THE GETTY VOCABULARIES
TECHNICAL SOLUTIONS OVERVIEW

Joan Cobb
IT Specialists Project Manager
Getty Information Technology Services

International Working Group Meetings
Friday, September 5, 2014
Current Technical Solutions

Contributions to AAT, TGN, ULAN
- Batch
- Online Forms

Contribution Validation Application (CVA)

Contributions from CONA
- Batch

Editorial System (VCS)

Linked Open Data

Bi-weekly exports to the public web sites

Web Service APIs

Yearly UTF-8 release
- XML & REL

Licensed Data Admin Application
LOD: General Information Site

Getty Vocabularies as Linked Open Data

The Getty vocabularies are constructed to allow their use in linked data. A project to publish AAT, TGN, ULAN, and CONA to the LOD (Linked Open Data) cloud is underway. The documents on this page contain news and presentations about releasing the Getty vocabularies as LOD. These materials are subject to frequent modification and addition.

- News and Status of the Project
- What is LOD?
- Introduction to Getty Vocabularies as LOD (PDF, 3.9 MB, 46pp)
- Linked Open Data Flair (PDF, 1 MB, 1pp)
- List of External Advisors (PDF, 8KB, 79p)

The AAT and TGN are now available as LOD. They are published under the ODC-by 1.0 license.

Developers and programmers, technical documentation to help you explore the AAT and TGN data is available at the SPARQL endpoint at vocab.getty.edu

News and Status of the Project

Ontology update: Note that the ontology has changed. Click on the “Developers and programmers...” link above for details.

Additional releases: Plans are in place to release the other vocabularies as LOD on the following dates: ULAN in January 2015 and the CONA demo in July 2015.

Other formats: In addition to LOD, we plan to continue providing the data in relational tables and XML releases. If any format is discontinued in the future, users will be given advance notice.

Working with the community: We have established an open community and we welcome collaboration. Previously we were getting feedback at an email address, but now the preferred way of communicating technical questions, comments, and usage stories is to use the SemanticWeb.com Q&A forum. We will be monitoring this forum regularly. The tags to use are: Getty, AAT, TGN, ULAN, CONA. The following links will pre-populate with the appropriate tags:

http://answers.semanticweb.com/questions/aat?tag=Getty_AAT
http://answers.semanticweb.com/questions/aat?tag=TGN
http://answers.semanticweb.com/questions/aat?tag=ULAN
http://answers.semanticweb.com/questions/aat?tag=CONA

For questions and comments about editorial content or general information regarding the Getty Vocabularies, please contact us at vocab@getty.edu

URLs: A description of the full set of URLs for all four vocabularies is available in "Getty Vocabulary Semantic Representation, section 1.5.3 GVP URLs and Prefixes."

The semantic links to the AAT and TGN subjects are now live. For example:

http://vocab.getty.edu/aat/300112492
http://vocab.getty.edu/tgn/7001393

Other vocabularies: Until the other vocabularies are available in LOD format, as requested by several in the community, we have implemented a redirect to the HTML version of the full record. Here are examples in each vocabulary:

http://vocab.getty.edu/aat/300125533
http://vocab.getty.edu/tgn/700061691

In order for data to be understood and processed automatically by computers, data in records or about resources must be expressed in a standard format. Each thing (for example, a museum object, a place, or a person) must be represented by a persistent identifier (known as a Uniform Resource Identifier (URI)). A Resource Description Framework (RDF) is a language or format for describing things as well as the relationships between things in terms of simple properties and values (known as ‘triples’), while things are represented using URIs. Among the most often-used formats for publishing art vocabularies are the Simple Knowledge Organization System (SKOS) and Web Ontology Language (OWL).

If data is to be open to the community for linking and discovery, traditional licensing and copyright practices for art information, images, and associated vocabularies must be adjusted. Data is considered open if the community is free to use, reuse, and redistribute the data, subject either to no restriction or to only the requirements of attribution or share-alike. Among the licenses most often applied to art information are Creative Commons and Open Data Commons licenses, each of which offers a range of levels of openness.

The Art & Architecture Thesaurus® (AAT), the Getty Thesaurus of Geographic Names® (TGN), and the Cultural Objects Name Authority® (CONA) are copyrighted by the J. Paul Getty Trust.

Revised 19 August 2014


© 2014 J. Paul Getty Trust, author: Joan Cobb. For educational purposes only. Do not distribute.
The Getty Vocabularies

Welcome to the SPARQL endpoint vocab.getty.edu

The Getty Vocabularies: The AAT, TGN, ULAN, and CONA contain structured terminology for art and other material culture, archival materials, visual surrogates, and bibliographic materials. Compliant with international standards, they provide authoritative information for catalogers and researchers, and can be used to enhance access to databases and Web sites. The Getty Vocabularies are produced by the Getty Vocabulary Program (GVP) and grow through contributions.

NOW Available:

- The Art & Architecture Thesaurus
  (Catherine wheel or rose wheel and architecture.)

- The Getty Thesaurus of Geographic Names
  London or Londinium? TGN

Ontology Update Note: There are changes to:

Documentation and Download:

- GVP Semantic Representation: Sections GVP URLs and Prefixes and Comprehensive Sample Queries are available
- GVP ontology: HTML, RDF/XML, and NTRiples
- Datasets: (N)TTriples Zip, see doc for
  - AAT: full zip (all statements), etc.
  - TGN: full zip (all statements), etc.
- VOID: void.ttl (see doc section Descriptions)
- AAT Relationship Types: Crosswalks
- General information about the LOD

Explore the Dataset:

- Text search: Use 'All/AAT/TGN' to search only terms. Use 'Full' to search
- SPARQL Query: Click on the 'SPARQL Query' to provide a lot more detail!

Questions and Comments:

For technical questions and comments, please post them on the SemanticWeb.com Q&A forum. We will be monitoring that forum regularly. The tags to use are: Getty, AAT, TGN, ULAN, CONA. The following links will pre-populate with the appropriate tags:

- http://answers.semanticweb.com/questions/ask/?tags=Getty,AAT
- http://answers.semanticweb.com/questions/ask/?tags=Getty,TGN
- http://answers.semanticweb.com/questions/ask/?tags=Getty,ULAN
- http://answers.semanticweb.com/questions/ask/?tags=Getty,CONA

For questions and comments about editorial content or general information regarding the Getty Vocabularies, please contact us at vocab@getty.edu.

Usage Stories:

It is our hope that we can all learn from each other, so please don't be shy about letting the community know how the LOD versions of the Getty Vocabularies are useful in your work. The best way to do this is to post to the SemanticWeb.com Q&A forum as described in the 'Questions and Comments' section above.

Openness:

The data is released under Open Data Commons Attribution License. Adherence to ODC Attribution instructions for the correct assertion of attribution is encouraged. The preferred form of attribution for AAT is:

"This [title or report or article or dataset] contains information from Art & Architecture Thesaurus (AAT), respectively Thesaurus of Geographic Names (TGN) which is made available under the ODC Attribution License."

Special Case: In circumstances where providing the full attribution statement above is not technically feasible, the use of canonical AAT/TGN URLs is adequate to satisfy Section 4.3 of the ODC Attribution License.

Future Publications

- The Union List of Artist Names (ULAN) ©
  - Tiziano Vecellio? ULAN is a structured vocabulary, including names, biographies, and other information about artists and architects.

- The Cultural Objects Name Authority (CONA) ©
  - Mona Lisa or La Gioconda? CONA, a new vocabulary now accepting contributions, includes titles, attributions, and other information for art and architecture.

Additional Links:

- The Getty Vocabularies
- The Getty Research Institute
- The Getty (Main)
Web Services - User’s Instructions

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

**Instructions include:**
- Examples
- Data Element Descriptions

**Web services:** Data files may also be accessed through Web services APIs (application programming interfaces), which provide access to data that is refreshed every two weeks. The APIs are available to any institution that holds a valid license. Please read details in the Web Services User’s Instructions (pdf).

Changes to the data structure: Changes have been made in the structure of all three vocabularies to allow for the seamless integration of multilingual terms. One significant change is that the Qualifier field is now linked to language (which has multiple occurrences per Term). In order to accommodate qualifiers in multilingual data, originally...
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
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)

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)

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)

http://www.getty.edu/research/tools/vocabularies/obtain/download.html
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
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
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
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
The Getty Vocabularies as Conduits to Other Resources
CONA Can Help Connect Resources

CONA Mapping Tool → CONA Interim XML → CONA Validation Application

Contributing Institution

SPARQL Queries To Synchronize Links

VOCAB.getty.edu

A contribution to CONA is a contribution to all 4 vocabularies

AAT XML → TGN XML → ULAN XML → CONA XML

Contributions Ready to Load into VCS

VCS Loading Process

Editorial System (VCS)

Bi-weekly LOD Publication
CONA-VCS: Live links to AAT, TGN & ULAN
A CONA contribution may also be a contribution to AAT, TGN & ULAN
Loading CONA data

Contribution Validation Application (CVA)

Raw Interim XML (un-linked)

Upload

Auto-processing

Has ambiguous or not found links?

NO

Auto-processed links

YES

Manually revised links

Pending

Manual review

Is record fixable?

YES

NO

Rejected Subject

Loader XML (with links)

Loader

VCS
CONA from CVA to Publication
Upload Raw Interim XML file

<table>
<thead>
<tr>
<th>File Name</th>
<th>Date</th>
<th>Time</th>
<th>Username</th>
<th>ID</th>
<th>File Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONA_Interim_100.xml</td>
<td>07/22/2011</td>
<td>07/22/2011</td>
<td>administrator</td>
<td>21</td>
<td>In review</td>
</tr>
<tr>
<td>CONA_Interim_7.xml</td>
<td>07/22/2011</td>
<td>07/22/2011</td>
<td>administrator</td>
<td>12</td>
<td>In review</td>
</tr>
<tr>
<td>CONA_Interim_12.xml</td>
<td>07/22/2011</td>
<td>07/22/2011</td>
<td>administrator</td>
<td>7</td>
<td>In review</td>
</tr>
<tr>
<td>CONA_Interim_21.xml</td>
<td>07/22/2011</td>
<td>07/22/2011</td>
<td>administrator</td>
<td>21</td>
<td>In review</td>
</tr>
</tbody>
</table>

Upload file progress: 100%
CONA from CVA to Publication
Manually review the list
Editors can define rules for batch processing of the set as needed.

Editors can define candidate records for AAT, TGN and ULAN as needed.
CONA from CVA to Publication
Approve the record

26. Portrait of Agostino Pallavicini
Scope Note:
Work Types: [search subject] [edit]
1. Painting [ ] preferred
   Link to: ○ painting (coating), coating (process), surface covering processes and techniques, additive processes and techniques, processes and techniques by specific type, processes and techniques, Processes and Techniques, Activities Facet, Top of the AAT hierarchies)
2. Oil on canvas [ ] preferred
   Link to: Discard [search subject]
3. Portraits [ ] preferred
   Link to: @ portraits [30001563] Works (Hierarchy Name: AAT hierarchies)

Person/Corporate Body Relationship:
1. Anthony van Dyck (Flemish, 1599-1641)
   Link to: Dyck, Anthony van [edit]

Styles: [search subject] [create subject]

Subject List of CONA_Interim_100.xml

<table>
<thead>
<tr>
<th>No.</th>
<th>Subject</th>
<th>Subject Status</th>
<th>Pending Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fragment of a Yellow Fresco Panel with Muse</td>
<td>Pending</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Fragment of a Fresco Ceiling with a Head of Medusa</td>
<td>Pending</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Fragment of a Mosaic Floor with Head of a Season</td>
<td>Pending</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Lidded Bowl (scuolai)</td>
<td>Pending</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Aryballos</td>
<td>Pending</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Tray</td>
<td>Pending</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Sprinkler Flask</td>
<td>Pending</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Family Group in an Interior</td>
<td>Pending</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>Interior with Soldiers and Women</td>
<td>Pending</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Madonna and Child</td>
<td>Pending</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Portrait of Agostino Pallavicini</td>
<td>Revised</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Madonna and Child with a Male Saint, Catherine of</td>
<td>Pending</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>Transport Amphora</td>
<td>Pending</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Statue of a Seated Cybele with the Portrait Head of</td>
<td>Pending</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>The Mass of Saint Gregory the Great</td>
<td>Pending</td>
<td>4</td>
</tr>
</tbody>
</table>
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
Current Technical Solutions

Contributions to AAT, TGN, ULAN

Contribution Validation Application (CVA)

Contributions from CONA

Linked Open Data

Editorial System (VCS)

Bi-weekly exports to the public web sites

Web Service APIs

Licensed Data Admin Application

Yearly UTF-8 release

XML & REL
Possible Mid Term Technical Solutions

Batch Contributions to

AAT, TGN, ULAN and CONA

Editorial System (VCS)

Vocabs on the Web

AAT, TGN, ULAN, CONA

vocab.getty.edu

SPARQL Queries

Bi-weekly publication to LOD

Semantic Downloads

JSON, RDF, N3/Turtle, N-Triples
full.zip and explicit.zip

Licensed Data Admin Application

Yearly UTF-8 release

XML & REL

Web Service APIs
Possible LongTerm Technical Solutions

Contributions to AAT, TGN, ULAN and CONA

Vocabs on the Web
AAT TGN ULAN CONA

vocab.getty.edu

Linked OPEN Data and Protected Linked Data

Editorial System (VCS)

Semantic Downloads
JSON, RDF, N3/Turtle, N-Triples full.zip and explicit.zip
Joan Cobb
IT Specialist Project Manager
Information Technology Services
1200 Getty Center Drive
Los Angeles, CA 90049
jcobb@getty.edu

The J. Paul Getty Trust