Culture." With keyword search capability, which is normally supported by a text search engine, it doesn't matter so much where these terms appear in the record. But with relational databases fields have to be very well-defined and predictable, because users have to select what field their data is in before conducting a query.

In addition, the large amount of *descriptive* text captured by a typical CCO record—or any museum catalog record containing descriptive text—would benefit from models that are widely used today by text search applications[6] and library catalogs, but which are in fact not *native to* relational database management systems[7] at all. Relational databases, or the search that is native to them, SQL, do not do relevance ranking. The objectives of *descriptive* cataloging require text search and a text format, along with relevance ranked results.

It might come as a surprise to many, given the numerous published examples of CCO-compliant cataloging records on the Getty's website and in the CCO manual itself, that there is no application—an actual database—at the Getty which uses CCO Work Records for cataloging and record retrieval.[8] The Getty Museum, which uses Gallery System's TMS, has not implemented CCO. Nor has CCO's recommended method of authority control, the Getty Vocabularies, which have now been released as Web Services, been critically evaluated. The AAT, recommended by

[6] A vector space model, a vector text search engine, a text search engine, search engine, and 'Google-like' search, are the same thing. There are many articles explaining vector space models for text search on the Internet , as it is taught in every Introduction to Information Retrieval classes; but my favorite source is "Building a Vector Space Search Engine in Perl" by Maciej Ceglowski (February 19, 2003), http://www.perl.com/lpt/a/713, because it explains term and category weighting, stemming, and relevance ranking in a relatively nonmathematical way.

[7] Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? http://www.cidrdb.org/cidr2005/papers/P01.pdf, p.2.; "Integrating Text and Data." David A. Grossman and Ophir Frieder, *Information Retrieval: Algorithms and Heuristics*. Springer, 2004

[8] Confirmed by email correspondence.

CCO for controlling concepts, was structured to help manual indexers to locate terms which can be used in combination with other terms (European + painting). Particularly as adjectives seem to be lumped into in one facet and nouns in another, this strategy of linking to term ids would seem to result in a bland genericism (e.g., why would one want to control terms such as "cup," "chair" or "painting"?) when used as a controlled vocabulary.

CCO is at this point a *de facto* cataloging standard, based on the fact that it supports CDWA Lite, an internationally accepted XML schema for museum data exchange and publishing; and also on the rationalization that it would be beneficial if the data content standard for catalog records corresponded to the data content standard for data exchange of the same.

A sample CCO Work Record is supposed to look something like this,[9] with Work Type, Subject, Style, Culture, and most other contextual data fields, containing repeating, authority-linked values:

> **Class** *[controlled]:* • paintings • European art
>
> ***Work Type** *[link]:* • polyptych • altarpiece
>
> ***Title:** Polyptych with Saint James Major, the Madonna and Child, and Saints **Title Type:** repository
>
> ***Creator Display:** Bartolomeo Vivarini (Italian, ca. 1432-1499)
>
> ***Role** *[link]:* painter *[link]:* Vivarini, Bartolomeo
>
> ***Creation Date:** 1490 *[controlled]:* • **Earliest:** 1490 • **Latest:** 1490
>
> ***Subject** *[links]:* • religion and mythology • Madonna and Child (Christian iconography) • Saint James Major (Christian iconography) • Jesus (Christian iconography) • Saint Mary Magdalene (Christian iconography) • Virgin Mary (Christian iconography) • Saint Bartholomew (Christian iconography) • Saint Peter (Christian iconography) • Saint Catherine (Christian iconography) • Saint John the Baptist (Christian iconography) • Saint John the Evangelist (Christian iconography) • Saint Apollonia (Christian iconography) • Saint Ursula (Christian iconography) • martyrs • saints • pilgrimage
>
> **Culture** *[link]:* Italian
>
> ***Current Location** *[link]:* J. Paul Getty Museum (Los Angeles, California, USA) • **ID:** 71.PB.30
>
> ***Measurements:** comprises 10 panels; overall: 280 x 215 cm (110 1/4 x 84 5/8 inches) *[controlled]:* **Extent:** components • **Value:** 10 **Type:** count | • **Value:** 280 **Unit:** cm **Type:** height | • **Value:** 215 **Unit:** cm **Type:** width
>
> ***Materials and Techniques:** tempera and gold leaf on panel
>
> **Material** *[links]:* • tempera • panel (wood) • gold leaf **Technique** *[links]:* • painting
>
> **Description:** The themes of martyrdom and pilgrimage are strongly represented in this polyptych. The central saint, Saint John Major, holds a pilgrim's staff and shell, references to a famous pilgrimage site dedicated to him, Santiago de Compostela; pilgrimage sites were also dedicated to several other of the saints depicted. All of the saints depicted were martyrs, with the exception of John the Evangelist, Mary Magdalene, and the Virgin Mary. However, two of those three may be linked to martyrdom: John the Evangelist was thought to have survived an attempted martyrdom; the female saint with a jar has been identified as Mary Magdalene, but she carries a martyr's palm, so perhaps that identification is mistaken.
>
> **Description Source** *[link]:* J. Paul Getty Museum online. www.getty.edu **Page:** accessed 15 October 2006

It is not clear whether the CCO Work Record is intended to be a fundamental unit of storage (stored as a BLOB or CLOB)[10] or else is meant to be broken out into rows in tables in a relational database and, in response to a query, reconstructed into some kind of text format for additional

[9] CCO website, http://vraweb.org/ccoweb/cco/example4.html.

[10] Binary Large Objects, Character Large Objects. BLOBs are used to store binary information, such as images, in relational databases. CLOBs are used to store character information, such as large text files. These may be indexed and searched differently from the rest of the fields in the database.

processing and display. Is the CCO record meant to be a result set in a web-based system with a relational database on the backend? Is search going to be performed against a collection of Work Records in text format, or against database records whose contents may spread across various tables?

Why does it matter? *Because any data standard designed for record retrieval is going to depend on the search strategies supported by the system*. Better retrieval ("enhanced end-user access") is one stated goal of CCO data standard, along with good descriptive cataloging and data interoperability ("shared documentation"), but a traditional (Boolean) relational database management system, what is recommended by CCO, is not up to the task of any data content standard that does not strictly enforce, on a field by field basis, *exactly* how those fields are to be populated. In a Boolean environment, field values need to be predictable—there is no "keyword" search. CCO makes the assumption that, at the end of the day, one can freely search on Subject, Style, Culture, Period, Movement, Medium, Classification, Work Type, etc., but in fact, for particularly for large collections, the contents of these fields are not able to either be controlled enough, or made predictable enough, to support effective retrieval by a relational database management system alone.

A simple way of thinking about the problem is this: It is not uncommon for traditional relational database applications to require searchers to specify the field where their data might be found (again, there is no keyword search feature, which is typical for relational databases), and precisely how the term appears in the database. Variant forms of a term are indexed separately--the name with a diacritical mark and the name without are indexed as separate terms. This is typical of a SQL-based or traditional relational database application. For this reason, relational databases are good as "known-item" retrieval tools rather than discovery/search tools. Relational databases presume familiarity with the data, and datasets capable of being structured in such a way that field values are predictable. It is for this reason purely Boolean systems were phased out years ago by vendors of library catalogs, and Boolean models are no longer the basis for modern information retrieval systems or content management systems.[11]

The alternative model, and one which is widely used in web and text search engines, is known as a vector-space model, or vector model—but most people just call it "Google-like search" or "text search" or "text search engine."[12] This is a more appropriate search technique for textual data, such as a catalog record with a lot of good descriptive textual information, and especially one with dynamic access to authority files. Through access to a thesaurus or taxonomy, a text

[11] Boolean models are no longer the basis for modern information retrieval systems, and even bibliographic databases have adopted "vector" models capable of considering documents that only partially match a query. Where in 1999 bibliographic systems were based on Boolean models, today all incorporate some degree of text handling; in some systems, MARC records are stored as BLOBs in the database.

[12] The key ingredient in Google-like searching is not the Web, but textual data (HTML docs, XML metadata or documents, and text in any indexable format). This approach eliminates many of the problems associated with keyword searching in a relational database. Vector models may also be extended to handle XML. See "A Vector Space Model for Information Retrieval,". http://nlp.stanford.edu/IR-book/html/htmledition/a-vector-space-model-for-xml-retrieval-1.html. Category weighting in addition to term weighting can emphasize terms which appear in parts of a record, such as the data in the structured fields of the record.

search engine can return related records, only ranked less highly as records containing an exact match.

My point is not to disparage existing systems, but to highlight some of the challenges applying a standard meant for XML to a relational database application. A data standard designed for populating records in a relational database is very different than one for cataloging in a text format. In the case of the former, the standard needs to compensate for the limitations of structured search strategy by enforcing rigid controls on fields, leaning heavily on controlled vocabularies, pick lists, and so forth, to ensure data consistency so the data can be pulled back out again.

In such a system, having just a few ways of classifying objects and a limited number of indexed fields is advantageous for record retrieval ("recall"), but is not good for descriptive cataloging, publishing, or supporting research, or recommended in any context where linguistic precision is desired. A data standard for system capable of word stemming,[13] term weighting, proximity searching and relevance ranking—for example, where a search on "Cubism" could also pull up records containing "Cubist," where a search on "amphorae" would retrieve "amphora," or where "arm chair" would be automatically indexed as one term because "arm" and "chair" commonly appear together in records, obviously requires a different standard for data entry. Word-stemming and other text handling approaches, now built into enterprise search applications, as well as many commercial database packages,[14] place less burden on catalogers to maintain a physical thesaurus to represent variant word forms, less burden on the database to process queries, and less burden on users to be able to predict how (format) and where terms are represented in the database.

CCO does not fit a traditional relational database model for record storage and information retrieval, because relational database searching has more stringent requirements for data than what the CCO framework supports. For any given object cataloged according to the CCO content standard, the **exact same authority controlled descriptor/index term** could appear in one of **several different parts of the Work Record in ways that cannot be anticipated by end users**, making it difficult even for those who know the standard to find what they are looking for through a structured query: e.g., adjectives such as "Pre-Columbian," "Hellenistic," or "Baroque," could appear in any number of CCO fields: Style and Period, Culture, or attached to

[13] Stemming is part of the indexing process in any text search. Words are stripped down to their roots (stemmed) and assigned weights proportional to the number of times the root term appears in a document, sometimes weighted also based where it appears (if in the title, for example), and finally, the degree of similarity/dissimilarity to the query term or terms. A vector is a mathematical measure of the degree of similarity to the query relative to all of the records in the system.

[14] Microsoft SQL Server uses a full-text search technology called SQL Server Full Text (SQL FTS), see "http://en.wikipedia.org/wiki/SQL_Server_Full_Text_Search"; "Oracle uses a technology called Oracle Text," http://www.oracle.com/technology/products/text/pdf/11goracletexttwp.pdf. These are ways of extending a traditional relational database to handle unstructured text and xml on particular fields of the database. In both relational database systems, fields containing unstructured text or xml require different search algorithms as well as a separate method of indexing in order to get good results. Even with these, organizations often resort to exporting data from their database into a separate text search system such Lucene to facilitate search objectives.

nouns in the Object Name, Work Type or Class. An Ionic column could be cataloged as an “Ionic column” or as a “column” with a Style of “Ionic;” and Class could be any broader category one chooses to add to the record, making its values completely unpredictable.

The lack of predictability on a field by field basis makes CCO particularly unsuited as a data content standard for Boolean approaches, likely resulting in too few results from term matching on authority controlled fields, and too many irrelevant records from term matching in the narrative free text fields.

Another reason CCO is not precise enough for a traditional relational database management system is that it recommends index terms (AAT) for authority control. Terms from the AAT are meant to be used in combination with other index terms to represent a single concept (European + Art, tin glaze + eathenware). Organized into facets, the AAT lumps common adjectives into one “Styles and Period” facet (the source for CCO elements Style, Period, Culture, Movement and Reign), and nouns into the Objects facet, except in instances where terms are bound into noun phrases. Because sometimes terms are bound (“Morris chair,” “Aeolic capital,” “boudoir print”) and other times unbound (Ionic + column, Greek + vase, Japanese + print) in an arbitrary way (meant to save space in the printed editions), it is hard to apply the AAT consistently for vocabulary control, let alone rely upon it for subject control. Common topics like “Greek vase” and “Japanese print” are not represented. Again, structured queries require highly predictable values, and it would be impossible to use the AAT as a data value standard in a structured query where “Ionic column” would require a different query formulation than “Aeolic capital” simply because of the organizational structure of the AAT.

Some of these criticisms should not be regarded as a shortcoming of CCO as a data standard for cataloging cultural objects, but of the presumption by its authors that a relational database is the most suitable model for it. In CCO, this assumption appears to be based merely on the need to provide some practical mechanism for linking (relating) Work Records to other Work Records, but more importantly, linking Work Records to Authority Records (“**Because of the complexity of cultural information and the importance of Authority Records, CCO recommends using a relational database**.”). Now that the Vocabularies are available in a text format (XML UTF-8) and also accessible as a Web Service, one wonders if this recommendation still holds.

The integration of unstructured or loosely structured text with database technologies for structured search is a huge topic in the field of computing today, with advancements having been made using text search engines[15] against “flattened” records (the data is pulled together from different tables in the database to create a longer record more suitable to text search) in a relational database, and/or else indexing columns of text or XML after pre-processing the data in those fields. An open source search engine, Apache Lucene/Solr, is now being used to try to apply Google like search syntax and indexing to databases, but people are often finding it

[15] Surajit Chaudhuri, Raghu Ramakrishnan, Gerhard Weikum, Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? http://www.cidrdb.org/cidr2005/papers/P01.pdf.

necessary to exporting the data from their database into Lucene for full text indexing and search (that is, maintain two separate systems) in order to achieve desired results.[16]

CCO as a Standard for Descriptive Cataloging

Museum collection management systems have served primarily the needs of registrars to accession objects, control inventory, track the location of objects, run reports, and to locate items already known to be in the collection. The ability to capture good descriptive data about an object has not been high on Registrars' wish lists in terms of system functionality, and it has probably been for this reason that TMS, the leading collection management software for art museums, has not offered the kind of enhanced search capabilities or required fields to support descriptive cataloging in the way CCO recommends. Unlike library catalogs, museums systems were not developed with public search in mind, and they do not support much descriptive metadata.

In fact, some see "descriptive cataloging" as just creating more work (As one person said after a presentation we gave, "I'm supposed to enter all of this into TMS just because someone might come along one day and look for it?"). In libraries, catalogers already assume that they are creating metadata for enhancing end-user access, i.e., creating records with authorized headings, descriptive notes, links to websites, and cross references, for the purpose of promoting public access and/or scholarly research. Similarly, and because the data format and content standard (the MARC record, AACR2) supported multiple entries, the approach to cataloging was oriented towards anticipating, within reason, the ways many users would approach a work, rather than devising prescriptive categories for classification (e.g., "Pastels and watercolors are to be classified as a Drawing").

Support for descriptive cataloging is not a big selling point for Registrars, often the ones who select and maintain the collections database in the museum, because description is viewed as more a curatorial than a registration function. (It is hard to imagine an RFP for a museum system specifying that the system had to support "descriptive cataloging" and "authority control"). And to the extent Curators are interested in the collection management system, it is often only how can be used to generate tombstone data for wall labels. They often do not see cataloging as their function either, because it is not perceived as a publication (like a published catalog).

The organizational structure of museums could itself present obstacles. The production and maintenance of a public access catalog could require the cooperation of several departments: registrars, publications, curators, visual resources (rights and reproductions), IT, and marketing. Because the larger purpose of CCO is not so much about the disposition of one institution's records, but about *interoperability* and record sharing, the cost of implementation—including additional staffing—would need to be offset by benefits to the institution, to the museum and arts community as a whole.

[16] Lucid Imagination (website devoted to enterprise support for Lucene and Solr) http://www.lucidimagination.com/Community/Hear-from-the-Experts/Articles/Search-Engine-versus-DBMS

When discussing CCO, there is a bigger picture which needs to be communicated, with the end result being the creation a global knowledgebase of cataloging records for researchers, curators and catalogers, to easily locate objects for exhibition, collaborate, and even download records for copy cataloging. CCO is just a start towards a more collaborative approach, the result of which might be a complete sea change in workflows and better catalog records for everyone.

Regardless of whether CCO is implemented, museum standards and systems should support each other, as well as better search/text handling capabilities to be able leverage the textual data in museum cataloging records as assets—as metadata—for improving search—rather than as it is in a Boolean system, a liability (the more free text is added, the less likely the system will return relevant results). It should be kept in mind that while subject experts commonly search by artist name and title, the public and nonsubject experts—even new staff trying to learn about the collection—are most likely going to enter a subject term or keyword: e.g., “impressionism,” Pre-Columbian,” “Art Nouveau,” “Last Supper,” “Dada,” “African-American art,” etc, to discover what is in the collection. This is why a more flexible system capable of text search is so important.

Applicability to Existing Collection Management Systems. Because of its self-describing nature, readability, and platform independence, XML is now used in many industries for data exchange. Even though XML follows a more hierarchical structure than a relational database, it is not hard to convert table structures to XML schema and vice versa, and there are many mapping tools available to get the job done.

However, this does not mean that any XML Schema can be mapped to legacy database with good results. CDWA Lite, the XML Schema which expresses CCO, may have been intended to be the *lingua franca* of catalog records, but legacy databases (that is, those behind an existing collection management system), may **have not have the underlying table structures that will support the CCO standard for data import, export, or cataloging**. A technically valid CDWA Lite XML record can be generated from it, but the elements or attributes will be empty, or under-populated, because it does not correspond to the schema of the database.

With funding from the Mellon Foundation, RLG (part of OCLC) and a company called CogApp have developed software called COBOAT, a “metadata extraction tool” to extract data from TMS to generate CDWA Lite XML records. A big part of this grant was to use this software to evaluate the compliance of seven museums to CCO by analyzing the result set, once the CDWA Lite XML data was extracted.[17] According to the press release (Feb. 2008):

> This initiative will result in the creation of a low-barrier/no-cost batch export capability out of the collections management system used by the participating museums (Gallery Systems TMS), as well as a test of data exchange processes using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The test will create a large research aggregation of museum records, which will be analyzed to determine in which areas museums should invest in upgrading their records, and in which areas automated processes can be utilized to harmonize descriptions

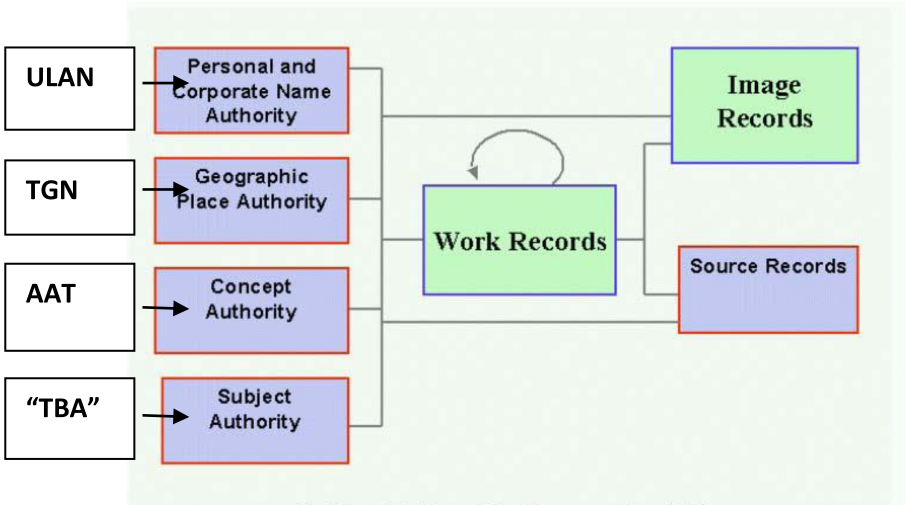
[17] http://www.oclc.org/research/activities/museumdata/default.htm

> for retrieval. Participating museums will also discuss the evidence about the relative utility of the aggregation with stakeholders from the museum, vendor and aggregator communities.[18]

A few months ago, when I began mentioning incompatibilities of TMS with one of the authors of CCO, I was told "Well, we don't know that, the results of the OCLC study are not in yet." Unless I am missing something, the results of this study seem to be predictable.

There is no good way to map a CCO/CDWA Lite record to a relational database if the database's tables have not been set up to hold the elements, attributes and links needed to express CCO. A table in the database may have been set up to hold only one data value per field, for example, in which case there is no suitable place to store any additional values or links. There may be no place to capture multiple classes, work types, subjects, styles, periods, and no way to link to authorities in those fields. Since data interoperability and record sharing is a goal of the CCO, it should be recognized that a significant barrier to adopting CCO is that legacy databases may not have the table structures in place to hold the content that supports the standard.

CCO recommends creating linked entries to AAT authority records for Work Type, Style and Period, Culture, Materials and Techniques; ULAN, the name authority for artist names; to TGN, the geographic name authority, for locations; and to some Subject authority for subject indexing:



Entity Relationship Diagram for CCO [19]

The idea is that when the authority record in any of the Getty Vocabularies is updated, for example, when an artist dies or name changes, all of the indexed terms in the Work Record will be globally updated by virtue of a link[20] to the term's id in the authority file. Also, when

[18] http://www.oclc.org/news/releases/200695.htm; A succinct overview of the scope, purpose and progress of the OCLC Museum Data Exchange Project can be found at http://hangingtogether.org/?p=368.

[19] *Cataloging Cultural Objects: Introduction and Application of CCO and CDWA* http://www.getty.edu/research/conducting_research/vocabularies/intro_to_cco_cdwa.pdf

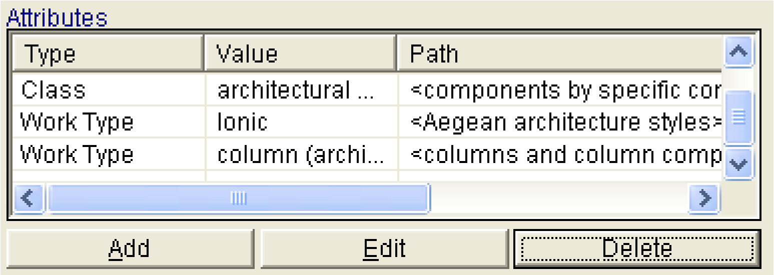
[20] Web-based systems can be programmed to pull in updated content from external websites by virtue of AJAX scripting, a combination of XML and Java Scripting. This type of approach, where one pulls in very specific content from other websites,

someone performs a query, the thesaurus structure can be leveraged to permit search expansion, so that synonyms and other related terms can be used to broaden the search.

In an XML record or a web-based application, it would be possible to link values *directly* to authority records in online thesauri. Self-updating links can be created through AJAX (**A**synchronous **J**ava **a**nd **X**ML) scripting,[21] but to make use of the thesaurus structure one would have to go through an API or Web Service.[22] One museum system, ADLIB, has developed an API to link to Iconclass, a hierarchical iconographic database. The Getty Information Service is developing a set of Web Services (APIs) for the AAT, TGN and ULAN which can be accessed by application developers for direct linking. In a web-based collection management system, one could create dynamic, self-updating links to one's own local thesaurus, although links to a common, external thesaurus is better for record sharing and interoperability. One institution doesn't want to link their records to another institution's thesauri or authority files, which would probably live behind a firewall anyway.

Links in TMS occur only through the ThesXRef Attribute field, a special grid-like form control (see below) that can be used to link values to terms in a thesaurus. Entries in the ThesXRef Attributes field must be linked to an internal version of the AAT, or custom authority file called the AUT. Links are actually joins between the Objects and Attributes tables in TMS and the table in the authority database where the term resides. It is a time consuming process to create linkages. The Getty name authority, ULAN, is not offered by TMS, perhaps because it would be difficult to parse the information out of the name authority record to populate the appropriate fields in the Constituents table—or else because it was thought to become more quickly dated than the other vocabularies, which are updated in TMS only when a customer does an upgrade.

In TMS, any element requiring *linking to a thesaurus,* or *requiring multiple values*—two aspects which comprise the very essence of the CCO standard—has to go inside the ThesXRef Attributes (Thesaurus Cross-Referenced Attributes) table rather than on the main data entry form:



Attributes

Type	Value	Path
Class	architectural ...	<components by specific cor
Work Type	Ionic	<Aegean architecture styles>
Work Type	column (archi...	<columns and column comp

sometimes referred to as a "mash-up," is at the heart of Web 2.0, but it is an aspect of CCO not discussed in the published standard, which is intended for relational databases.

[21] See Bert Degenhart Drenth, "Using Web Services for Terminology Control." Paper presented at CIDOC, 2008. http://cidoc.mediahost.org/content/archive/cidoc2008/Documents/papers/drfile.2008-06-82.pdf.

[22] "A Web Service makes a programmatic application interface (API) accessible to remote applications over the Internet just as an HTML server makes an application user interface accessible to browser clients over the Internet. Web services are accessed using the XML-based Simple Object Access Protocol (SOAP). The data is also returned as XML. Therefore, by using Web Services, XML data can be transparently imported to the underlying database." Importing XML documents into Relational Databases using Java (white paper), Ale Gicqueau, Business Development Manager HiT Software.

Not only are essential, required CCO record elements be marginalized as attributes, but the semantic relationship between terms is lost whenever a noun phrase (e.g., "Ionic column") is created by combining terms from different facets of the AAT. In the example above, these should not be two Work Types, but one Work Type comprised of two AAT terms, "Ionic" and "column." Class is in there because CCO recommends multiple values for Class, and two Classes have been assigned.

With the majority of the CCO Work Record—Classification, Work Type, Style, Subject, Culture, and other descriptive cataloging elements—shoe-horned into a small subfield, where the data becomes difficult to see, search (because a user has to select "ThesXRefAttribues") and incorporate into reports—it is hard to say that TMS supports the CCO standard, or descriptive cataloging in general. Because without using ThesXRef Attributes, the system *does not allow enough access points to be assigned* to properly catalog an object. TMS's table structure makes it impossible to create repeating values per field (multiple Classes, Subjects, Work Types, Styles, etc.) like CCO prescribes. One value per field is not enough for descriptive cataloging purposes, especially if the objective is good metadata for search and retrieval. As large institutions begin to move their collection information online, they are purchasing additional products—from content management systems to DAMS--to facilitate the creation of richer metadata for better search and retrieval.

For example, at the Tate, another TMS institution, an indexing project was undertaken to allow users to search by Subject (which includes thematic as well as historical and contextual information http://www.tate.org.uk/servlet/SubjectSearch), but a separate product was purchased, and it was never included into the core cataloging functionality of TMS. Similarly, the Met, which also uses TMS, offers a Subject index http://www.metmuseum.org/toah/hi/a.htm not limited to iconography, but includes media, genre, and historical terms linked to exemplary objects from the collection. Museums are having to buy or develop intermediary applications for the purpose of enriching metadata in order to overcome the limitations of TMS when it comes to descriptive cataloging.

Hierarchal Classification of Works of Art and Artifacts. Gallery System's TMS is the most widely used collection management system for art museums. While some customization is possible, its default data entry screen largely determines the kind of information museums may capture about objects. TMS offers only one Classification field, capable of holding one value. Given the one value limit, which was probably designed for "labels" and not metadata, it might be tempting to subdivide broader the classifications into narrower ones, an approach which is encouraged by TMS's trainers. Indeed, CCO says that "class terms may represent a hierarchy." However, TMS treats the subdivided classifications as literal strings, not as hierarchies. So, if one were to apply subdivisions to the Classification field, for example, for ceramic pottery:

Ceramics
Ceramics—Earthenware
Ceramics—Polychrome
Ceramics—Porcelain
Ceramics—Stoneware

a search on "Ceramics" would not retrieve the records containing the subdivided terms. Adding a third level of classification, for example to designate culture or geographic location, would result in drop outs from the broader heading created by classification term1 + term2. The database would need to be programmed (called a "hierarchical search strategy") to reflect the logic that subdivided terms are *hierarchically related* to the term "Ceramics."

In a working group discussing ways to implement CCO standards in TMS, someone posed the question why is it "these discussions" about CCO always come down to the topic of Classifications. What is your institution doing for Classification? The thought process seems to be *if only our method of classification were more precise*, our records would be found. Classification lists are exchanged among institutions, but never seem quite good enough. The answer is really quite simple, and it is not "get more curatorial input."

There is no hierarchical classification system that is nuanced enough for cultural objects, *if* objects can only reside in one location in the scheme. This is because classification schemes offering *one point access* are good only for inventory control (i.e., for answering, "How many X's in the collection do we have?"). Indeed, art objects are the museum's inventory, but the purpose and mission of the museum is to emphasize the "art" part. Museums are academic institutions and its mission to promote scholarship should be reflected in its data. People show up to experience great art, not to experience cups, chairs, bowls, and paintings in the abstract. Good metadata should provide the context for art to be regarded as such, and all hierarchical systems objectify art as physical and functional objects. If we are going to support scholarship through our catalog records, it is imperative that we start thinking in terms of "classes," and not classification.

Classification seems the most likely field for expanding intellectual access to the collection. CCO suggests that multiple classes be assigned to objects in order to relate them to a broader context, "categorizing on the basis of similar characteristics, including materials, form, shape, function, region of origin, cultural context, or historical or stylistic period."[23]

In cataloging examples for CDWA Lite, the Getty Research Institute usually applies two or three Classes to a work record. Some represent the scholarly tradition to which the object belongs (e.g., European art, Asian art, Native American art, decorative arts, etc), and others to capture the class of object based on function (furniture, painting, sculpture, musical instrument). One museum using TMS created two Classifications through a user defined form in an effort to try to capture the intellectual tradition and a formal category for the object.

[23] CCO, 235

The challenge in creating pre-defined lists of terms is—no, not "need for more curatorial input" —that Class is supposed to stand in relation to Work Type, allowing a broader context for the object. If one is applying the AAT descriptors as if enumerated terms are the only authorized terms,[24] as seen in the many cataloging examples at the CDWA website,[25] then Work Type will be either very specific ("hatcha" for an ancient Incan ceremonial ax) or very generic ("cup" for a Mayan drinking vessel). CCO suggests that Class is based on the level of description in Work Type, making it hard to come up with a pre-defined list. If the object is of a Work Type "painting" would classing it as "paintings" provide *enough* of a context? What may be enough for registration purposes ("How many X's do we have?") or serendipitous browsing ("Let's see if anything comes up!"), will not be enough for scholarly research.

We could customize TMS to accommodate multiple Classes by creating user defined fields on a user defined form. But then we would be put into a position of having to anticipate how many "Class" terms would be needed to describe every kind of object in the collection. For example, in the case of a Native American instrument called a *kizh kizh dihi*, one might want to add three descriptors, requiring three separate Class fields:

Class 1	decorative arts
Class 2	musical instruments
Class 3	Native American art

Not only would we also need to configure the Query Assistant to search Class1, Class2, and Class3 so the user could find all of the items which belong to a particular class, but also this approach would need to be duplicated for every CCO element that needs to support multiple values, such as Subject, making swiss cheese out of our input form. This would seem an inelegant solution to arrive at this XML equivalent:

```
<cdwalite:classificationWrap>
        <cdwalite:classification termsource="AAT" termsourceID="aat300033168">
        decorative arts</cdwalite:classification>
        <cdwalite:classification termsource="AAT" termsourceID="aat300041620">
        musical instruments</cdwalite:classification>
        <cdwalite:classification>
        Native American art</cdwalite:classification>
</cdwalite:classificationWrap>[26]
```

Despite the efforts of authors of CCO to create a universal standard for cataloging cultural objects, it is at its core an XML standard for populating records with less structure and more flexibility than what some relational database systems require. In fact, the Getty Museum itself

[24] The AAT's editors intended for terms to be used in combination with other terms, not to limit usage to the terms which are enumerated in the index.

[25] http://www.getty.edu/research/conducting_research/standards/cdwa/examples.html

[26] *www.getty.edu/research/conducting_research/standards/**cdwa/cdwalite**.pdf*

has not successfully implemented CCO in its own TMS catalog. Despite the many examples[27] of cataloging records in XML and XHTML from the Getty Research Institute to illustrate the standard they are promoting, there is no sign of a working application, to prove, for example, that the AAT can provide for authority control. When I asked Patricia Harpring in an email "What is the Getty doing with all of those sample CCO and CDWA cataloging records?" the response was vague:

> The CDWA examples are only examples -- not in a specific editorial or collection management system. That's the whole point of having a standard. Your data can be in any system. A standard should be system-independent. It doesn't matter what system it would be in because if you have good data that is parsed into discrete fields, you can parse it into any format, including CDWA Lite (which is expressed in XML). I'm sure people use even a single system, such as TMS, differently. Again, doesn't matter - because you can map to the standard when you exchange data or participate in a consortium.[28]

The assignment of multiple Classifications, or Classes, is not currently supported by TMS as port of its core functionality. In order to follow the Getty standard--in order to convert the museum *registration database* into a *catalog*--it is essential to offer more points of access to the Work Record and also offer a method of authority control on each of the indexed fields on the default entry form, so that catalogers have a way of efficiently and consistently populating each of the fields.

Using the AAT for Authority Control over a Museum Catalog

Cataloging Cultural Objects can be credited for introducing to the museum community the concept of authority control, which, despite its being an integral part of the cataloging process in libraries, is not well understood even by most librarians.[29] However, it is easier for librarians to understand the concept than museum professionals because librarians have the advantage of well developed systems, standards, and shared vocabularies (authority records) which work in concert with each other to achieve "controlled access." Authority control has been around for at least 40 years in libraries. Because of this, and because of advances in search technologies, users of library collection management systems are afforded a fairly intuitive and optimal search experience.

Most libraries use the Library of Congress Authorities which includes subjects, names, and titles. LC authority records may be freely downloaded into library system software and easily updated independently of new releases of the automation system software. Although different systems

[27] For example: Categories for the Description of Works of Art, http://www.getty.edu/research/conducting_research/standards/cdwa/examples.html; Patricia Harpring, *Cataloging in the Museum CCO, CDWA, and CDWA Lite for Cataloging Museum Objects*, June 2009, http://www.getty.edu/research/conducting_research/vocabularies/cco_cdwa_for_museums.pdf; CDWA Lite Specifications, http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.pdf.

[28] Patricia Harpring, email received, Nov. 23, 2009.

[29] Authority control is only well understood by systems librarians and catalogers, "Technical Services" Librarians. Therefore museums professionals may not approach it with the same rigor or critical acumen since the concept itself is probably new and unfamiliar. Robert E.Wolverton, Jr. "Becoming an Authority on Authority Control: An Annotated Bibliography of Resources." *Library Resources & Technical Services* 50.1 (2006): 31-41. *Academic Search Complete*. EBSCO. Web. 1 May 2010.

implement authority control in different ways, authority records are sometimes linked to catalog records. This enables unproductive search terms to be redirected to authorized heading (e.g., **Ivory Coast**. See **Côte d'Ivoire**) and suggest to users ways that their searches can be expanded through related headings, e.g., **Art Nouveau**. See also ***Jugendstil*** (Germany and Austria), **Vienna Secession** (Austria), **Aesthetic Movement** (England), etc. Where "Sees" can retrieve a result set based on a redirected query, "See alsos" require a user interface to present users with choices.

Public, semi-public, or even easy access to collection information has not been a high priority in museums. The modern museum collection management system is a database application useful for tracking inventory, not intended to support research. Its cataloging module is an input form permitting one value per field, not designed for creating access points, performing "descriptive cataloging for enhanced end-user access," or implementing the kind of principles or controls alluded to in the CCO manual. Because the concept of authority control, the AAT, and to some extent, the CCO manual itself, has come from library science (CCO was published by the American Library Association), no one has taken note of the fact that the AAT—which was created as a reference tool for manually indexing monographs and articles—has been repurposed as an authority file.

Art museums have invested in the same software, Gallery System's TMS. TMS is a relational database application meant to be searched by a trained individual capable of translating complex queries into Boolean logic (and sometimes it is just easier to construct a SQL query). Public access to collection data, or even the need for user friendly interfaces which can allow people to efficiently access collection information, is such a new way of thinking for museums that many do not even expect their collection management system to offer an easy way to search it. The attitude seems to be "You cannot expect it to do everything." In museums, it is not uncommon for IT to leverage newer web and search technologies (XML, Lucene, Solr) against the collection database to make it searchable by people outside of the Registrar's Office. At larger museums, there may be three parallel applications: one for registration/data entry (the collection management system); one intended for search by museum staff, accessible through an intranet (a web application with text search capability run against a derivative, flattened database); and one for the public that is accessed through the museum's website (also run against a non-relational version of the collection database). A text search interface on the database greatly expands access to collection information because it supports keyword search, stemming and relevance ranking. Relevance ranking requires that records in databases be assembled (into records) at the time of indexing, which is why a relational structure can be problematic.

While sophisticated text processing and indexing afforded by search technologies does not replace the need for authority control, or what Murtha Baca calls "vocabulary-assisted search," when implemented properly, text search often does a good job of compensating for the lack of it, for example, mitigating the need for a thesaurus to capture lexical variations (e.g., *amphora, amphorae*; *Impressionist, Impressionism*; *still-life, still life*). Relevance ranked results, partial matching, stemming and phonetic search are popular features of text search engines. In addition, synonym files can be loaded into the search engine to equate "seat" and "chair," "sofa" and "couch," for indexing purposes, further reducing the need for a database thesaurus.

The Getty Research Institute, which manages the Getty's Vocabulary program and developed CCO, has started advocating, as a way of enforcing authority control over a museum catalog, the practice of linking terms to the AAT and other Getty Vocabularies. For an extra fee, the AAT and TGN as a database come bundled into TMS, even though museum collection managers and registrars are typically unsure of how they are to be used—for good reason. The thesaurus's organization doesn't correspond to the fields in TMS's objects module. Because the facets of the AAT do not correspond to the field values in TMS, even if one were to use the AAT as an authority file, terms could still appear in different parts of a catalog record—not good for a relational database which has no keyword search feature. Styles, Periods, Cultures, Movements, Schools are lumped together in the Styles and Periods facet AAT (e.g., "Renaissance Baroque Styles and Periods") in such a way that they cannot very well control the data values in TMS.

It is certainly not possible to link the values in the data entry fields to terms in the AAT as the CCO manual prescribes. It is not all that obvious whether placing terms into TMS attribute field to link terms to their corresponding value in the thesaurus—what the vendor recommends—is worthwhile, when these values would duplicate the values which were already entered in the standard data entry fields, and, because of the post coordinate structure of the AAT, linking would be done to discrete terms (Baroque, painting), not headings (Baroque painting). The AAT is the wrong tool for authority control over a museum catalog.

Authority control is a mechanism for ensuring that when a user performs a query, that all relevant records are retrieved. It includes data normalization (vocabulary control), as well as capturing logical relationships among data values, such that persons, places or things denoted with different terminology--different names, particularly in the event of name changes--may be grouped, and unlike things or different persons named the same way may be differentiated. The easiest way to think about authority control is that it serves to collocate related terms and disambiguate unrelated terms in a database or catalog.

Authority control requires two things: 1. the consistent use of authorized terms when entering data into controlled fields. Search strategies are predictable, because a concept is always designated using the same terms and entered into the same place in a record. This is called vocabulary control. Authority control also requires a method for providing end-user access to records with significant relationships to each other. Library catalogs accomplish this though authority records whose purpose is to redirect users from a deprecated or unauthorized terms to the preferred term or heading (see), and to capture implicit relationships among controlled headings (see also). The relationships or cross references among terms is sometimes referred to as the "syndetic" structure of a catalog.

As mentioned above, CCO recommends using the AAT, Art and Architecture Thesaurus, for authority control of terms. In various illustrations, terms are actually hyperlinked to permanent IDs in the AAT. However, the AAT was from its inception a thesaurus of indexing terms whose structure was designed to assist indexers locate terms used to describe art and other cultural objects. It was conceived at a time when indexing of publications in art history was not done by machines (today database vendors Gale, Ebsco, ProQuest and Wilson do the indexing of scholarly publications for us), but by librarians in the library. It is a thesaurus consisting largely of lexical and linguistic variants whose hierarchical structure does not capture relationships which exist among terms residing in different parts of the thesaurus.

The AAT began in the late 1970s to try to create a system for indexing periodicals and monographs in the field of art history. The project director, Toni Petersen, attempted to restructure relevant parts of the Library of Congress Subject Headings into a hierarchical thesaurus. This hierarchical arrangement was thought to make it easier for indexers to find suitable terminology. Explaining his thought process at the time, Petersen says . . . "as librarians move closer to recognizing the need for chapter level indexing of monographic literature, the importance of a system which provides varying levels of specificity for both the indexing of periodicals and monographs was very much on our minds."[30]

By forcing LCSH into this hierarchical rather than an alphabetical arrangement, certain modifications were made to the data, including removing LCSH's syndetic structure. It was admittedly an experiment, Petersen says, to see if the design of the thesaurus, its hierarchical structure, could to a significant extent replace cross references--the see and see also's which makes LCSH so useful for authority control over a library catalog. This assumption proved, by Petersen's own admission, to be incorrect:

> We saw that we would have to deal with the main term as an entity in itself, and that where it came out in its hierarchical array would determine the narrower and broader term structure. Presumably, if LCSH's syndetic structure were perfect, the *see alsos* given would naturally turn up as related or narrower terms and *see also froms* as broader or related terms. Unfortunately, it doesn't work out that way at all" . . ."The variety of types of references present as *see alsos* precludes any sort of machine conversion . . .and LCSH's syndetic structure fell like a house of cards."[31]

Petersen admits that the hierarchical format could not capture the complex relationships of LC's syndetic structure; but this was not viewed as a serious drawback because the AAT was intended as an indexing tool, *not* an authority file. LC subject headings were broken down into discrete terms to better fit into the themes of the facets and conserve space for printed publication.

[30] Toni Petersen. "The AAT: A Model for the Restructuring of the LCSH." *The Journal of Academic Librarianship*, vol. 9, no. 4, 1983, p. 208.
[31] Petersen, p. 209.

Rather than expressing single concepts, terms in the AAT were meant to be combined. Petersen illustrates this point in the preface of the AAT:

> "For example, *Wooden doors* is an LCSH heading, as is *Renaissance painting*. In the AAT, because of its faceted structure, *wood* is found in the Materials hierarchy, *doors* in the Built Works Components hierarchy, *Renaissance* in the Styles and Periods hierarchy. Indexers are free to use terms separately or to combine them into headings that they pre-coordinate at the time of indexing to match the item they are describing."[32]

This is why the AAT is an inappropriate tool to use for authority control over a museum catalog. Linking to "Baroque" and linking to "painting" does nothing for authority control for the concept of "Baroque painting." The fact that terms in the AAT are meant to be combined to create more specific terms (e.g., *Baroque + painting*), makes it problematic even as a controlled vocabulary, which is the foundation for a controlled system. Due to AAT's faceted structure, the necessity for terms to be put together by the cataloger to form a complete concept (like Greek vase, Japanese print, or Native American art), it has limited ability to provide control over concepts.

The early editors of the AAT intended for noun phrases to be created or put together by catalogers at the time of doing data entry.[33] There is no descriptor "Ionic column" in the AAT, for example; it has no place in the hierarchies. One cannot link directly to it as a concept or authorized term comprised of "Ionic" + "column." If terms are combined by the cataloger to create a complete picture of what a thing is, then *there is no authority control.*

The control afforded by the AAT is limited to terms—particularly alternative spellings and synonyms—not to concepts. There are few cross references to related or broader subjects. Even with all authority controlled fields linked to the AAT, as CCO recommends, and the capability of broadening the search by moving up the AAT—a feature not supported by TMS , nor required by CCO—it would still be impossible to come up with the record for a "Panathenaic amphora" by searching on the terms "Greek vase."

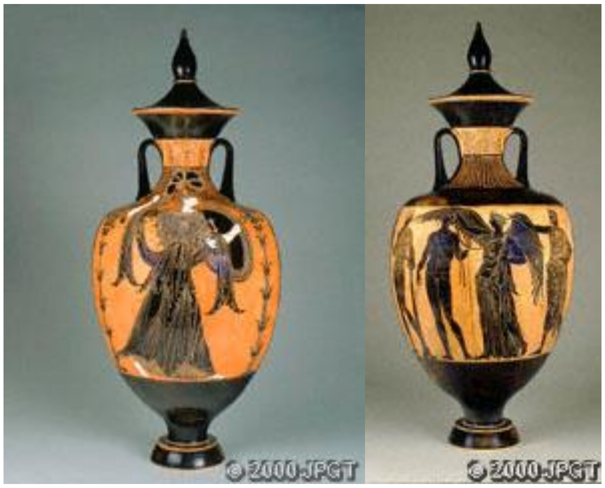
The following record, which is a published example from the Getty[34] of how one should catalog a Greek vase according to CDWA/CCO, serves to illustrate one thing: when it comes to a *Panathenaic amphora*, we all refer to this object as a **"Greek vase**," and yet it is nowhere to be found in the record or the AAT:

[32] "History of the AAT," *Art & Architecture Thesaurus*, Volume I, Oxford University Press, 1990, p. 6.

[33] "In constructing the AAT it was quickly discovered that much of the terminology within its scope related to object names. It was soon apparent that noun phrases denoting objects of art and architecture often contain adjective (designating material, style, technique, and function, among others) and that these adjective recur in infinite combinations throughout the built environment (*Victorian houses*), material culture (*painted furniture*), and the fine arts (*charcoal drawings*). Rather than enumerate the nearly infinite number of object and subject descriptions needed by thesaurus users, the AAT decided to provide building blocks of these descriptors in a faceted vocabulary. . . ." (AAT, p. 27). As a result of the faceted approach, many common phrases, such as "painted Victorian furniture" are not explicitly listed as thesaurus descriptors since they often contain words that are listed in different facets.

[34] http://www.getty.edu/research/conducting_research/standards/cdwa/examples.html

Greek Vase



Object/Work♦	***Catalog Level:*♦** item ***Type:*♦** Panathenaic amphora	Controlled list Authority
	Components/Parts-Quantity: 1 ***Type:*** amphora ***Components/Parts-Quantity:*** 1 ***Type:*** lid	Controlled format Authority
Classification♦	***Terms:*♦** ceramics Greek and Roman art	Controlled list
Titles or Names♦	***Text:*♦** Prize Vessel from the Athenian Games ***Preference:*** preferred ***Type:*** repository ***Text:*♦** Panathenaic Prize Amphora and Lid ***Preference:*** alternate ***Type:*** former	Free text Controlled list
Creation♦	***Creator Description:*♦** attributed to the Painter of the Wedding Procession as painter (Greek vase painter, active ca. 360s BCE); signed by Nikodemos as potter (Attic potter, active ca. 362 BCE)	Free text
	Qualifier: attributed to ***Identity:*♦** Painter of the Wedding Procession ***Role:*♦** painter ***Identity:*♦** Nikodemos ***Role:*♦** potter	Controlled list Authority Authority
	***Creation Date:*♦** 363/362 BCE ***Earliest:*♦** -0363 ***Latest:*♦** -0362	Free text Controlled format
	Creation Place/Original Location: Athens (Greece)	Authority
Styles/Periods/ Groups/Movements	***Indexing Terms:*** Black-figure Attic	Authority

Measurements♦	***Dimensions Description:***♦ height with lid, 89.5 cm (35 1/4 inches); circumference at shoulder, 115 cm (15 1/16 inches)	Free text
	Value: 89.5 ***Unit:*** cm ***Type:*** height ***Value:*** 115 ***Unit:*** cm ***Type:*** circumference	Controlled format and Controlled lists
Materials and Techniques♦	***Description:***♦ wheel-turned terracotta, sintering	Free text
	Material Name: terracotta ***Technique Names:*** turning sintering vase painting	Authority
Inscriptions/Marks	***Transcription or Description:*** signed by Nikodemos	Free text
Subject Matter♦	***Extent:*** general ***Indexing Terms:***♦ religion/mythology object (utilitarian) ceremonial object ***Extent:*** side A ***Indexing Terms:***♦ Athena Promachos (Greek iconography human female ***Extent:*** side B ***Indexing Terms:***♦ Nike Victor competition human females prize	Authority
Descriptive Note	***Text:*** Side A: Athena Promachos; Side B: Nike Crowning the Victor, with the Judge on the Right and the Defeated Opponent on the Left. The figure of Athena is portrayed in an Archaistic style. The particular use of Nike figures atop the akanthos columns flanking Athena allow scholars to date this vase to precisely 363/362 BCE. The Panathenaia, a state religious festival, honored Athena; the festival included athletic, musical, and other competitions. Amphorae filled with oil pressed from olives from the sacred trees of Athena were given as prizes in the Panathenaic Games.	Free text
	Citations: J. Paul Getty Museum online ***Page:*** accessed 10 February 2005	Authority Free text
Current Location♦	***Repository Name/Geographic Location:***♦ Getty Villa Malibu, J. Paul Getty Museum (Los Angeles, California, United States) ***Repository Numbers:***♦ 93.AE.55	Authority Free text
Image Credits: *Prize Vessel from the Athenian Games* (terracotta), 363/362 BCE. Attributed to the Painter of the Wedding Procession (as painter); signed by Nikodemos (as potter) ; artist: unknown Athenian; terracotta; J. Paul Getty Museum (Los Angeles, California), 93.AE.55. © The J. Paul Getty Trust. All Rights		

Searching in the AAT, one finds only variant forms of the term *Panathenic amphora*:

Terms:

Panathenaic amphorae (**preferred**,C,D,U,English-P)
Panathenaic amphora (C,AD,U,English)
Panathenaic amphorai (C,UF,U,English)
amphorae (Panathenaic amphorae) (C,UF,U,English)
amphorae, Panathenaic (C,UF,U,English)
amphorae, type c neck (C,UF,U,English)
amphorae, type IIc (C,UF,U,English)
amphorae type IIc (C,UF,U,English)
amphorai type IIc (C,UF,U,English)
neck amphorae type c (C,UF,U,English)
neck amphorai type c (C,UF,U,English)
panathenaic amphorae (C,UF,U,English)
Panathenaic amphoras (C,UF,U,English)
type c neck amphorae (C,UF,U,English)
type IIc amphorae (C,UF,U,English)
ánforas panatenaica (C,D,U,Spanish-P)
ánfora panatenaica (C,AD,U,Spanish)

This shows the limitations of using the AAT for authority control over a museum catalog. With no reference to “Greek vase” in the record or the AAT, even with the most robust text search engine, how might this object be found in a museum catalog?

Cataloging means not simply identifying what an object is, but anticipating the ways users might be most likely to search for the record. The public doesn’t know the term “amphorae.” They know “Greek vases.” If a search on Greek vase doesn’t produce the record for an amphora, there is a problem with the system, and that problem can be described as lack of authority control. The AAT is a fantastic indexing tool, but it was never intended for the purpose of providing authority control over a museum catalog.

Given its limitations, there is even greater need for additional metadata to be placed into museum catalog records, as a way to create access points, guarantee consistency in the form of headings (Aeolic capitals, Ionic column, etc), and allow people searching the system to find records from a variety of perspectives.

Arguing Over Quiddities: Object Naming (CCO Work Type)

When it comes to cultural artifacts, saying what a thing is is always a matter of interpretation. Object naming, or labeling, often requires an understanding of what features (iconographic, functional, etc.) the object shares in common with other objects, either from the culture of origin or how it relates to other objects in the museum’s collection. Saying what (Latin: *quid*) an object is, for example, whether an object is a “statue,” a “totem” or a “wood carving” can be controversial, which is where the expression “arguing over quiddities”—arguing over the whatness of things—comes from.

This may account for the reason why the Museum of Fine Arts Houston and many other museums using TMS do not use, or use only sparingly, the Object Name field, which is the closest thing to CCO's "Work Type" element. According to CCO, **Work Type** most closely characterizes the work, using the most specific terms possible. It is a required element in CCO. In CCO, Work Type is said to "give logical focus to a Work Record." Work Type is supposed to rely upon a controlled vocabulary, the AAT concept authority records.

The AAT is a thesaurus of descriptors divided into seven facets (Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials and Objects), and 33 hierarchies, which permits terms to be *combined* to form a compound term, e.g.: Ionic (Style Facet) +column (Objects Facet). However, there exist many compound descriptors in the AAT as well.

Deferring to the AAT as an authority does not solve the problems of how to populate the Work Type field. The AAT does not provide subject headings, but indexing terms, a loosely hierarchical mixture of physical attributes, materials, styles, and objects chiefly defined by function. Some terms are post-coordinated or unbound (e.g., "red," "Victorian"), others pre-coordinated or bound ("Panathenaic amphora,""Windsor chair").

In no other field does the faceted structure of the AAT seem to be more at odds with the objectives of descriptive cataloging of art than the Object Name field, where the cataloger is continuously torn between a common sense designation of a thing and maintaining a controlled vocabulary (having to select whatever terms avail themselves to him in the form of noun phrases in the AAT). As mentioned above, with the most commonly occurring adjectives having been sequestered into separate facets, the remaining nouns are more like a functional description (cup, bowl, chair, paintings) than a specific name.

In designing the AAT, editors gathered terms from hundreds of sources. Noun phrases containing common adjectives were broken apart to save space, so that the objects in the objects facet are often too generic and, if applied generically, would tend to overly emphasize (by the way the hierarchy is organized) a utilitarian or functional purpose, even for objects which might have played a more ceremonial, symbolic or decorative purpose. As one example, the richly painted *Panathenaic amphorae*, which were awarded as prizes in athletic competitions, are contextualized in the AAT as (utilitarian) storage vessels, containers—and not as "vases," denoting a more ornamental function. When nouns phrases referring to art and cultural artifacts are stripped of their adjectives, objectification takes place. A "Victorian painted chair" is intellectually something very different thing from a "chair," but "chair" would be the name or best characterization of the object according to CCO. If so many objects would end up with generic object names, what is to be gained by making this a required, authority controlled field, as CCO prescribes? Why isn't Work Type just another narrower Classification?

The AAT is not intended to provide authority control for a collection of catalog records, but rather serves to help catalogers and indexers to locate and select the most appropriate terms to describe objects. In fact, AAT discourages catalogers from being *overly influenced* by the thesaurus's faceted structure for term selection, promoting combination of terms from separate facets to achieve greater accuracy. Authority control can be said to extend only to a term in the very limited context of its own hierarchy, but not to combined terms as they might appear in the Work Type field of the museum catalog.

CCO is frustratingly vague on how to approach what is, arguably, the most important element, Object Name, using the AAT. In one place it implies that compound terms could be used for Work Type provided that *each term* is linked to the AAT.[35] It makes no sense to separately link each of the combined terms back to their respective authority files, in the way CCO illustrates, because authority control does not extend to terms in combination with other terms.

An alternative is to create new terms from existing terms and add them to a local thesaurus, or local copy of the AAT. But how could combined terms be reconciled with the existing structure of the AAT? In all likelihood the new compound term would not fit very well into the pre-existing hierarchies, which is the reason it isn't there to begin with.

If cataloging policy is set that only a single descriptor will be selected from the AAT for Work Type, that is, use AAT only in a pre-coordinated fashion, the consequence will be that objects in the catalog will be named inconsistently or too generically (e.g., "column," instead of "Ionic column"), which will have broad implications for search. Users will not be able to anticipate, how an object might be named. The standard advises using the most specific terms possible to describe the object, permitting broader access through additional layers of classification (via the Class and other elements):

Examples for Work Type:
canopic jar, not container
scroll painting, not painting
engraving, not print
prayer book, not book
obi, not Japanese sash
albumen print, not print
Other examples: lithograph, altarpiece, decoupage, watercolor, Attic helmet

Neither AAT nor CCO instruct catalogers to use the terms in a pre-coordinate or post-coordinate way, or in how to combine the terms, for example, if one should say "Japanese print" or "Japanese woodblock print" or "woodblock print" or "Japanese woodcut" or just "woodcut."

CCO recommends only:

1. that the most specific terms be used, reserving broader categories for Class;

[35] "In cataloging, it may be necessary to combine discrete terms into compound terms. Combining compound terms in free-text fields for display in the Work and Image Record is recommended. . . Some institutions may not have free-text fields, and thus may need to combine the terms in the Concept Authority into compound terms in the Work Record. If so, ideally each part of the phrase, such as red silk in the materials field, should retain its original links to the discrete parts of the concept authority." CCO, p. 334.

2. that data values for Work Type be taken from a controlled list such as the AAT Concept authorities.

3. The values in Work Type be linked to the AAT or some other authority file. If more than one term is used, they should each be linked to the authority file.

While TMS does not permit multiple values in its default Data Entry Form, the CCO standard actually permits multiple Work Types to be added to an object record, as in the following example:

```
cdwalite:objectWorkTypeWrap>
        <cdwalite:objectWorkType termsource="AAT"
termsourceID="aat300127141">cartes-de-
        visite </cdwalite:objectWorkType>

        <cdwalite:objectWorkType termsource="AAT"
termsourceID="aat300265164">boudoir
        photographs </cdwalite:objectWorkType>
</cdwalite:objectWorkTypeWrap>
```

There are many compound, pre-coordinated, or bound terms in the AAT, for example, “Panathathenaic amphora,”“albumen prints,” “boudoir prints,” and “cartes-de-visite.”

However, there is no such authority record for a Japanese woodcut, Chinese scroll, Greek vase, or Ionic column; there are only prints done in Japanese style, or columns which reflect the Ionic order. Searching the AAT Online on “Greek vase” will not produce “Panathenaic amphorae,” for example, and this is precisely the way our own AAT authority controlled catalog would work without customization to facilitate more inclusive searching. Querying the AAT Online provides indication of what might result in a productive search in our own catalog if AAT is used “out of the box,” i.e., in a post-coordinated fashion.

If we choose to use the Object Name field for Work Type, we will either need to edit our local AAT to include hundreds of search equivalent terms, or else compromise the way we name objects so that names can only be what resides in a single facet of the AAT. This is what the original editors of the AAT advised against, letting the structure of the AAT overly influence the way one approaches a work.

Linking to the AAT for Work Type, as CCO recommends, presents a number of challenges, the most significant being that the terms in the AAT can be, and are intended to be, *combined* to form a more complete description of what a thing is: Ionic + columns, artists + diaries, etc. Yet linking to two or more descriptors to their respective sections to accommodate authority control seems wrong: not only is it labor intensive to link the discrete terms, but one should be performing authority control on the *common way of designating the thing as a whole*, not on separate terms. It is a waste of time to have combined descriptors linked to different parts of the thesaurus.

If the style Ionic is deprecated and changed to Ionian, for example, one would still call an "Ionic column" and "Ionic column," and not an "Ionian column." If "Oriental" were deprecated, for "Asian," one would still call an Oriental rug and Oriental rug (actually Oriental rug is an elemental descriptor in AAT). It is truly impossible to rely upon AAT for authority control using a combination of terms drawn from different facets. AAT's authority control extends only to elemental descriptors in the context from which they are drawn, and does not carry over to the objects named by them in combination.

It does not seem to be beneficial, from the standpoint of search, to provide thesaurus linking on multiple discrete terms because one cannot leverage the hierarchical structure of the thesaurus in any meaningful way for query expansion. "Greek" + "vase" represents a particular kind of object, but "Greek" by itself and "vase" by itself are not meaningful, anymore than the terms in "Morris chair" or "albumen print" would be significant if broken apart into discreet terms.

Beyond the Tombstone: Assigning Descriptive Metadata to Works of Art

CCO is a descriptive cataloging standard, not a museum registration standard. Registration is about controlling inventory and recording very basic object information to be used for identification.[36] Cataloging, on the other hand, is all about information retrieval, enhancing access to information through good metadata and authority control. To this end, CCO and CDWA encourage catalogers to add contextual data—style, period, group, movement, school—to records, and to use the AAT for authority control. CCO says: "Record one or more terms that denote the style, historical period, group, movement, or school, whose characteristics are represented in the work being cataloged." [37] It also encourages the addition of subject matter.

The question needs to be asked, where does the museum cataloger obtain the information and the authority within the institutional structure to populate the contextual data fields in a CCO Work Record? An immediate but overly optimistic response would be "the Curators." For a number of reasons, not just lack of time or interest (although these should not be underestimated), there is reluctance among Curators to "burden" objects with art historical terminology and labels. Although of course it varies from person to person, many Curators are concerned with the documentation of objects, and believe that the object record should be *free from bias*. This includes the assignment of contextual information. Compounding the problem is that cultural and historical periodization became particularly unfashionable in the 80s due to academic trends which emphasized the artificiality and implicit biases of all labels. Postmodernism and deconstruction have now become labels in their own right, and yet in their wake many have been left with a heightened sensitivity when it comes to assigning adjectives to cultural objects.

[36] Registration as defined by The New Museum Registration Handbook (1998), is "the process of developing and maintaining as immediate, brief, and permanent means of identifying an object for which the institution has permanently or temporarily assumed responsibility.

[37] CCO, p. 160.

There is a professional divide between Curators and Catalogers in this regard: Curators are concerned with controlling interpretation, and catalogers are concerned with permitting broad access to the information contained in a database. The latter sometimes means adding descriptive metadata to the record and approaching things as would someone unfamiliar with the collection, that discipline's terminology or with that object's history. This also means that how a thing is most commonly known or referenced has to be included in the record, which a Curator might object to because lends legitimacy to something that is not, in their eyes, correct. The way dealers describe pieces to market them, for example, to make items seem more appealing, valuable, approachable or understandable to the public, goes against the grain of a Curator: "There really is no such thing as a *Morris* chair. He never called it that in his catalogs," a Curator informs me, even though *I* call it a Morris chair. Curators may choose not to title a chair a Morris chair on a wall label, but it should be included in the metadata for the record because that is the object's familiar name, just like we record the most familiar name of the artist as the preferred name.

However, a database schema which supports only one Style and one Period for a work, such as TMS, rather than accepting multiple values, only compounds the problem by contributing to the perception that we are all about "labeling" objects in ways Curators do not agree with. TMS's contextual fields are incapable of accepting capturing multiple, authority controlled values: an object cannot be "Rococo" and "Neo-Classical"; "Georgian and Palladian" in TMS. It has to be only one, compromising our efforts to accurately describe a work, assign access points, and expand end-user access as instructed by CCO.

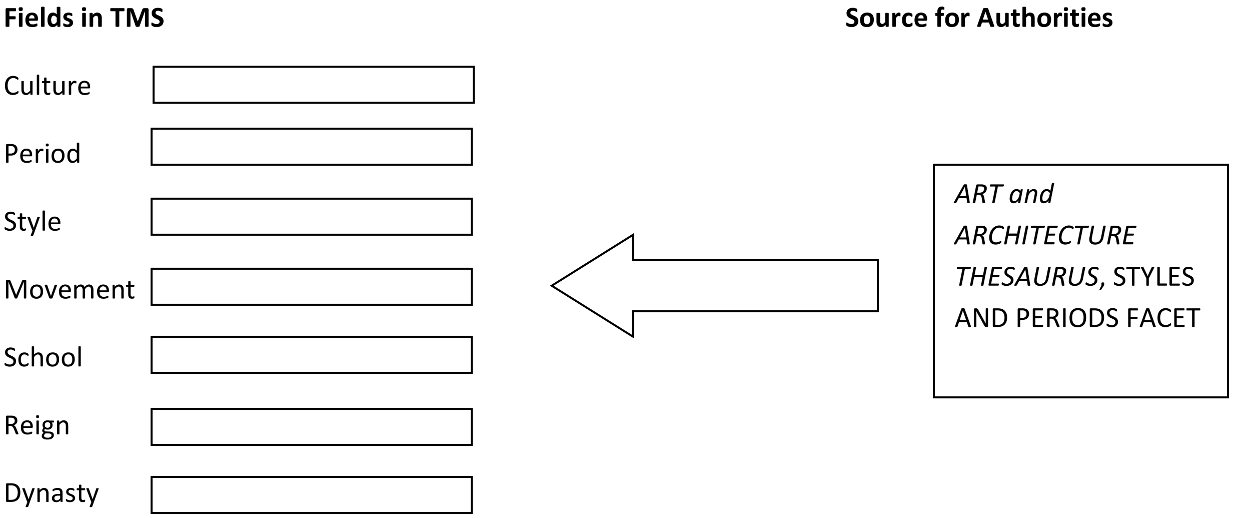
And although CCO recommends using the AAT for authority control over the Period and other contextual data fields, the larger companion standard, CDWA, **explains that Period, Style, Culture Groups, Reign, Dynasty, Movement and School, cannot realistically be treated as separate entities from the standpoint of a field definitions**.[38] In other words, there is no way to differentiate a Style from Period, or a Period from a Movement, or a Style from a Movement, which is why all of these terms are put together in one facet of the AAT under guide terms like: <*Chinese styles and periods*>*,* <*modern North American decorative art styles and movements*>, and other combinations of <*Periods and Cultures*>, <*Style and Period*><*Styles and Movements*>, and why CDWA Lite schema simply puts style and period together into a generic, repeatable element:

Tagging examples:

```
<cdwalite:styleWrap>
<cdwalite:style>Renaissance</cdwalite:style>
</cdwalite:styleWrap>
<cdwalite:styleWrap>
<cdwalite:style termsource="AAT" termsourceID=" aat300021147">Baroque </cdwalite:style>
<cdwalite:style termsource="AAT" termsourceID=" aat 300021080">Louis XIV </cdwalite:style>
</cdwalite:styleWrap>
```

[38] http://www.getty.edu/research/conducting_research/standards/cdwa/definitions.html#styles

When museums are setting up SQL databases like Access for the purpose of cataloging, or writing style manuals for cataloging in TMS, it must be kept in mind that databases typically require **strict field definitions** in order to return good results, particularly if one is limited to using structured search parameters. SQL requires that users select the field where the data can be found, but with a schema like this, there is no way to control the values in these fields because it is not possible to strictly define the data in a consistent or predictable way:


Fields in TMS
Source for Authorities
Culture
Period
Style
Movement
School
Reign
Dynasty
ART and ARCHITECTURE THESAURUS, STYLES AND PERIODS FACET

Originally, many of the terms in the AAT were subject headings in the Library of Congress System, which corresponds to a more flexible data format (MARC record) where one does not have to describe the metadata in order to assign it to a record: all topical subject headings go into a generic, repeatable subject headings field. In library systems, the heading “Arts and Crafts Movement” goes into a 650 field, and there is no need to indicate if the heading reflects a style, or a movement, or both:

```
100 1  Parry, Linda.
245 10 Textiles of the arts and crafts movement /|cLinda Parry.
250    New ed.
260    London :|bThames & Hudson,|cc2005.
300    160 p. :|bcol. ill. ;|c26 cm.
500    Originally published: London : Thames & Hudson, 1988.
504    Includes bibliographical references (p. 156) and index.
650  0 Textile fabrics|zEngland|xHistory|y19th century.
650  0 Arts and crafts movement|zEngland.
```

Taking terms from a designated source does not make records in a database authority controlled; data values also have to be consistently placed in predictable fields in the database so that records can be efficiently retrieved. This is particularly true of relational database applications, which require the user to select the correct field to perform a query. There is no keyword search in a traditional relational database application.

The AAT can help a cataloger to identify terms pertaining to *Renaissance-Baroque styles and periods,* but not how to distinguish a Style from a Period for doing the actual data entry in the fields of a relational database. Authority control has to be applied *per field*, not on the record level. We need to know what terms go into what fields if data is to be efficiently entered and retrieved in a relational database system. This is something the authors of the CCO standard never address, despite their insistence that CCO is compatible with any system, and especially a relational database application.

Assigning Subject Matter to Objects. Many institutions have devised schemes for searching by subject when they move a subset of their collections online, anticipating that this is how the public would most likely want to search the collection. The AAT does not consistently index subject, thematic, or iconographic terminology, and there is no existing vocabulary for providing subject access to works of art. The assignment of subject headings is accomplished more easily in libraries because books, unlike works of art, tend to be self-describing.

Curators may be unwilling to participate in assigning subjects to their objects, because there is a concern that identifying the subject of a work focuses the viewer's attention, inadvertently *frames the interpretation* of the object. Yet, the task of assigning subjects cannot be completely left up to volunteers ("What do you see?") because an informed eye is often required to correctly identify the subject matter of a work of art.

As Erwin Panofsky, in his landmark *Studies in Iconology* illustrates, *The Last Supper* can be 13 men seated around a table, or "a group of men eating dinner," or could be a portrayal of a particular religious and historical event, "The Last Supper" (the iconographic meaning); and then, this particular rendition might have additional layers of *iconological* meaning which can be understood only by someone familiar with the Biblical narrative *and* art historical precedents. Assigning subject matter to works of art, while seemingly straightforward, can become incredibly complex. Like a fly or snail in still life table setting, which can change the subject from a bouquet of flowers to a Vanitas painting, sometimes the smallest details in the picture are loaded with significance which only a trained eye can detect or account for.

Conclusion:

Even if CCO is not being used by museums as a cataloging standard, but as a data exchange standard (for populating CDWA Lite) for contributing records to ARTStor or some other consortium, it is impossible to map data from a museum collection database to XML *if the data isn't there to begin with*. Databases to support museum cataloging need to be able to support the seamless import and export of well-formed CDWA Lite (CCO) records for resource sharing and collaboration. **To achieve this goal, the schema of the database needs to be able to accommodate all of the elements, attributes, and repeating values of the XML schema.**

A fully CCO compliant museum collection management system should be able to import and export CDWA Lite, which means that the collection management system would provide support for :

> multiple classes (art historical traditions, functional classification)
> multiple work types
> multiple styles, period and chronological designations
> multiple subjects (iconographic themes, etc)
> hyperlinks to bibliographic sources and repositories
> dynamically self-updating content through "links" to authority records

Regardless of what cataloging system or standard a museum employs, the AAT is inadequate for authority control over a museum catalog because the terms within it are frequently too generic (cup, bowl, vase, etc), and meant to be combined with other terms in the thesaurus in order to achieve required specificity. To use it for authority control, one's local thesaurus would need to be extensively customized, not only to support local practices and specializations, but also to contain "pre-coordinated" (already combined) AAT terms and to ensure that specificity is not sacrificed for consistency of data.

Using a search engine capable of keyword search and relevance ranking, rather than a Boolean model employed by traditional relational databases, would allow more natural language approaches and require less precision in terms of formatting of data and their placement in different areas of the CCO Work record. Text search engines, which commonly incorporate word stemming techniques (would retrieve "amphora" with "amphorae"), synonym expansion (could retrieve "cups" and "vessels"), and indexes terms not strings (would retrieve both "Greek, Ancient" and "Ancient Greek," Chimu and Chimú, gold leaf and goldleaf), would also reduce dependenct on an external thesaurus (and/or Data Standard) considerably, and on catalogers to spend time capturing lexical variants.

Before we "learn how to share"[39] our records though metadata harvesting or crosswalks, museums first need systems and standards that match up so we can create rich cataloging records that are worthy of sharing in the first place. However, even with a collection management system capable of accepting the kind of metadata recommended by CCO, and even with a perfect set of authorities in place, it remains to be seen whether the type of cataloging which requires the assignment of descriptive, and on some level *interpretive* metadata to art objects, can work in a museum setting, given that the type of cataloging advanced by the CCO standard, while common in libraries, on some levels goes against existing museum practices. Or I should say, that for it to work, there needs to be a clear mandate not

[39] Waibel, Günter, Ralph LeVan, Bruce Washburn, Learning How to Share: Final Report to The Andrew W. Mellon Foundation The published results of a grant funded study in which Patricia Harpring, lead author of the CCO standard, was paid to analyze the CCO-ness of harvested catalog data from nine institutions. http://www.oclc.org/research/publications/library/2010/2010-02.pdf. The fields that were analyzed limited to creator name, creator nationality, title, worktype, display date, dimensions, and medium, rather than trying to apply the entire standard. In conversation with Ms. Harpring, I was mentioning to her the incompatibilities of CCO with TMS, and she responded, "We don't know that yet, the results are not in."

just to make the collection database authoritative, but to make collection records more accessible to people outside the museum's walls.