THE GETTY VOCABULARIES TECHNICAL UPDATE

Joan Cobb
Gregg Garcia
Information Technology Services
J. Paul Getty Trust

International Working Group Meetings
January 7-10, 2013

International Terminology Working Group

August 23-26, 2010

November 7-10, 2011

Monday

Overview of the technical aspects that apply to all four of the vocabularies.

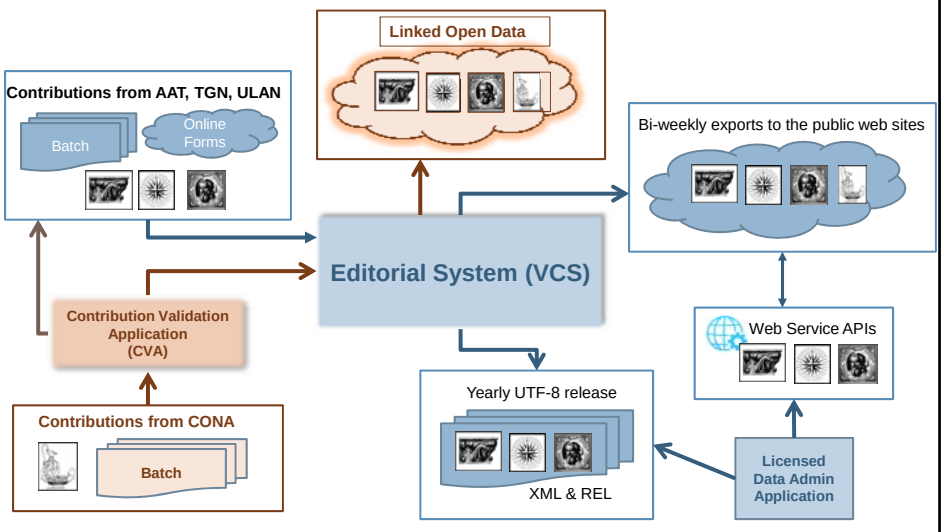
AAT

TGN

ULAN

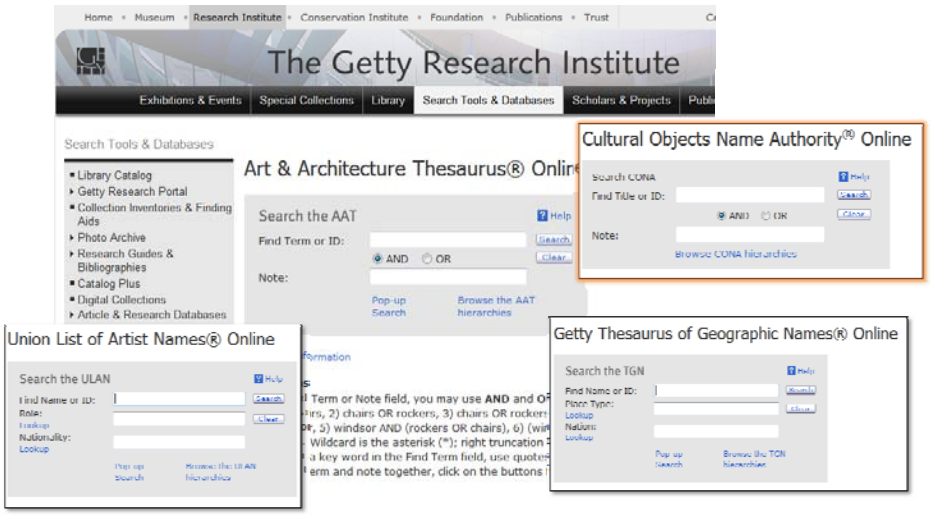
CONA

Technical solutions



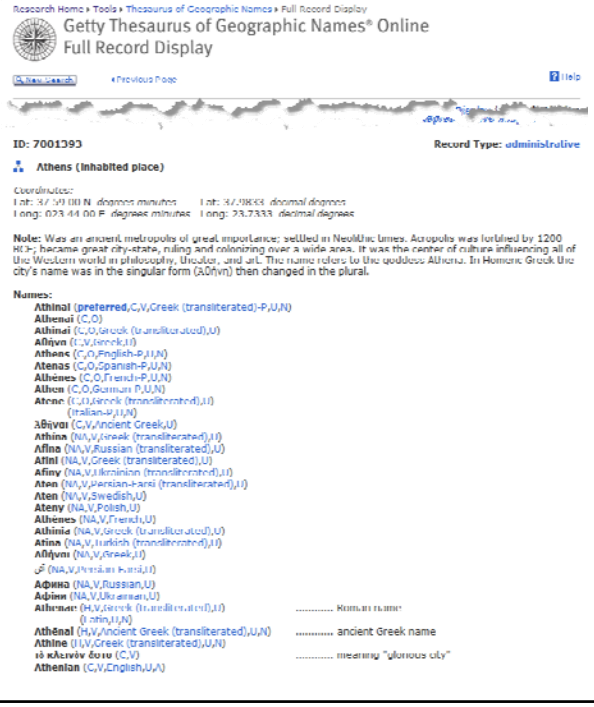
Vocabularies on the Web

Automatically updated bi-weekly



All of our vocabularies are multilingual

William James Stillman; Acropolis, houses of Plaka and the Areopagus, seen from the cathedral of Athens; Gary Edwards collection of photographs of Greece; 1869; Albumen print mounted on red-ruled board; Research Library, The Getty Research Institute (Los Angeles, California); 92.R.84.



Japanese and Hebrew

Library of Congress Authorities online

Research Home ▸ Tools ▸ Union List of Artist Names ▸ Full Record Display

Union List of Artist Names® Online
Full Record Display

New Search | Previous Page | Help

Click the icon to view the hierarchy.

ID: 500060426 **Record Type: Person**

Hokusai, Katsushika (Japanese printmaker and painter, 1760-1849)

Note: He is famous for works of ukiyo-e ('pictures of the floating world') in painting and printmaking. He had a long career of over 70 years in diverse media and styles. He studied painting of Ming period (1368-1644) in China to the styles of the Kanō school, Sumiyoshi school, Rinpa painting, and his contemporaries of the Edo period (1600-1868) in Japan; he also learned from Western style painting. Object types and media included nikuhitsuga (polychrome or ink paintings); surimono ('printed things'; small-edition woodblock prints) and nishikie (polychrome prints); woodblocks for eirihon (illustrated books) and kyōka ehon (illustrated books of poems called kyōka); and printed book illustrations for kibyōshi ('yellow cover' books, often moralizing tales and adventures) and yomihon ('reading books,' sometimes historical novels). He was one of the main shunga (erotic picture) artists of the Edo period. Hokusai is thought to have made in all at least 30,000 drawings and the illustrations for 500 books. Since the late 19th century, his work has had a significant impact on Western artists, including Gauguin and van Gogh.

Names:

Hokusai, Katsushika (**preferred**,V,index)	
Katsushika Hokusai (V,display)	
Hokusai (V)	.... name taken by the artist in 1798
葛飾 北斎 (V,Japanese,NA,U)	
Shunrō (V)	.... gō (artist's name), used in his years of training, when painting hosoban (narrow prints)
Sōri (V)	.... gō (artist's name), used in early career, named taken from his Rinpa-school master Tawaraya Sōri
Kakō (V)	
Tatsumasa (V)	
Gakyōjin (V)	
Taito (V)	.... name used since 1810, when creating illustrated picture books
Iichi (V)	
Manji (V)	
Tokitarō (V,BN)	
時太郎 (V,BN)	
葛飾北斎 (V)	
葛飾北斎 (V)	
葛触北斎 (V)	
葛飾北斎 (V)	
葛飾北斉 (V)	
葛飾北齋 (V)	
葛飾北齋 (V)	
葛飾北斎 (V)	
近藤市太郎 (V)	
הוקוסאי (V)	
קאטושיקה, הוקוסי (V)	

Chinese

Dutch

Spanish

Research Home ▸ Tools ▸ Art & Architecture Thesaurus ▸ Full Record Display

Art & Architecture Thesaurus® Online
Full Record Display

New Search | Previous Page | Help

Click the icon to view the hierarchy.

ID: 300014916 **Record Type: concept**

lacquer (coating) (<coating by composition or origin>, coating (material), ... Materials (Hierarchy Name))

Note: Term applied broadly to paint or varnish that dries with a high gloss, to a coating made from shellac, or to any coating that dries by the evaporation of solvent. In most lacquers resin is dissolved in a volatile solvent or a drying oil and hardens by evaporation.

Terms:

lacquer (coating) (**preferred**,C,U,English-P,D,U,N)
lacker (C,U,English,UF,U,N)
漆 (C,U,Chinese (traditional) P,D,U,U)
漆器 (C,U,Chinese (traditional),AD,U,U)
塗漆 (C,U,Chinese (traditional),AD,U,U)
人然漆 (C,U,Chinese (traditional),AD,U,U)
大漆 (C,U,Chinese (traditional),AD,U,U)
qī (C,U,Chinese (transliterated Hanyu Pinyin) P,D,U,U)
qi (C,U,Chinese (transliterated Pinyin without tones)-P,D,U,U)
ch'i (C,U,Chinese (transliterated Wade-Giles)-P,D,U,U)
lak (C,U,Dutch-P,D,U,U)
laque (C,U,French P,D,U,N)
lacca (vernice) (C,U,Italian-P,D,U,N)
laca (barniz) (C,U,Spanish-P,D,U,N)

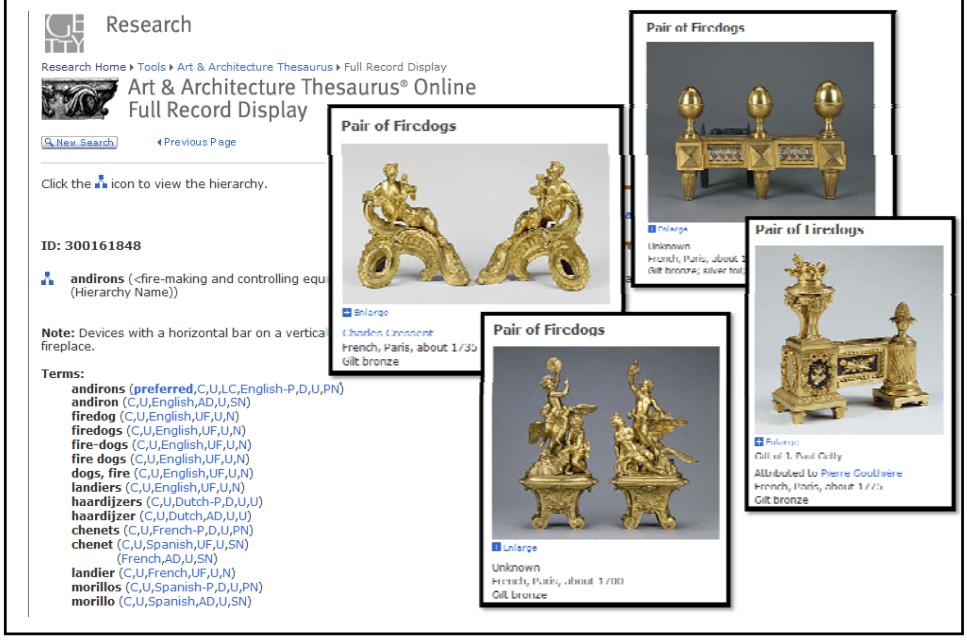
Additional Notes:

Chinese (traditional) ..... 此詞彙廣泛用以指塗佈於表面光滑的物件或表面，以產生有光澤的保護膜，或由精溶劑揮發而乾燥的物件塗層。在人造樹脂出現中，植物學家曾經將此詞彙試驗過中，然後由經驗由經驗。

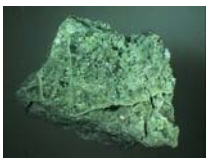
Dutch ..... Algemene term voor verf of vernis die met een hoogglans opdroogt en voor schellak of iedere andere deklaag die droogt door verdamping van het oplosmiddel. In de meeste laksoorten is hars opgelost in een vluchtig oplosmiddel of drogende olie, waarbij het hars verhardt door verdamping.

Spanish ..... Término aplicado ampliamente a la pintura o barniz que se seca y alcanza considerable brillo, para el revestimiento fabricado con goma laca o para cualquier revestimiento que se endurece por evaporación de su solvente.

Ability to Associate Images



Multiple Parents



ID: 300011121 Record Type: concept

jadeite (mineral, inorganic material, ... Materials (Hierarchy Name))

Note: Mineral consisting of sodium aluminum silicate, usually containing small amounts of iron, calcium, and magnesium. It constitutes the more valuable variety of jade, ranging in color from pale to dark green that is commonly called jade; nephrite is the other mineral called "jade." Jadeite is highly valued for its translucent glassy appearance; it is found chiefly in China, Myanmar (formerly Burma), and Mexico.

Terms:
jadeite (preferred,C,U,LC,English-P,D,U,N)
Imperial jade (C,U,English,UF,U,N)
硬玉 (C,U,Chinese (traditional)-P,D,U,U)
翡玉 (C,U,Chinese (traditional),AD,U,U)
玉翠 (C,U,Chinese (traditional),AD,U,U)
ying yü (C,U,Chinese (transliterated Hanyu Pinyin)-P,D,U,U)
ying yu (C,U,Chinese (transliterated Pinyin without tones)-P,D,U,U)
ying yü (C,U,Chinese (transliterated Wade-Giles)-P,D,U,U)
jadeïet (C,U,Dutch-P,D,U,U)
jadeiet (C,U,Dutch,UF,U,N)
jadéite (C,U,French-P,D,U,N)
Jadeit (C,U,German-P,D,U,N)
jadeita (C,U,Spanish-P,D,U,U)

Facet/Hierarchy Code: M,MT

Hierarchical Position:
Materials Facet
.... Materials (Hierarchy Name) (G)
..... materials (matter) (G)
..... <materials by composition> (G)
..... inorganic material (G)
..... mineral (G)
..... jadeite (G)

mineral

Additional Parents:
Materials Facet
.... Materials (Hierarchy Name) (G)
..... materials (matter) (G)
..... <materials by composition> (G)
..... inorganic material (G)
..... rock (G)
..... <rock by form> (G)
..... stone (rock) (G)
..... <stone by quality> (G)
..... gemstone (G)
..... <gemstones by specific type> (G)
..... jadeite (G)

gemstone

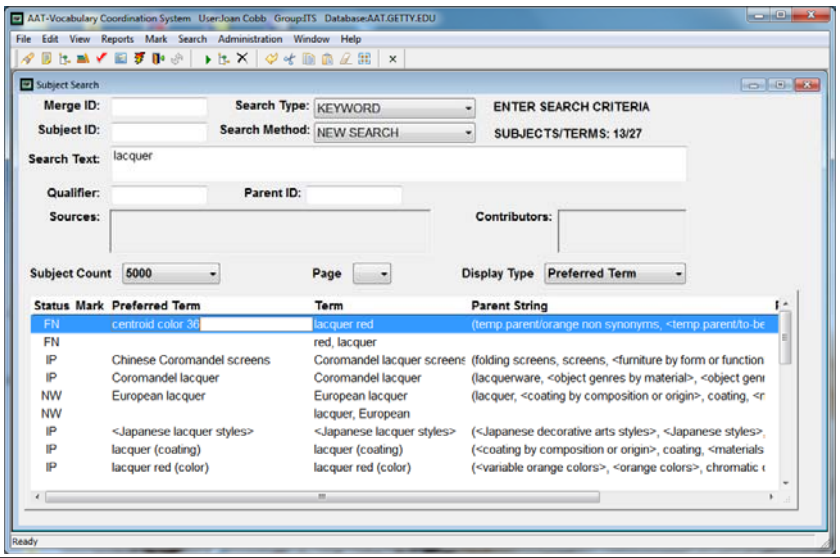
Materials Facet
.... Materials (Hierarchy Name) (G)
..... materials (matter) (G)
..... <materials by composition> (G)
..... inorganic material (G)
..... rock (G)
..... metamorphic rock (G)
..... jade (rock) (G)
..... jadeite (G)

rock

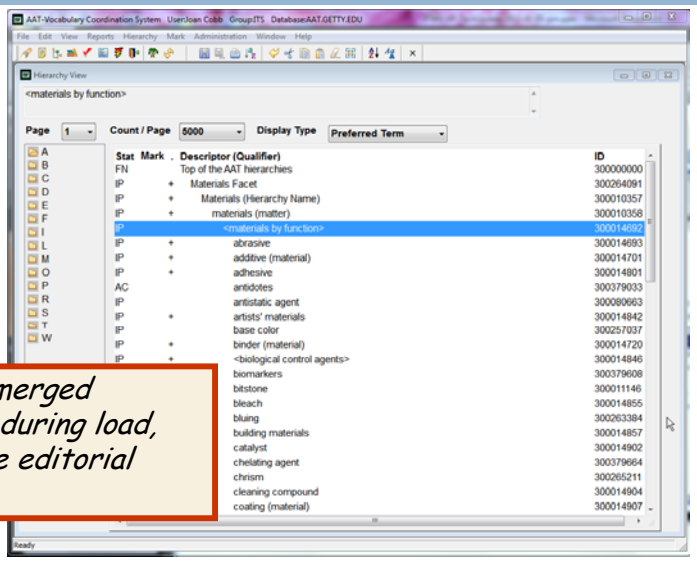
Ability to Track Contributors and Sources

Editorial System

Subject Search

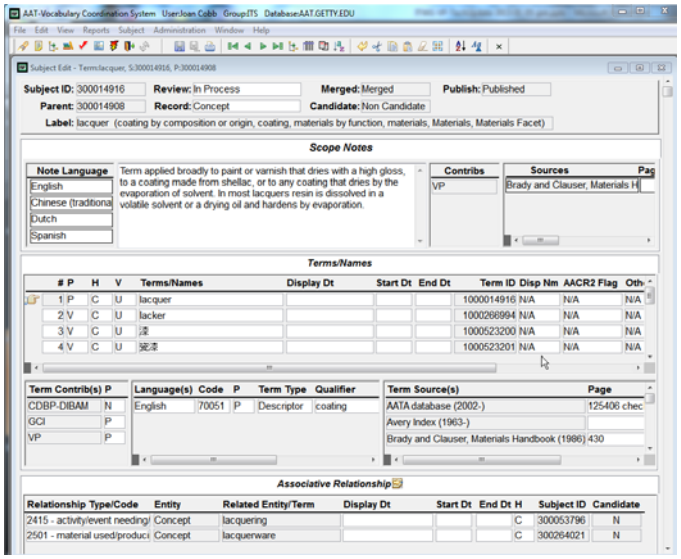


Editorial System -
Hierarchcy View



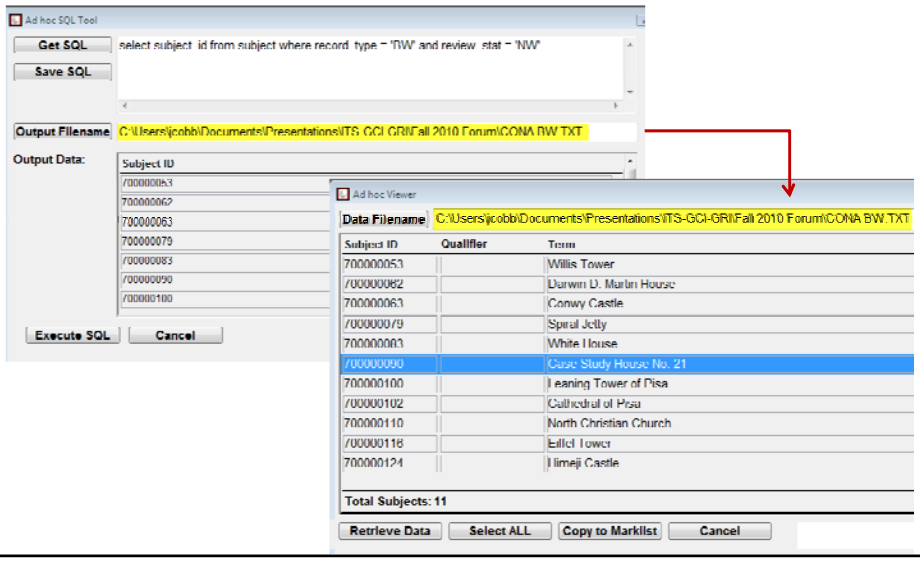
Some can be merged automatically during load, others require editorial intervention.

Editorial System
Top part of a full record data entry screen



Editorial System

Ad hock query tool



Batch loading of data

- Information about how to contribute is available on our Web site

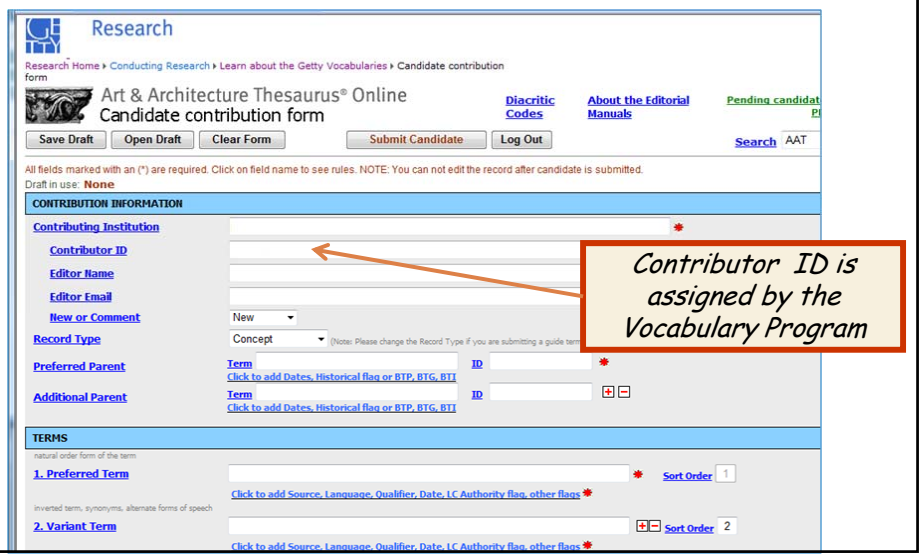
  http://www.getty.edu/research/tools/vocabularies/contribute.html

- Bulk contributions
  - Can range from a few thousand to several million records
  - Every new contribution requires close coordination
- Online Web candidate contribution forms
  - URL: vocabcontrib.getty.edu
  - Uses the same XML schema as bulk contributions
  - Harvested on the 1st of every month

Contribution XML schemas

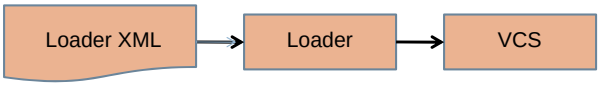
- Basic Schemas
  - Getty_Vocabulary_AAT.xsd
  - Getty_Vocabulary_TGN.xsd
  - Getty_Vocabulary_ULAN.xsd
  - Getty_Vocabulary_CONA.xsd
- Schemas containing specific enumeration values are updated every month
  - AAT : associative_type, contrib, language, lookup_lists, source
  - TGN: associative_type, contrib, language, lookup_lists, source plus place_type
  - ULAN: associative_type, contrib, language, lookup_lists, source plus event, nationality, place, and role
  - CONA: associative_type, contrib, language, lookup_lists, source plus event, nationality, role, and class

Online contribution forms for AAT, ULAN & TGN



Loading the data

- Programs to load the data are very complex
  - Several rounds of testing are necessary
  - Actual load takes from several hours to several days



- Merging
  - Rules are different for each vocabulary
  - Need to accommodate differing standards
  - Some of the merge routines can be skipped with translations because the matching records are already known
- Unicode
  - Contributions can be submitted in any encoding

Loading the data

Multi-lingual data

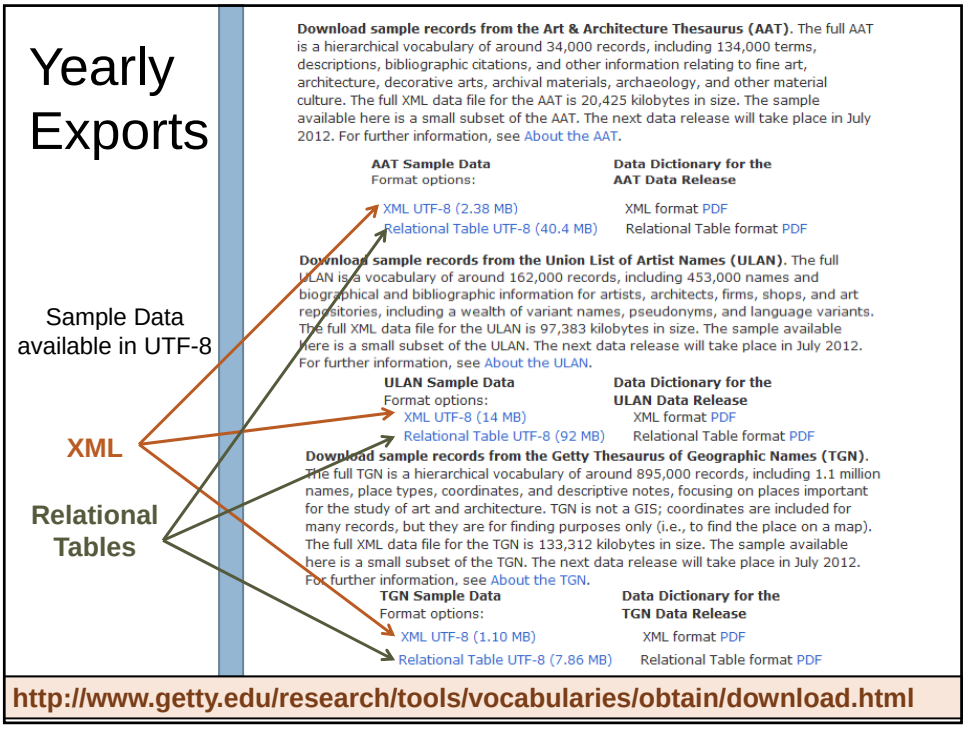
- Worked through a number of issues related to ordering and classifying terms in different languages
- Adjusted the data structure to make the term qualifier, term type and part of speech language specific
- Modified the rules combining terms
- Updated the way the information is presented to users
- Changes to the core editorial system require changes to everything else

Yearly Exports

Sample Data available in UTF-8

XML

Relational Tables



Download sample records from the Art & Architecture Thesaurus (AAT). The full AAT is a hierarchical vocabulary of around 34,000 records, including 134,000 terms, descriptions, bibliographic citations, and other information relating to fine art, architecture, decorative arts, archival materials, archaeology, and other material culture. The full XML data file for the AAT is 20,425 kilobytes in size. The sample available here is a small subset of the AAT. The next data release will take place in July 2012. For further information, see About the AAT.

AAT Sample Data
Format options:
XML UTF-8 (2.38 MB)
Relational Table UTF-8 (40.4 MB)

Data Dictionary for the AAT Data Release
XML format PDF
Relational Table format PDF

Download sample records from the Union List of Artist Names (ULAN). The full ULAN is a vocabulary of around 162,000 records, including 453,000 names and biographical and bibliographic information for artists, architects, firms, shops, and art repositories, including a wealth of variant names, pseudonyms, and language variants. The full XML data file for the ULAN is 97,383 kilobytes in size. The sample available here is a small subset of the ULAN. The next data release will take place in July 2012. For further information, see About the ULAN.

ULAN Sample Data
Format options:
XML UTF-8 (14 MB)
Relational Table UTF-8 (92 MB)

Data Dictionary for the ULAN Data Release
XML format PDF
Relational Table format PDF

Download sample records from the Getty Thesaurus of Geographic Names (TGN). The full TGN is a hierarchical vocabulary of around 895,000 records, including 1.1 million names, place types, coordinates, and descriptive notes, focusing on places important for the study of art and architecture. TGN is not a GIS; coordinates are included for many records, but they are for finding purposes only (i.e., to find the place on a map). The full XML data file for the TGN is 133,312 kilobytes in size. The sample available here is a small subset of the TGN. The next data release will take place in July 2012. For further information, see About the TGN.

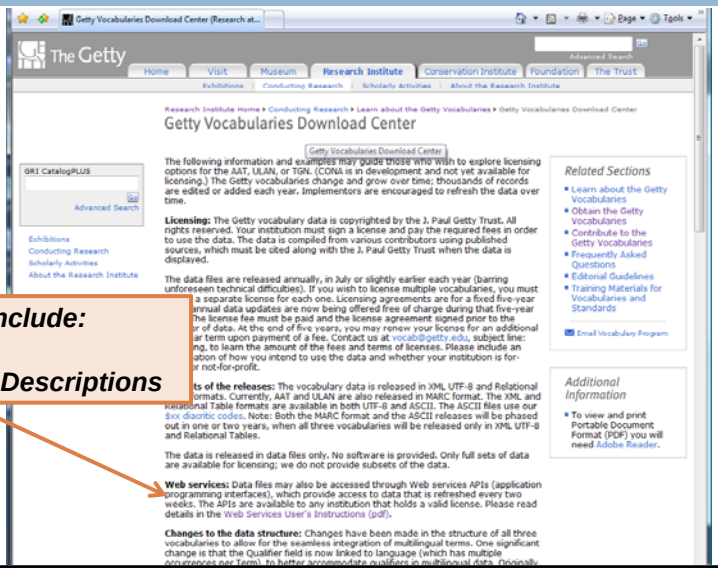
TGN Sample Data
Format options:
XML UTF-8 (1.10 MB)
Relational Table UTF-8 (7.86 MB)

Data Dictionary for the TGN Data Release
XML format PDF
Relational Table format PDF

http://www.getty.edu/research/tools/vocabularies/obtain/download.html

Web Services - User's Instructions

http://www.getty.edu/research/tools/vocabularies/obtain/download.html



Instructions include:
Examples
Data Element Descriptions

Web Services

Access to basic information

- Information that is available on our public web site is also available via our Web Services
  - Term Match – uses the same criteria
    - AAT – Term and Note
    - TGN – Name, Place Type, Nation
    - ULAN – Name, Role, Nationality
  - Full record for a particular subject ID
  - Preferred and non-preferred parent hierarchies for a particular subject ID
  - Immediate preferred and non-preferred children for a particular subject ID

Web Service

Ability to synchronize

- List of records that have been merged along with their old and new IDs
- Current ID for a given subject ID, accounting for whether or not a record has been merged
- List of records that have been modified based on a date range and one of the following parameters
  - Overall subject record edits
  - Added, deleted, modified terms
  - Scope note edits
  - Moved records
  - New records

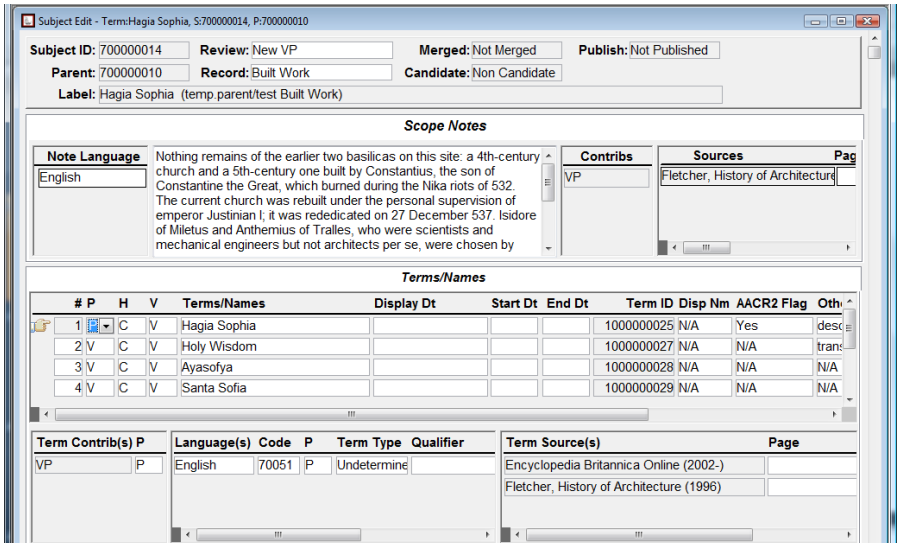
Tuesday

CONA

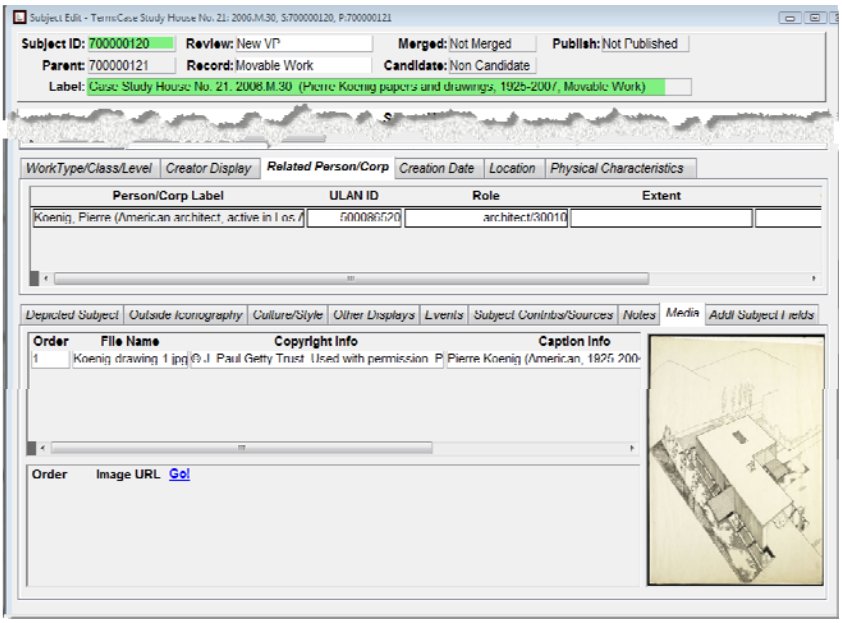
Overview of the technical aspects that are particular to the Cultural Objects Name Authority®

Editorial System

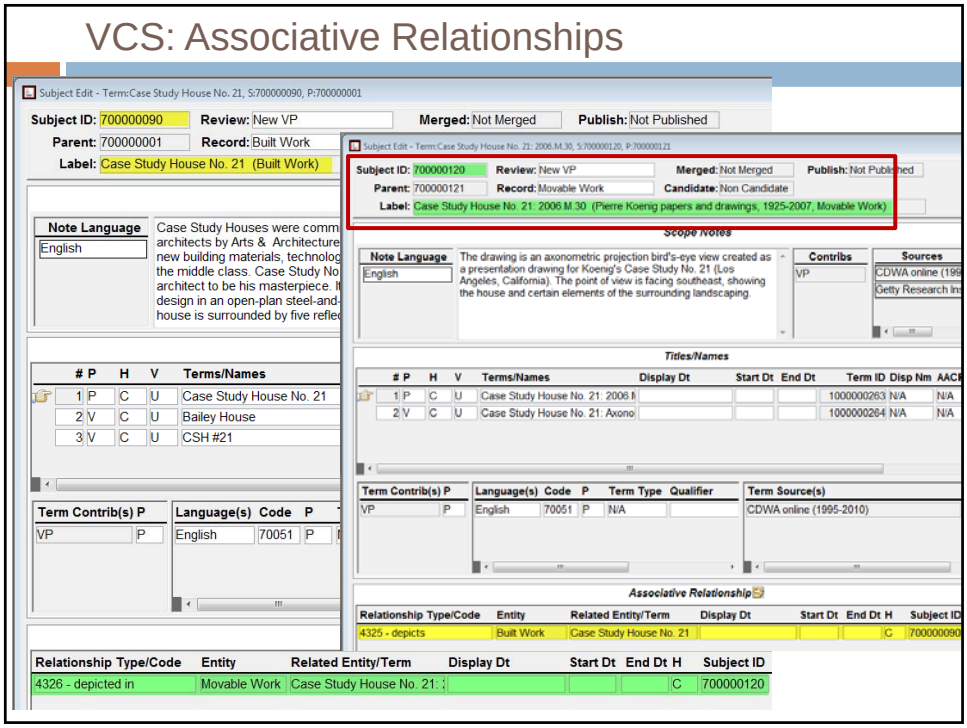
Full record data entry screen



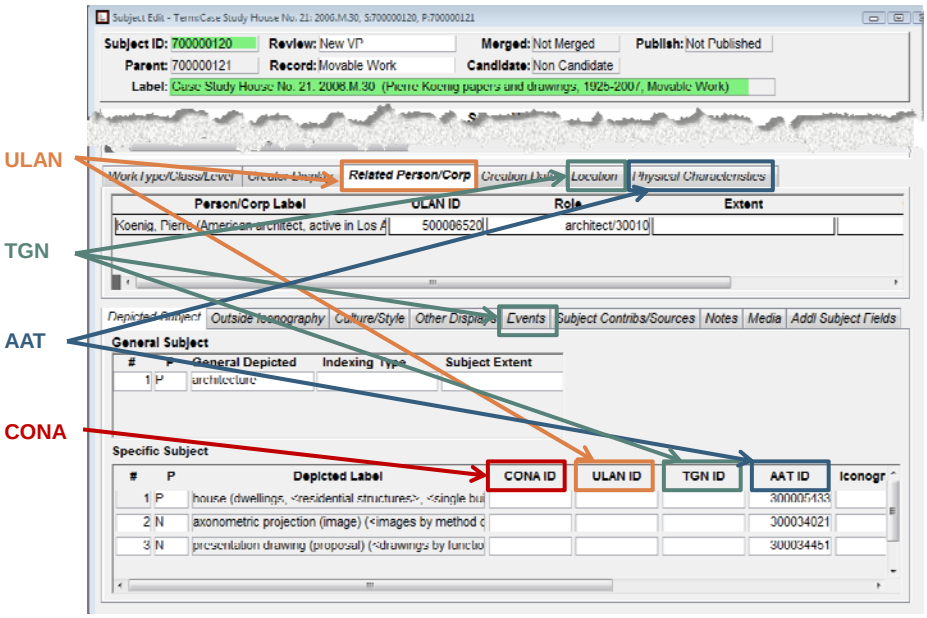
CONA-VCS: Associate an image



VCS: Associative Relationships



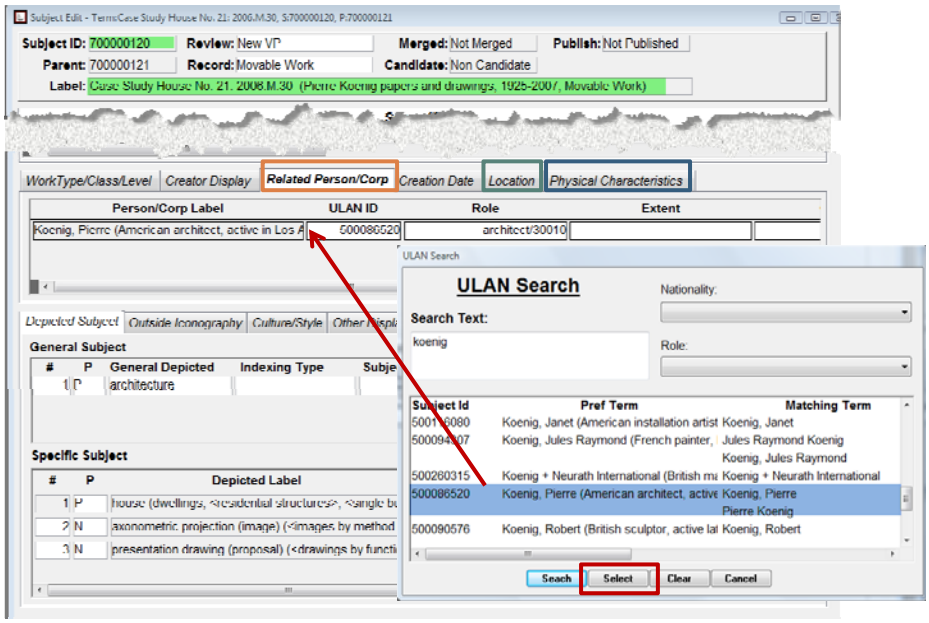
CONA-VCS: Live links to AAT, TGN & ULAN



Examples of New Web Services for In-house Reports and CONA CVA

- **GetChildCount** – returns an indexing alpha character along with the count of the number of subject records with a preferred term starting with that letter
- **GetChildrenReports** – returns basic child record information given parent ID, alpha index character, language code and candidate status
- **GetMergedSubjectIDsReports** – returns the count, subject IDs and merge dates of all records merged into the given current subject ID
- **GetMergedSubjectReports** – returns full record information for a merged subject record
- **GetParentHierarchy** – returns total parent levels along with pertinent parent/child information for each level
- **GetSourcesReports** – returns full source information based on the date of edits or creation, alpha index, and editor
- **GetSubjectBriefReports** – returns a limited set of full record data for a current subject ID
- **GetSubjectIDsDiacritics** – returns a list of subject IDs given a comma delimited list of Getty diacritic codes; also allows user to input a hierarchy parent ID and candidate status
- **GetSubjectIDsReports** – returns a list of subject IDs given a broad range of criteria: creation/edit date, alpha indexing character, editor, candidate status, parent hierarchy ID, and contributor
- **GetSubjectReports** – returns full record information given a current subject ID
- **CreateSubject** – creates a candidate record for a given preferred term and parent ID
- **TGNGetSubjectIdsCoordinates** – returns a list of subject IDs given a set of coordinate points, parent key and candidate status
- **CONAIconographyMatch** – term search matching service for the Iconography module within CONA
- **CONACreateIconography** – creates a candidate Iconography record given a preferred term

CONA-VCS: Search ULAN-VCS

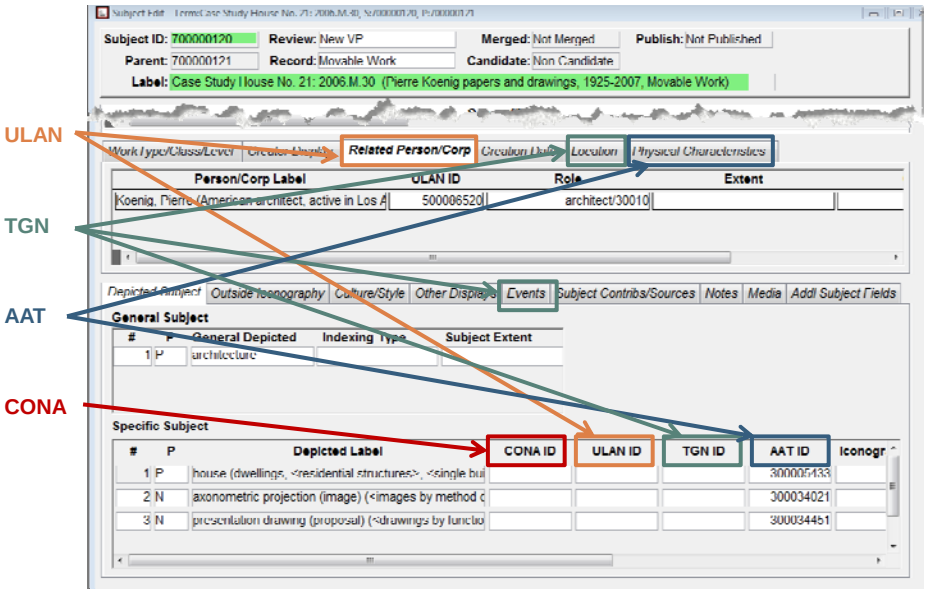


Loading the data

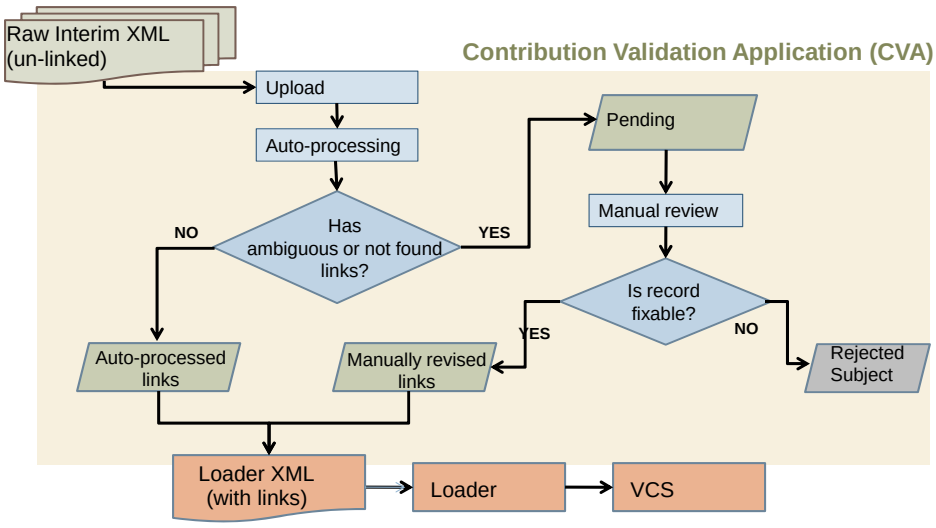
- Programs to load the data are very complex
  - Several rounds of testing are necessary
  - Actual load takes from several hours to several days

- Merging
  - Rules are different for each vocabulary
  - Need to accommodate differing standards
  - Some of the merge routines can be skipped with translations because the matching records are already known
- Unicode
  - Contributions can be submitted in any encoding

A CONA contribution may also be a contribution to AAT, TGN & ULAN

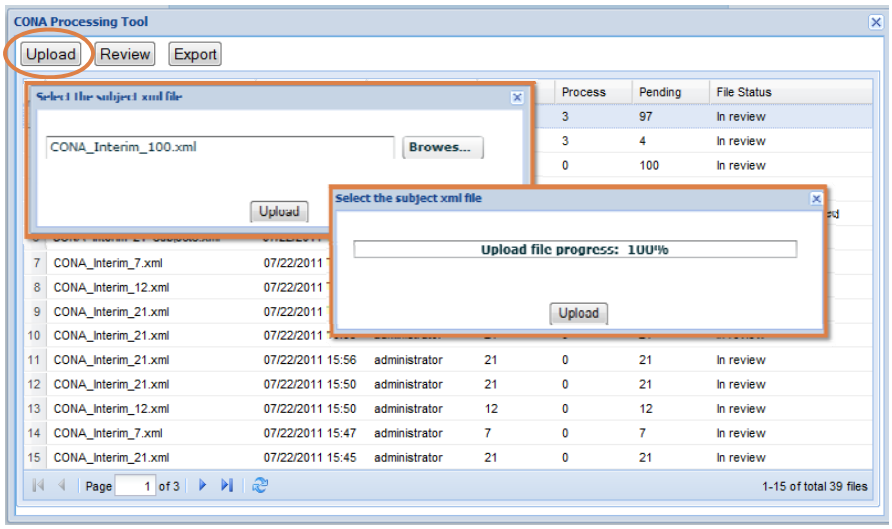


Loading CONA data



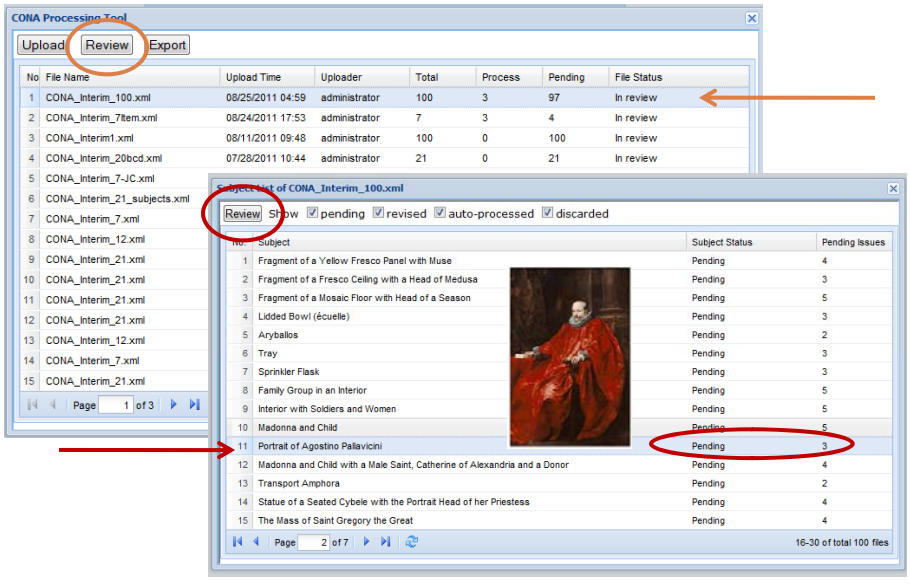
CONA from CVA to Publication

Upload Raw Interim XML file



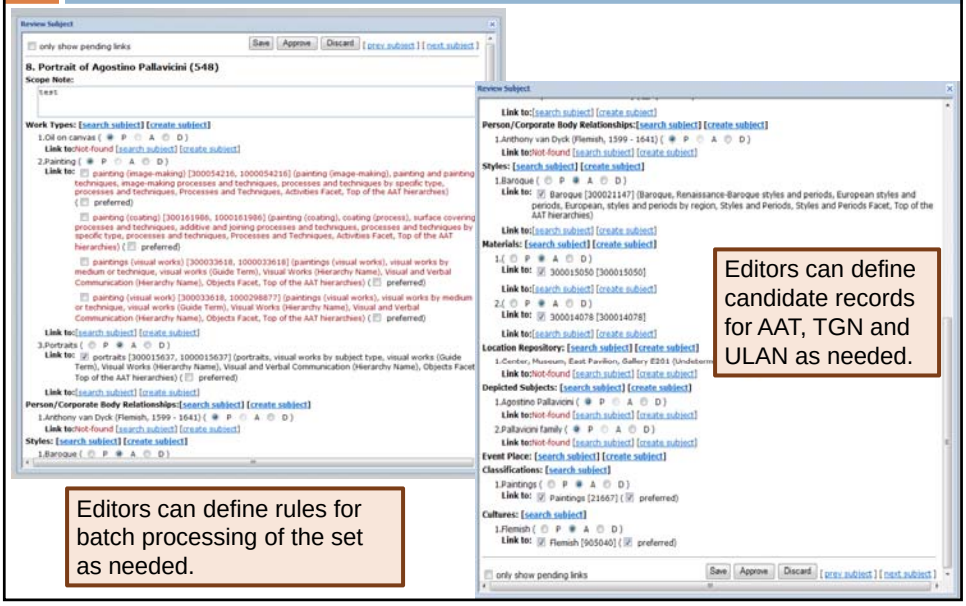
CONA from CVA to Publication

Manually review the list



CONA from CVA to Publication

Review the record

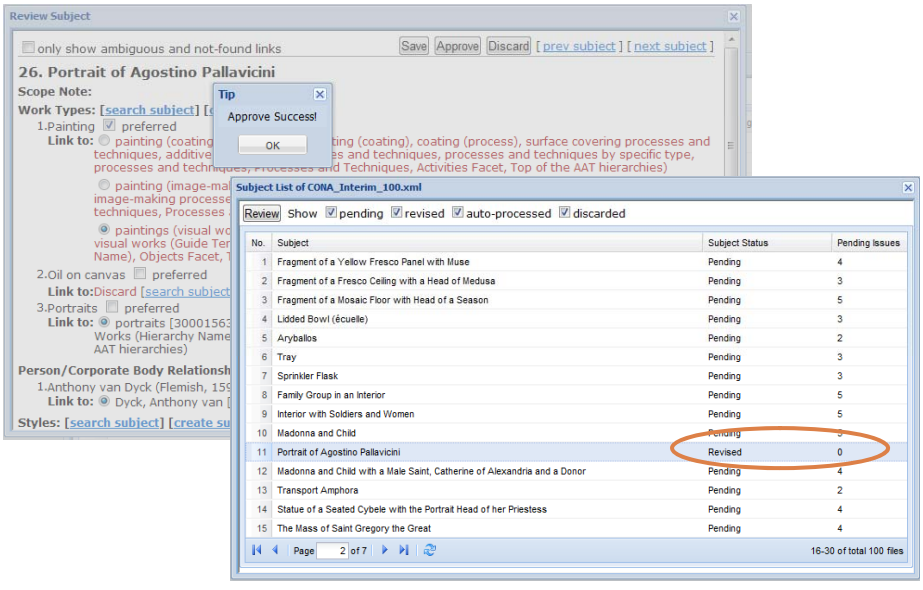


Editors can define candidate records for AAT, TGN and ULAN as needed.

Editors can define rules for batch processing of the set as needed.

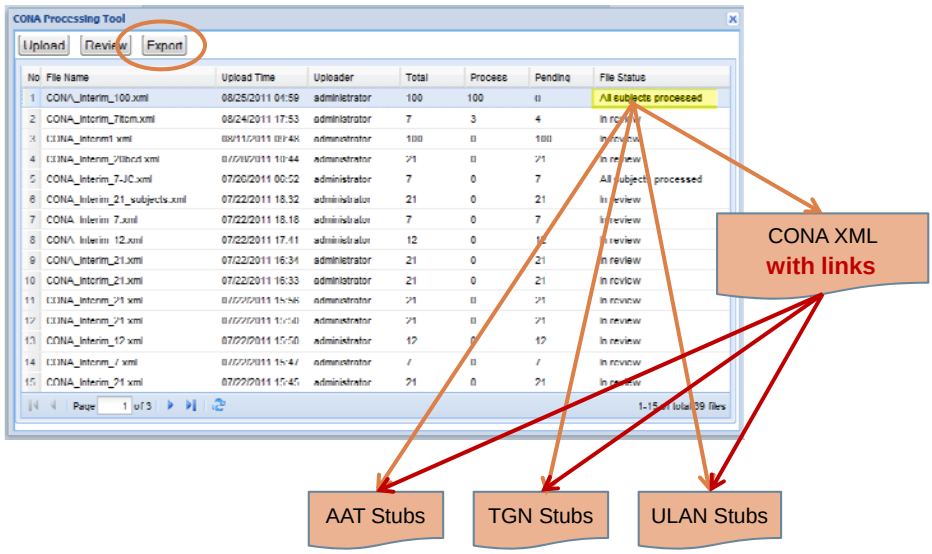
CONA from CVA to Publication

Approve the record



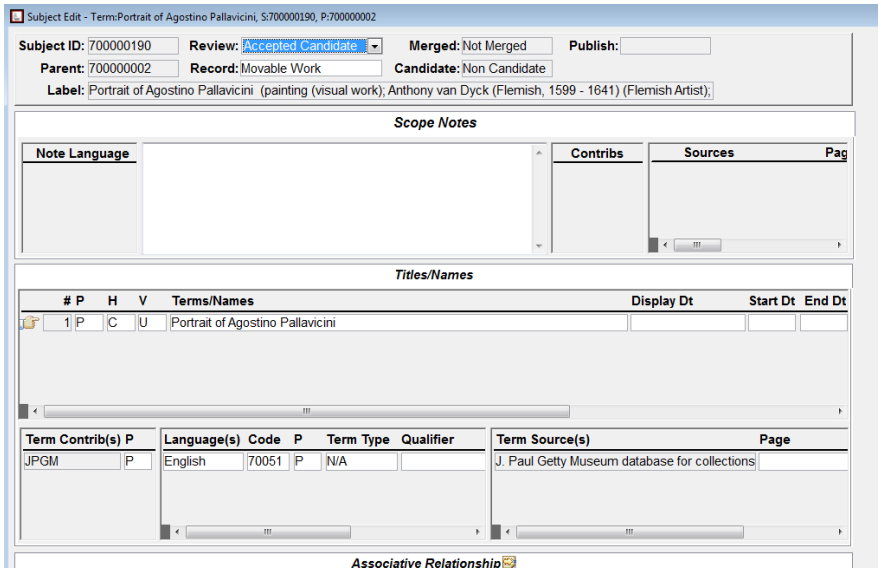
CONA from CVA to Publication

Ready to load into our editorial system



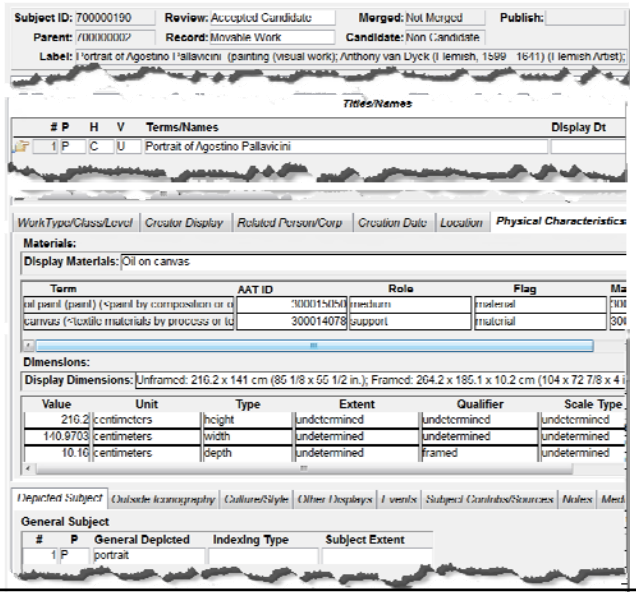
CONA from CVA to Publication

Edit the record in the editorial system if necessary



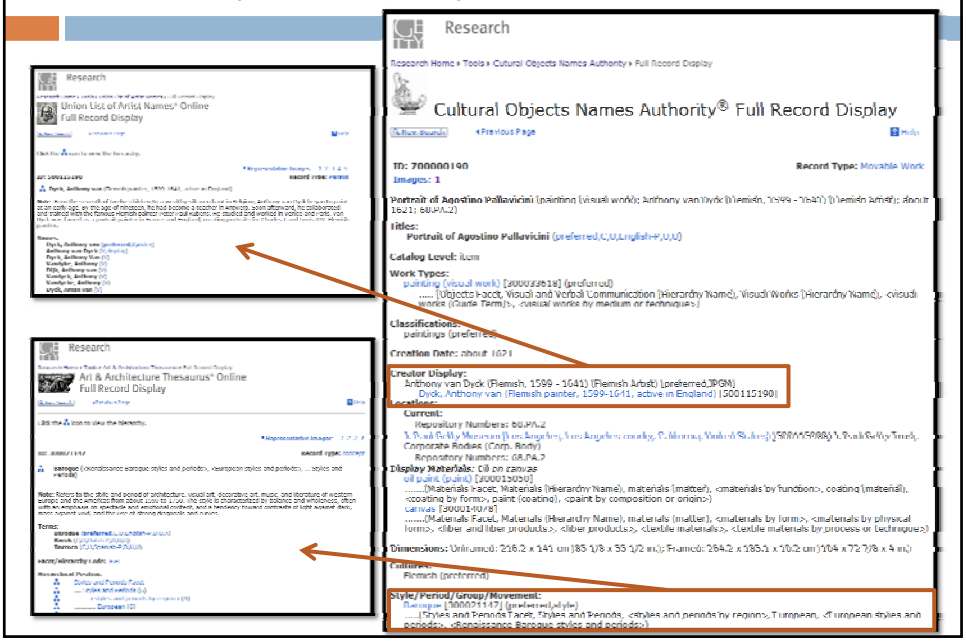
CONA from CVA to Publication

The links established in CVA are already there



CONA from CVA to Publication

Records are published bi-weekly to the web



Wednesday

Overview of some of the basic concepts relating to Linked Open Data.

A brief glossary –
to start us out on the same page

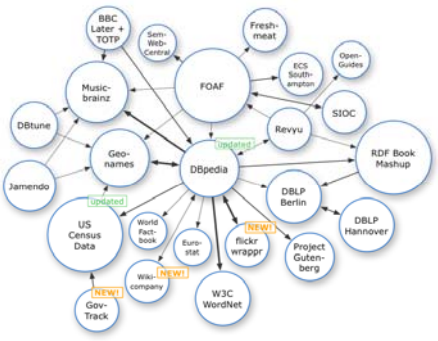
- Semantic Web
- Linked Open Data (LOD)
- Resource Description Framework (RDF)
- LOD Ontologies & Vocabularies
- Uniform Resource Identifier (URI)
- Web Ontology Language (OWL)
- Simple Knowledge Organization System (SKOS)
- Triplestore
- SPARQL Protocol RDF Query Language (SPARQL)
- Creative Commons (CC)

Semantic Web

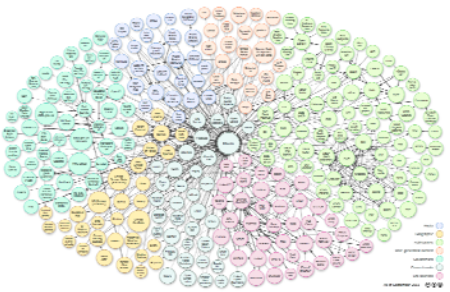
- A collaborative movement led by the World Wide Web Consortium (W3C),
- Promotes common data formats on the Web.
- Involves publishing in languages specifically designed for describing data: RDF, OWL, XML.
- Aims at converting the current web dominated by unstructured and semi-structured documents into a web of linked data,
- History
  - Concept of the Semantic Network Model was formed in the early 60s as a form to represent semantically structured knowledge.
  - Term 'Semantic Web' was coined by Tim Berners-Lee, the director of the W3C.
  - The original 2001 *Scientific American* article by Berners-Lee described an expected evolution of the existing Web to a Semantic Web.
  - He defined the Semantic Web as "web of data that can be processed directly and indirectly by machines." .

LOD: Doubled in size every 10 months since 2007

View of LOD as of October 2007
2 billion RDF triples
Interlinked by over 2 million RDF links



View of LOD as of September 2011
31 billion RDF triples
Interlinked by over 504 million RDF links



Linked Open Data (LOD)

- Linked Data
  - A method of publishing structured data so that it can be interlinked and become more useful.
  - It builds upon standard Web technologies such as HTTP and URIs.
  - But, rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers.
- Open data
  - The idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.
- Linked Open Data (LOD)
  - A community project, sponsored by the Semantic Web Education and Outreach group of the W3C, to extend the Web with a data commons by publishing various open datasets as RDF on the WEB and by setting RDF links between data items from different data sources.
  - Both linked and open.

Resource Description Framework (RDF)

- One of the key ingredients of Linked Data.
- Provides a generic graph-based data model for describing things, including their relationships with other things.
- Three column format – known as a *triple – forms the fundamental building block of semantic representations.*
  - *Subject - corresponds to the entity – a 'thing'*
  - *Predicates – property of the entity - names, birth/death dates*
  - *Objects - Subjects in another triple OR Literal values*
- Conceptualized anything (and everything) in the universe as a resource.
- Specifically supports evolution of schemas over time without requiring all the data consumers to be changed
- RDF has varying representations (e.g., XML, text, JSON)

LOD Ontologies & Vocabularies

- Ontology
  - Used to capture knowledge about some domain of interest. An ontology describes the concepts in the domain and also the relationships that hold between those concepts.
- Vocabulary
  - A vocabulary in the sense of knowledge systems or ontologies are controlled vocabularies. They provide a way to organize knowledge for subsequent retrieval.
- Semantic Web Ontology
  - This is a formal specification of a shared conceptualization of the information to be published.
  - Examples of the vocabulary used in ontologies: class, object property, role, category, property, resource, etc.

Uniform Resource Identifier (URI)

- Identifies a resource either by location (URL), or name (URN). It usually includes some type of unique and persistent identifier.
- URIs are not URLs but every URL is a URI
- The URI for a resource represented in an RDF statement is called the URI reference (URIref) for that graph node.
- URIs are used as 'strong keys' that are assigned to each of the nodes so that we can refer to them consistently across all the triples that describe their relationships.
- The URN that will be the base URI for AAT will be http://vocab.getty.edu/aat.

Web Ontology Language (OWL)

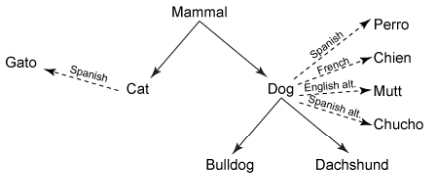
- Developed by the W3C
- Built as an extension to RDFS
- Provides a precise vocabulary that enables more powerful reasoning and inference over relationships with which to knowledge can be represented.
- Can be seen as a social contract between a data provider and a data consumer.

Simple Knowledge Organization System (SKOS)

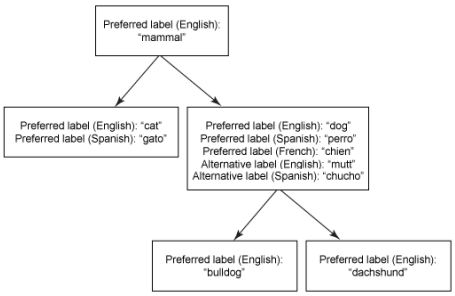
- A language designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary.
- Based on RDF
- SKOS is currently developed within the W3C framework
- Can be used on its own, or in combination with other ontologies.

LOD: Simple SKOS example

Sample concepts relationship in a pre-SKOS taxonomy



Sample concepts relationship in SKOS



http://www.ibm.com/developerworks/xml/library/x-skostaxonomy/index.html

Triplestores

- Are purpose-built databases for the storage and retrieval of the type of unstructured information express in RDF triples that makes up the Semantic Web.
- Are not relational databases.
- Query language used with triplestores is typically SPARQL
- In addition to queries, triples can usually be imported/exported using RDF and other formats.
- Some triplestores store billions of triples.

SPARLQL Protocol RDF Query Language

- Was made a standard by the RDF Data Access Working Group (DAWG) of the W3C.
- Is considered one of the key technologies of the semantic web.
- Allows users to write unambiguous queries.
- SPARQL endpoint is a service that accepts SPARQL queries and returns results.
- Queries can be distributed to multiple SQARQL endpoints.

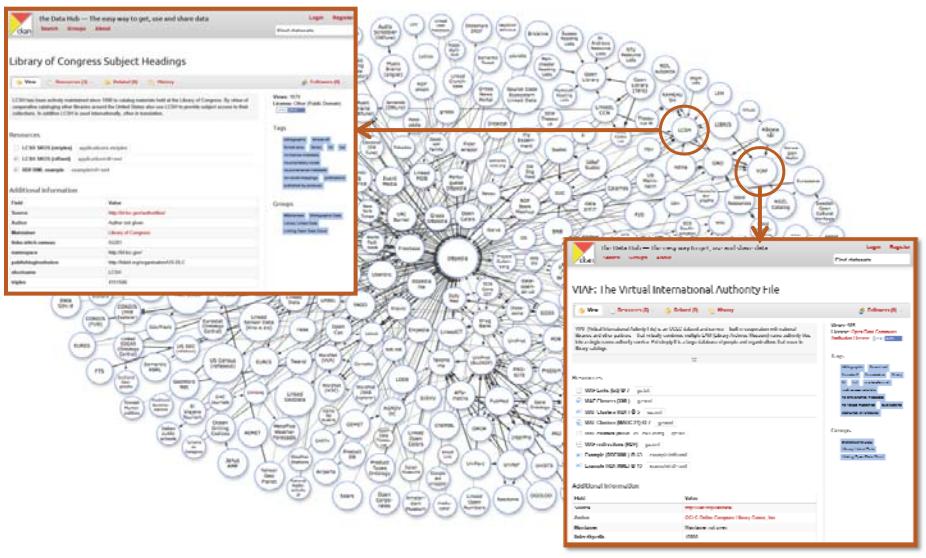
Why do we care

- The way data is being published on the web is currently in transition.
- A current trend in managing art information is to increasingly make data art, architecture, and cultural heritage objects available as Linked Open Data.
- This applies to the information directly describing the objects, but also the vocabularies used in the descriptions.
- For years now, the Getty Vocabulary team has been receiving legitimate requests to publish our thesauri as LOD.
- It is time to stop asking “why” and to start exploring “how”.

Explore the LOD Cloud Diagram

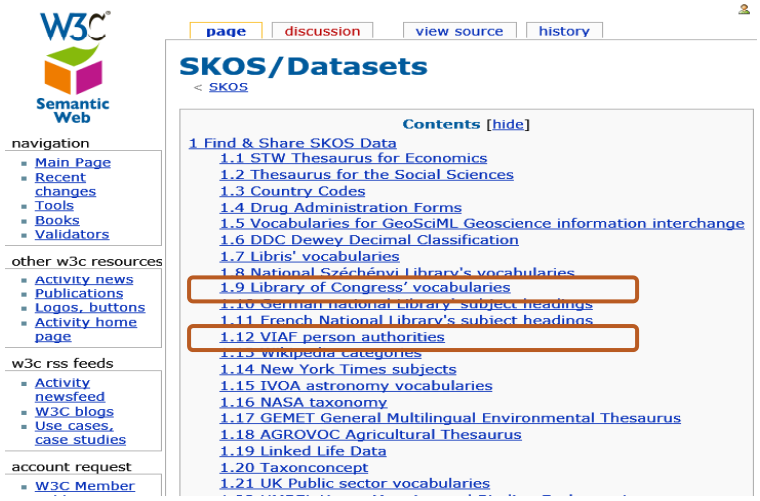
- This web page is the home of the diagram.
- The image show datasets that have been published in Linked Data format, by contributors to the Linking Open Data community project and other individuals and organizations.
- Clicking on the image will take you to an image map where each dataset is a hyperlink to its homepage
- URL: http://richard.cyganiak.de/2007/10/lod

Interactive LOD Cloud Diagram

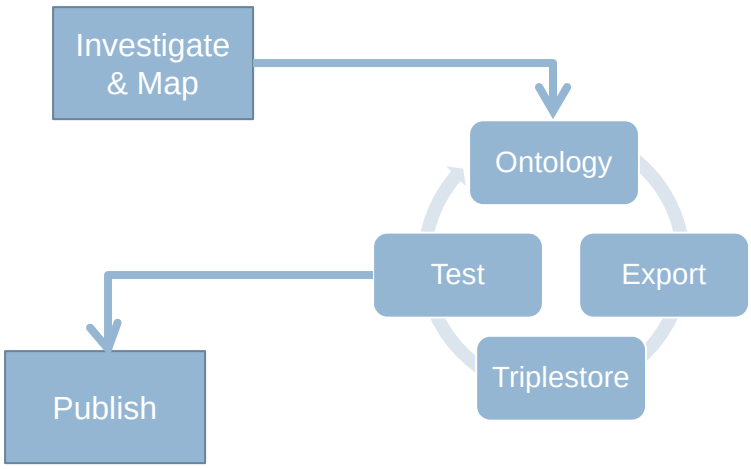


Explore Existing SKOS Datasets

URL: http://www.w3.org/2001/sw/wiki/SKOS/Datasets



Our process



Miscellaneous hints from various sources about semantic data models

- No one gets it right the first time. Be happy if it seems like you are getting closer by the third time around.
- Semantic data modeling is very different from relational data modeling.
- It is a collection of ontologies that work together to represent the slice of the world you are focusing on.
- They are extremely open-ended and highly collaborative.
- They borrow vocabularies whenever possible from other ontologies. The more you borrow, the more portable and extensible your data becomes.

Contact Information

Joan Cobb: jcobb@getty.edu
Gregg Garcia: ggarcia@getty.edu

Information Technology Services
J. Paul Getty Trust