Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies

Murtha Baca,* and Melissa Gill**

* Head of the Digital Art History Program, Getty Research Institute 1200 Getty Center Drive, Los Angeles, CA 90049-1679 <MBaca@getty.edu>
** Digital Projects Manager, Getty Research Institute 1200 Getty Center Drive, Los Angeles, CA 90049-1679 < MGill@getty.edu >

Murtha Baca, Head of the Digital Art History Program at the Getty Research Institute (GRI) in Los Angeles, holds a PhD in Art History and Italian Language and Literature from UCLA. She is also an adjunct professor in the Graduate School of Education and Information Studies at UCLA, teaching graduate seminars on metadata and indexing and thesaurus construction. Her research and publications have focused largely on descriptive metadata and controlled vocabularies for art, architecture, and material culture. Dr. Baca is currently editing a new edition of *Introduction to Metadata*, to be published by the Getty Research Institute in 2015.

Melissa Gill holds an MLIS from the University of Washington, where she is also finishing her Master in Art History. Her research interests include digital humanities and cultural heritage description and access. She is currently the Digital Projects Manager at the Getty Research Institute (GRI) in Los Angeles.

Baca, Murtha, and Gill, Melissa. **Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies.** *Knowledge Organization. 42(4)*, 232-243. 18 references.

Abstract: This paper gives an overview of the history, development, and structure of the electronic thesauri produced and maintained by the Getty Research Institute (GRI). We describe the evolution of the *Art & Architecture Thesaurus* (*AAT*®), the *Getty Thesaurus of Geographic Names* (*TGN*®), and the *Union List of Artist Names* (*ULAN*®) as multilingual, cross-cultural knowledge organization systems (KOS); the factors that make them unique; and their potential, when expressed as Linked Open Data (LOD) to play a key role in the Semantic Web.

Received: 4 May 2015; Accepted 5 May 2015

Keywords: Linked Open Data, Getty Vocabularies, multilingual, sematic web, metadata

1.0 Introduction

In our current digitally entrenched epoch, we have experienced an overwhelming growth of information resources that are purportedly available to anyone with Internet access, but are actually daunting if not impossible to find and navigate in the vast universe of the World Wide Web. Libraries, archives, and museums, in an effort to make digitized and born-digital content globally accessible, have embraced the emergent trend of unified online discovery systems to facilitate cross-repository searching of heterogeneous collections (Long and Schonfeld 2014, 46). Federated resources such as ARTstor, the Google Cultural Institute, the Digital Public Library of America, and Europeana[1] aggregate descriptive metadata and digital surrogates representing diverse repositories, languages, and cultures. In large, varied data resources of this kind, keyword searching is woefully inadequate for comprehensive information retrieval (Bates 1998, 1185). If they are to be truly available and effective, these systems must reconcile the challenges inherent in international cultural heritage domain knowledge, which by its very nature is multilingual, multicultural, interdisciplinary, and manifested in diverse formats. (Hyvönen 2010, 5). We believe that controlled vo-

cabularies, as knowledge organization systems (KOS) that achieve consistency and provide a wide range of access points for resource description and enhanced precision and recall in information retrieval, will play a key role in an increasingly multicultural and multilingual information ecosystem. The controlled vocabularies produced by the Getty Research Institute (GRI) are multilingual, semantically structured thesauri that can be powerful tools for enriching knowledge and providing meaningful links for cultural heritage information resources. This paper will provide an overview of the development and structure of the Getty vocabularies, their recent release as Linked Open Data (LOD), and their potential role in the Semantic Web. For the sake of clarity, we have included a brief glossary at the end of this paper that defines key concepts.

2.0 The Getty Vocabularies: Overview and Core Data Structure

In response to the cultural heritage documentation community's need for controlled vocabularies specifically relating to art, architecture, and other material culture (for which authority files such as the *Library of Congress Subject Headings* and *Thesaurus for Graphic Materials* were useful, but not fully adequate), in the 1980s the J. Paul Getty Trust began a program to develop thesauri for the cultural heritage domain. From the beginning, the Getty sought partnerships with users and other stakeholders, including art and architectural historians, architects, librarians, visual resource curators, archivists, museum specialists, and specialists in thesaurus construction, with the goal of creating resources applicable to the diverse interests and requirements that would allow cross-collection retrieval. The first thesaurus developed under the Getty's aegis was the *Art & Architecture Thesaurus*® (*AAT*), which includes terms, descriptions, and other information for generic concepts related to art, architecture, conservation, archaeology, and other cultural heritage. Over time, the work of the Getty Vocabulary Program was broadened to include a structured vocabulary containing names and biographical information for artists, patrons, and other agents in the cultural realm, resulting in the *Union List of Artist Names*® (*ULAN*). The third Getty vocabulary was a thesaurus containing names and other information for inhabited places, geographic features, and archeological sites: the *Getty Thesaurus of Geographic Names*® (*TGN*). The Getty vocabularies are faceted thesauri in compliance with ISO and NISO standards for thesaurus construction. The first two editions of the *AAT* (1990, 1994) and the first edition of the *ULAN* (1994) were released in printed volumes. As information technology evolved and the Internet burgeoned, the *AAT*, *ULAN*, and *TGN* were published through an online search interface, and also made available as full datasets for licensing by libraries and other cultural institutions, as well as by commercial entities such as collection management software vendors. Currently, the licensed, full datasets in relational tables and XML formats are released annually, with data updated every two weeks in APIs (application programming interfaces) and in the online search interface. Also, the Getty vocabularies are now available as Linked Open Data releases, which are also updated every two weeks.

The Getty vocabularies all share a core data structure: they map to a common schema, and are interconnected technically as well as semantically. Unlike the Library of Congress authorities, the *AAT*, *ULAN* and *TGN* are true thesauri, not lists of subject headings. Unique concepts in each vocabulary are represented by records, which contain terms or names, notes, dates, bibliography, and other information about the concept. One preferred term or descriptor is used as a default term to represent the concept in online displays. Terms may comprise a single word (e.g., "Baroque"); other terms may be "bound terms" (e.g., "rose window"), which are multiple-word terms expressing a single, unique concept. The terms and records are explicitly and semantically linked through the equivalence, hierarchical, and associative relationships that are inherent in the structure of a thesaurus. For example, in the *AAT*, the record for the object known as a "rhyton," a distinctive drinking vessel often shaped like an animal horn used in Ancient Greece and the Middle East, includes equivalence relationships for additional terms such as the plural "rhyta," and variant terms "rheons," "rhytons," and "ritons." The concept record has a hierarchical relationship under the broader parent term "drinking vessels." It is distinguished by an associative relationship to the related but different object "stirrup cup" (Figure 1).

The vocabularies' polyhierarchical structure also allows for concepts to be linked to multiple parents, thus one concept may appear in multiple hierarchical views (Harpring 2013, 42) (Figure 2).

The temporal nature of cultural information, including changes in nomenclature and interpretations of meaning over time, is represented in the vocabularies through the aggregation of current as well as historical terminology. For example, in the *Getty Thesaurus of Geographic Names*®, the record for the Indian city of Kolkata designates the transliterated Bengali name "Kolkata," approved in 1999, as the preferred name along with the historical names (flagged as such within the database) "Calcutta," "Fort William," and "Kalikātā" (Figure 3).

Vocabularies that aggregate variant terminology referring to a single concept can significantly enhance both the precision and recall of online searches by leading users to relevant resources that they would not have otherwise found without the use of the additional access

ID: 300198841 **Record Type:** concept

rhyta (drinking vessels, <vessels for serving and consuming food>, ... Furnishings and Equipment (Hierarchy Name))

Note: Refers to vessels from Ancient Greece, eastern Europe, or the Middle East that typically have a closed form with two openings, one at the top for filling and one at the base so that liquid could stream out. They are often in the shape of a horn or an animal's head, and were typically used as a drinking cup or for pouring wine into another vessel. Drinking was done by holding the rhyton above the drinker's head and catching the stream of liquid in the mouth.

Terms:

rhyta (**preferred**,C,U,LC,English-P,D,U,PN)
(Spanish,UF,U,PN)
(Greek (transliterated)-P,D,U,PN)
rhyton (C,U,English,AD,U,SN)
(Spanish,AD,U,SN)
(Greek (transliterated),AD,U,SN)
Rhyton (C,U,English,AD,U,SN)
rhytons (C,U,English,UF,U,N)
(Spanish-P,D,U,PN)
(French-P,D,U,PN)
rhea (vessels) (C,U,English,UF,U,N)
rheons (C,U,English,UF,U,N)
rheon (C,U,English,UF,U,N)
ῥυτόν (C,U,Ancient Greek,UF,U,U)
rhŭtón (C,U,Ancient Greek (transliterated),UF,U,U)
菜坦酒杯 (C,U,Chinese (traditional)-P,D,U,U)
角狀杯 (C,U,Chinese (traditional),UF,U,U)
角杯 (C,U,Chinese (traditional),UF,U,U)
lái tǎn jiǔ bēi (C,U,Chinese (transliterated Hanyu Pinyin)-P,UF,U,U)
lái tǎn jiǔ bēi (C,U,Chinese (transliterated Hanyu Pinyin),UF,U,U)
lai tan jiu bei (C,U,Chinese (transliterated Pinyin without tones)-P,UF,U,U)
lai t'an chiu pei (C,U,Chinese (transliterated Wade-Giles)-P,UF,U,U)
rytons (C,U,Dutch-P,D,U,U)
ryton (C,U,Dutch,AD,U,U)
ritons (C,U,French,UF,U,N)
riton (C,U,French,UF,U,N)
Rhyta (C,U,German-P,D,U,PN)
ritóns (C,U,Spanish,UF,U,N)
ritón (C,U,Spanish,UF,U,SN)
escanciadora (C,U,Spanish,UF,U,SN)

Facet/Hierarchy Code: V.TQ

Hierarchical Position:

Objects Facet
.... Furnishings and Equipment (Hierarchy Name) (G)
..... Containers (Hierarchy Name) (G)
..... containers (receptacles) (G)
..... <containers by function or context> (G)
..... culinary containers (G)
..... <containers for serving and consuming food> (G)
..... <vessels for serving and consuming food> (G)
..... drinking vessels (G)
..... rhyta (G)

Related concepts:

distinguished from .... **drinking horns**
..... (drinking vessels, <vessels for serving and consuming food>, ... Furnishings and Equipment (Hierarchy Name)) [300043229]
distinguished from .... **stirrup cups**
..... (drinking vessels, <vessels for serving and consuming food>, ... Furnishings and Equipment (Hierarchy Name)) [300197140]
distinguished from .... **sturzbechers**
..... (beakers (drinking vessels), drinking vessels, ... Furnishings and Equipment (Hierarchy Name)) [300197148]

Figure 1. The *Art & Architecture Thesaurus*® web display for "rhyta" illustrates (1) equivalence, (2) hierarchical, and (3) associative relationships.

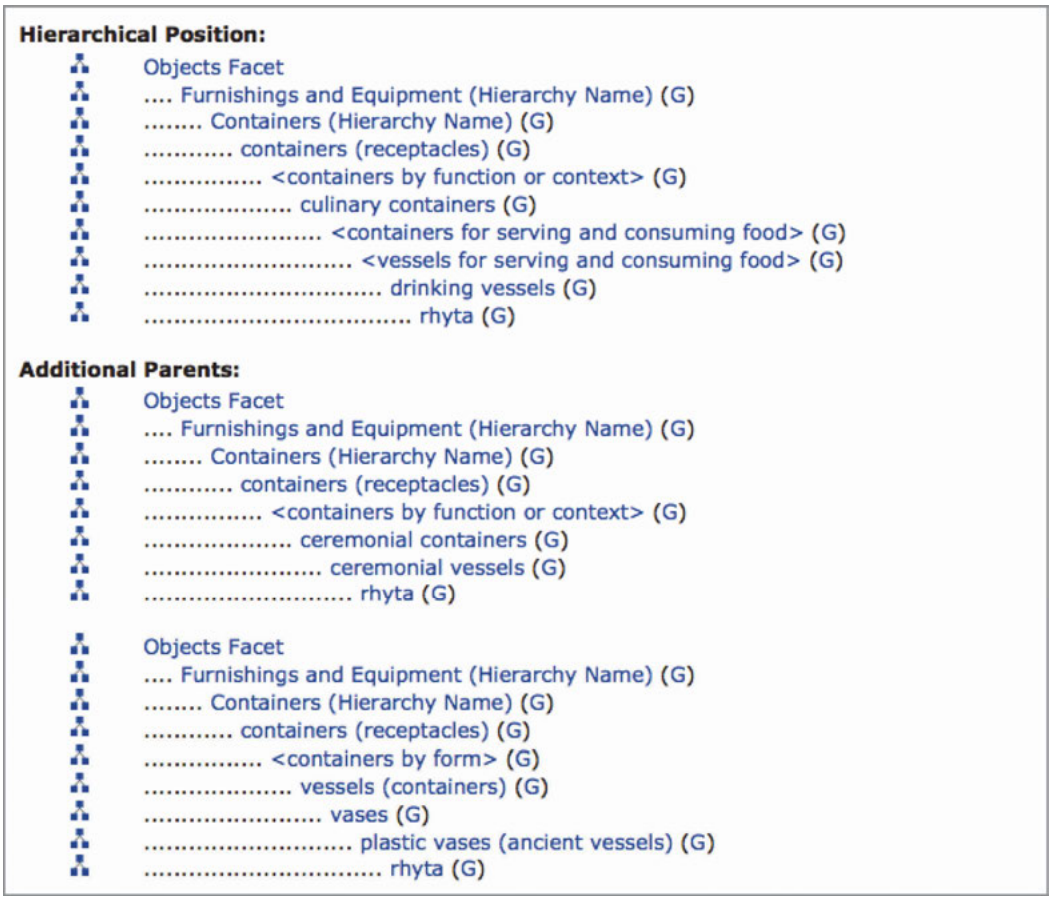

Figure 2. The *Art & Architecture Thesaurus*® web display for "rhyta" illustrates polyhierarchical relationships.

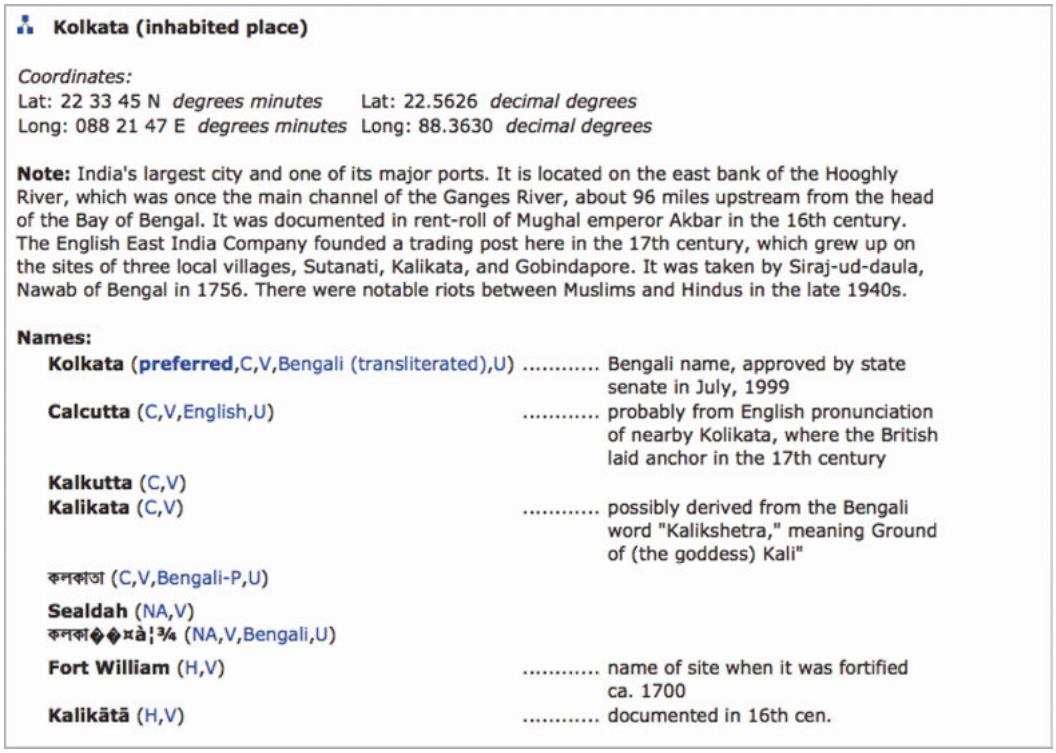

Figure 3. The *Union list of Artist Names*® web display for "Calcutta" illustrates the clustering of historical and contemporary terminology associated with the same concept (in this case, a place).

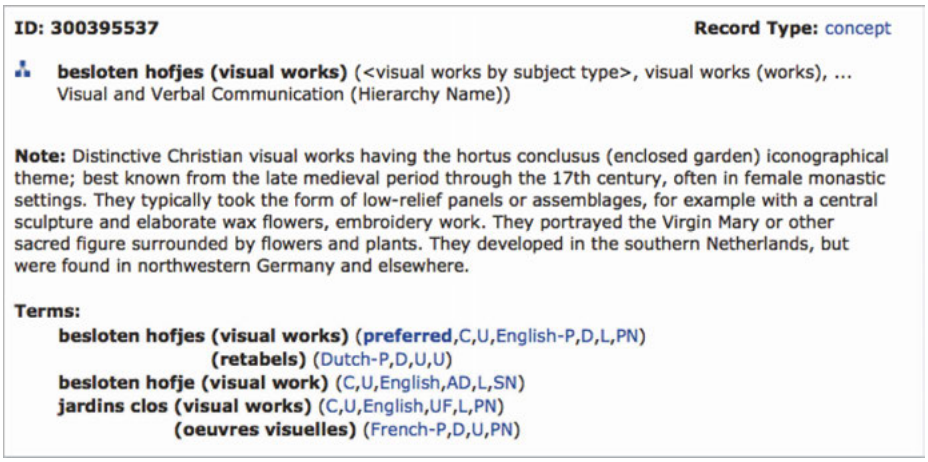

Figure 4. The *Art & Architecture Thesaurus*® web display for "besloten hofjes" illustrates the linguistic variants added by the partner institutions.

points provided by the vocabulary. The Getty vocabularies cluster together historical and other linguistic variants relating to a single concept, which when applied to a variety of collection metadata can create additional access points for users in search and retrieval.

3.0 Issues and Challenges in Creating Multilingual Thesauri

The Getty vocabularies grow through contributions from institutions and organizations from the international knowledge organization community, including repositories of art and cultural heritage as well as projects concerned with indexing and cataloging art and architecture. Major contributing organizations that have undertaken complete translations of the *AAT* include the Academia Sinica in Taipei; the State Museums of Berlin; the Centro de Documentación de Bienes Patrimoniales (a subdivision of the Dirección de Bibliotecas, Archivos y Museos, known as DIBAM) of Chile; and the Netherlands Institute for Art History (RKD). Partnerships with other cultural organizations greatly enhance the development of multilingual terminology, through large-scale translation projects as well as smaller batch contributions. The Getty vocabularies contain multilingual equivalents for generic concepts, such as liturgical "reliquaries" ("reliquiari" (Italian), "箱物聖" (Chinese-traditional), "shèng wù xiāng" (Chinese-transliterated Hanyu Pinyin), "reliquiaria" (Dutch)) as well as proper names, such as for the Renaissance Italian artist "Leonardo da Vinci" ("Léonard de Vinci" (French) and "チンィヴ・ダ・ドルナオレ" (Japanese)). The language of the term is often labeled with a language flag. Multilingual controlled vocabularies enrich descriptive metadata, provide additional important access points, and enhance online search and retrieval for collections of metadata encoded in different languages (Harpring 2013, 178).

The process of translating the Getty's *Art & Architecture Thesaurus*® into a target language is labor-intensive, requiring teams of experts in language, content, and thesaurus construction. When the members of translation project are also attempting to map their legacy local vocabularies to the existing *AAT*, occasionally they discover that a required concept is missing in the *AAT* and a new concept must be submitted for inclusion. For example, the Netherlands Institute for Art History, which has done a complete Dutch translation of the *AAT*, contributed a new record for "besloten hofjes," a type of low relief or assemblage developed in the southern Netherlands during the Middle Ages (Figure 4).

The addition of new concepts to the *AAT* may result in the creation of entirely new hierarchical branches that must be integrated into existing hierarchies or facets. For example, the Academia Sinica, which is responsible for the Chinese *AAT* translation project, recently introduced eight new concepts for Chinese festivals; the inclusion of these new concepts necessitated the creation of a hierarchical level for "cultural holidays," which was placed under the existing "holidays" in the *AAT*'s "Activities" facet. These types of translation projects promote multicultural documentation and access with the integration of new concepts and terminology into an existing, English-language-centric controlled vocabulary.

Even in cultures that share a common language, the same term may represent different concepts. For example, the *AAT* distinguishes between the term "retablo" used in Spain to denote a large altar screen or appendage ("reredos" or "retable" in English) versus the kind of small devotional panel painting called a "retablo" in Latin America (Baca 2014, 121-124). This is an example of why the

unique numeric IDs used by the Getty vocabularies are so important—they uniquely identify the concept, even if the terms representing the concept are homographs. Other resources, such as the Virtual International Authority File (VIAF) [2], also use unique numeric IDs. In an increasingly linked retrieval environment, if cultural heritage resources have included the unique numeric IDs rather than simply the text of an *AAT* or other indexing term, search portals will be able to better utilize true semantic retrieval.

4.0 What Makes the Getty Vocabularies Unique?

As mentioned above, the *AAT*, *ULAN*, and *TGN* are true thesauri, with all the power of the equivalence, hierarchical, and associative relationships. Unlike subject headings, or traditional dictionaries, the Getty vocabularies encode one unique concept per record, thus disambiguating homographs and avoiding false matches when users search on related or even identical terms. For example, the term "landscape" can represent two distinct concepts, a built environment and a visual representation of the environment; these concepts are represented as separate, distinct records with terms that are homographs in the *AAT*. In this case, the records are linked as "related concepts," and disambiguated by the associative relationship type "distinguished from," in addition to the use of qualifiers "environments" and "representations." In general, however, homographs do not have an associative relationship simply because the terms are spelled the same; there must be a direct relationship between the concepts. For example, the records for "drum" may refer to a musical instrument, while a second record refers to "drum" as a component of a column, while a third record refers to "drum" as the base for a dome. These records are not linked by associative relationships in the *AAT*, although they would be listed under a single entry in a dictionary.

Qualifiers, scope notes, and the placement of an *AAT* concept in the context of the hierarchy help disambiguate terms for users. For machines, it is the unique, persistent ID of every concept that identifies the concept; even if the terms change or the hierarchical position is altered, the ID remains.

The equivalence, associative, and hierarchical relationships encoded in the *AAT*, *ULAN*, and *TGN* make each vocabulary semantically and technically linked within itself; for example, in the *ULAN* the record for "Michelangelo Buonarroti" contains associative relationship links to the records for people to whom this artist was associated (Figure 5).

The vocabularies are also linked to each other; for example, the *ULAN* record for "Michelangelo Buonarroti" also contains links to the *TGN* records for Rome and Florence, places where the artist was active (Figure 6).

The Getty vocabularies also contain a wealth of bibliographic information. Each name or term in a vocabulary record is linked to one or more contributors, as well as to bibliographic sources that serve as literary warrant for usage of the term, illustrating the academic authoritativeness and research value of the vocabularies. The data structure and basic principles under which the vocabularies are constructed and maintained emphasize multilinguality and multiculturality. They are compiled through contributions over time, constantly growing with the addition of new terms and concepts. Their growth is inherently "social," in that they are built up primarily from contributions from trusted partner institutions (but not "crowd sourced" via contributions from the general public, which would significantly reduce their authoritativeness). Last but not least, the Getty vocabularies are freely available as reference and cataloging tools on the GRI's website, and now in the form of Linked Open Data as part of the Getty's Open Content program, which is aimed at making not only high-resolution images and associated descriptive metadata, but also large research datasets, available without restrictions.

5.0 Linked Open Data and the Getty Vocabularies: Linking and Enriching Cultural Heritage Information

The Getty vocabularies are multifunctional; they function as knowledge bases, data value standards for cataloging and resource description, and tools for enhancing online search and retrieval. Now they can be exploited in new ways for retrieval and discovery with the release of of the *AAT*, *TGN*, and *ULAN* as Linked Open Data (LOD). As discussed by Zeng and Chan (2004, 370), LOD represents a shift toward networked knowledge organization systems (NKOS) in the age of the Internet. The Semantic Web, as an extension of the World Wide Web, aims to add a semantic layer of machine-readable, standardized data into the Web's existing architecture (Berners-Lee et al. 2001). Making the vocabularies available as openly accessible linked data is in keeping with the Getty's Open Content policy. The time was also right due to the growing number of museum and library datasets being published as Linked Open Data. The *AAT* and *TGN* were released as LOD in 2014 and the *ULAN* as of April 2015. All three vocabularies are published under the Open Data Commons Attribution License (ODC-By) v1.0.

As Linked Open Data, the Getty vocabularies are expressed as structured and openly reusable machine-readable data that information systems can interpret and use to create semantically relevant relationships across other linked datasets. The data are described using the principles of the Resource Description Framework (RDF),

Figure 5. The *Union list of Artist Names*® web display for "Michelangelo Buonarroti" illustrates associative links to the various individuals to whom this artist was associated, also represented by unique records in the same thesaurus.

a standard model for data interchange on the web that employs uniform resource identifiers (URIs) to identify the name and location of resources (any piece of data in the vocabularies) and expresses entities and the relationships between them as triples, or subject-predicate-object statements (Harpring 2013, 233), thus creating a semantic network of information. The Getty Vocabulary Program's LOD datasets are machine accessible at http://vocab.getty.edu/, with sample data and full documentation on how implementers can access and utilize the data, and they are available from the program's website at http://www.getty.edu/research/tools/vocabularies/. The Getty's LOD datasets are downloadable in several RDF-based formats: Turtle, JSON, RDF/XML, and N-Triples.

Wherever possible, the Getty technical staff expressed the data as LOD utilizing a set of standard ontologies—Simple Knowledge Organization System (SKOS) and SKOS-XL for representing thesauri information, Dublin Core (DC) for common properties, W3C Geo Ontology (WGS) for geographic information, Friend of a Friend

Figure 6. The *Union list of Artist Names*® web display for "Michelangelo Buonarroti" illustrates links to the *Thesaurus of Geographic Names*® concepts for his place of birth, death, and related events.

(FOAF) and Bibliographic Ontology (BIBO) for sources and contributors, Provenance (PROV) for revision history, and RDF, RDFS, OWL and XSD for system properties. For data that could not be mapped to an existing standard ontology, the Getty technical team developed a specific ontology, called GVP. The ontology stack taken as a whole creates a complete semantic representation of Getty vocabulary data, which is especially rich, deep, and multifaceted. Full documentation on the vocabularies' semantic representation is available at http://vocab.getty.edu/doc/.

Expressing the complex intricacies of the *AAT*'s multilingual translations as Linked Open Data posed a challenging undertaking for the Getty technical team and external collaborators. Language translations of concepts in the *AAT* contain loan terms, or words borrowed from other languages that become naturalized in the borrowing language (National Information Standards Organization and American National Standards Institute 2010, 32). For example, the French concept, "trompe-l'eoil," (literally, "deceive the eye"), used to describe a two-dimensional image rendered to appear as occupying three-dimensional space, is a loan word found in English, in addition to the Dutch and Spanish translations that have been contributed to the *AAT* by partner institutions. The four languages are flagged for the same term "trompe-l'eoil." This is problematic for the Semantic Web, which expects each term to be unique with an individual identifier. The Getty technical team had to resolve how to encode the language flags so that information systems could interpret them as discrete entities, while also making it clear that the loan words are identical, in their spelling and meaning, across several languages. This was resolved by representing each instance of the term as an individual URI composed of a core numeric identifier and the IANA (Internet Assigned Numbers Authority) language code. With the example "trompe-l'eoil," the term is expressed as four separate URIs, each containing the same identifier (1000056506) followed by the language code: *AAT*_term:1000056506-fr, *AAT*_term:1000056506-en, *AAT*_term:1000056506-es, *AAT*_term:1000056506-nl. This allows information systems to read language codes, which can be used for metadata enrichment and to enhance multilingual information services.

The Getty LOD project capitalizes on the vocabularies' existing semantic structure. Although the *AAT*, *TGN*, and *ULAN* have always been linked together conceptually, technical mapping required comprehensive harmonization across the three vocabularies. For example, Place Type list values in the *TGN* and Nationality/Culture/Race/Ethnicity and Roles lists values in the *ULAN* were mapped to existing concepts in the *AAT*, or if those concepts were nonexistent, they were added as new concepts. The "Languages" controlled list values found in all three vocabularies were also added to the existing "languages and writing systems" hierarchy in the *AAT*. ISO 639 alpha-2 and alpha-3 codes were also added as variant descriptors; for example, the *AAT* record for the Romance language "Spanish" includes the ISO codes "es" and "spa" as variant terms. The conceptual links established across the Getty vocabularies are now tangibly linked within the vocabularies' data and as LOD.

The linked data cloud has experienced exponential growth since Tim Berners-Lee first wrote about the Semantic Web in 2001. The evolution of the Linked Open Data Cloud diagram exemplifies the escalating presence of published linked datasets on the Web. In 2007, only twelve linked open datasets existed; by 2011 this number had grown to 295, and as of April 2014, 570 datasets are linked and published in the cloud, including the *AAT* (Figure 7a & 7b).

Many libraries, archives, and museums are now proceeding to structure and publish their collection information as machine-readable, linked datasets for better integration into the Web and re-use by other organizations. Knowledge organization systems, described as "value vocabularies" in the Semantic Web domain, play a crucial role in the success of Linked Data by acting as "hubs," using HTTP URIs to connect concepts, names, and works across datasets from different communities and domains (Bermes 2011, 7; Hooland and Verborgh 2011, 224). Enhancing cultural heritage metadata with the URIs

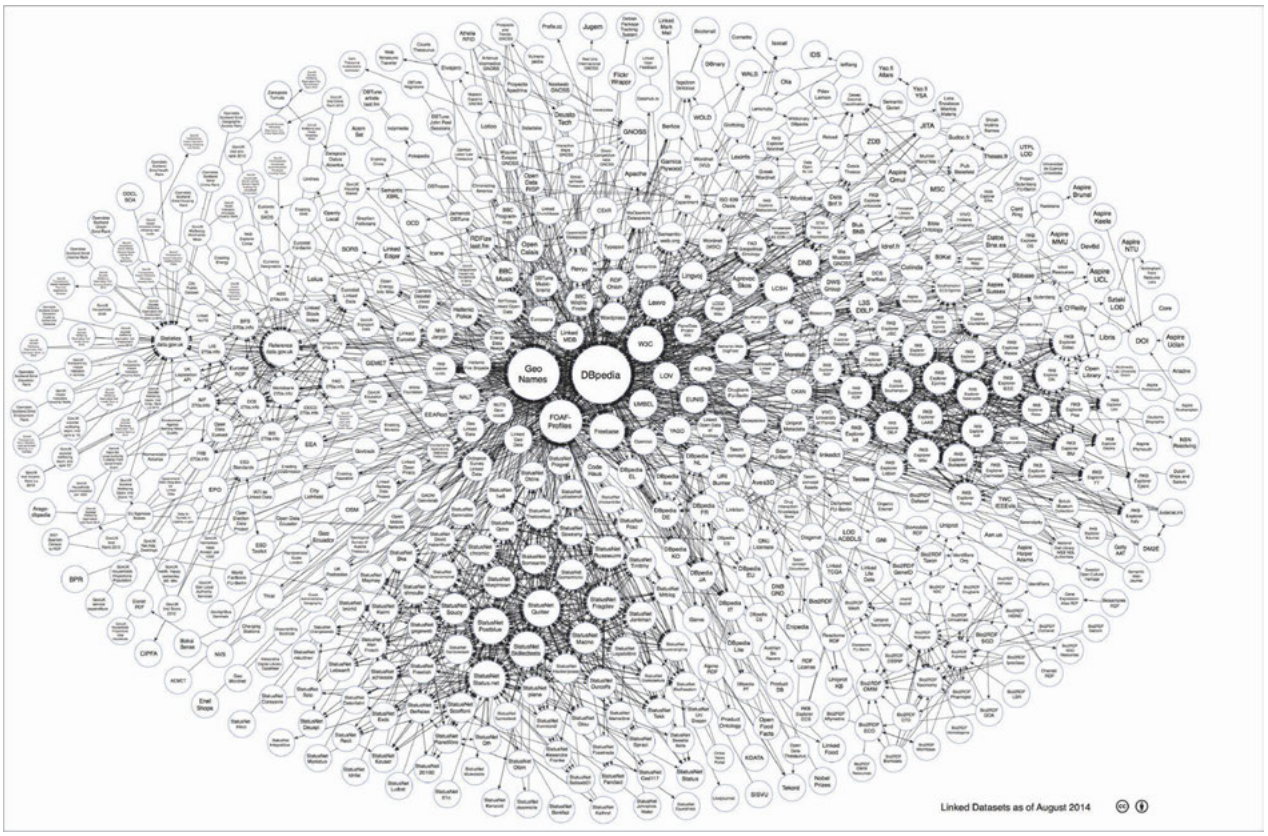

Figure 7a. Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

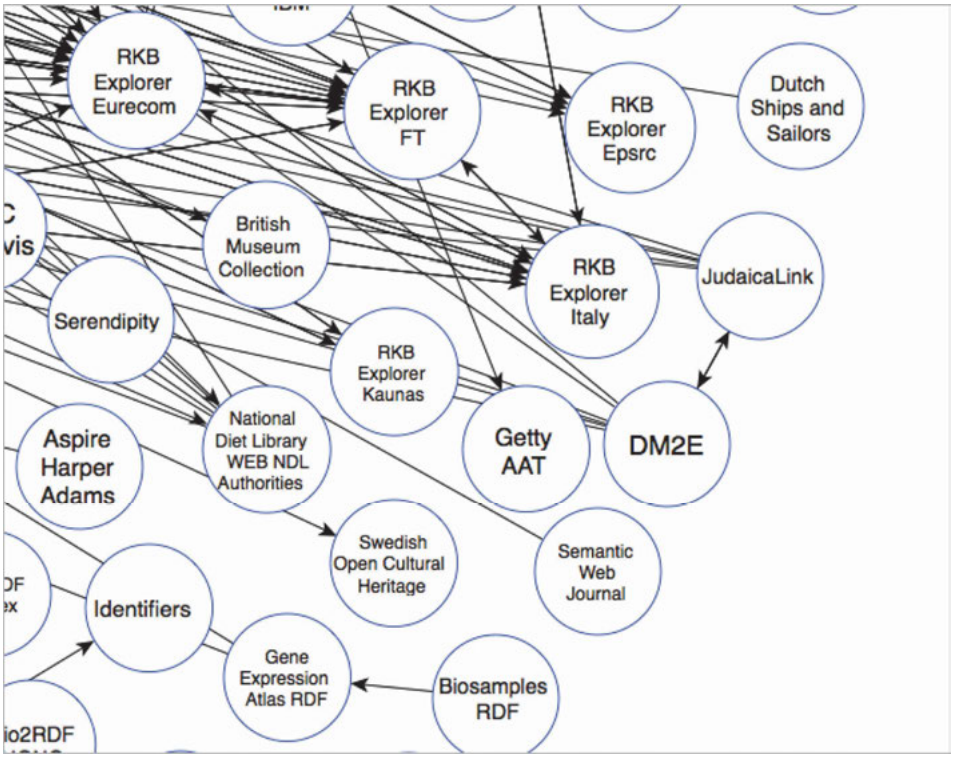

Figure 7b. Linking Open Data cloud diagram [detail], 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

from relevant controlled vocabularies and authorities expressed as Linked Open Data is essential if the Semantic Web is to fulfill its promise of creating an interconnected, discoverable information ecosystem.

The growing presence of online search systems for cultural heritage information, and more recently cross-domain and cross-cultural information portals, highlights the importance of multilingual access in a networked, global environment. Multilingual access in information systems involves three aspects: multilingual interfaces, multilingual search and browsing, and multilingual result display and translation. (Stiller et al. 2012, 4). Multilingual value vocabularies hold the potential to enrich existing metadata and support the search and browsing of resources described in different languages. Cultural institutions and repositories working with LOD stress that linked, open, interoperable, and multilingual vocabularies are of paramount importance for augmenting semantic and multilingual searching (Charles et al. 2014, 3). The Rijksmuseum, for example, first began working with Linked Open Data to support multilingual access to its collections online, intending to integrate and exploit the multilingual labels found in semantic vocabularies (Dijkshoorn et al. 2014, 5). At the time of this writing, there are few examples available that demonstrate comprehensive semantic and multilingual search as supported by linked data vocabularies. Europeana's recent leveraging of the *AAT* as LOD in their portal search illustrates how NKOS can enrich metadata records and augment search and retrieval across multilingual datasets.

Europeana, an online information portal, provides access to millions of cultural heritage resources by aggregating metadata from museums, libraries, and archives across Europe. Metadata from contributing institutions is linguistically diverse and hence poses several multilingual challenges, such as how to integrate search across multiple languages and ensure that users can interpret and understand metadata records retrieved in unfamiliar languages (Charles et al. 2014, 3). Europeana relies on knowledge organization systems to resolve multilingual issues by using an internal metadata enrichment tool to leverage value vocabularies available in the linked data cloud. For metadata records with encoded *AAT* URIs (from the Rijksmuseum, Museo Galileo, Erfgoedplus.be, and institutions forming the Partage Plus project) the enrichment tool makes use of variant terminology, language labels, and semantic data from the Getty's Linked Data service (Charles and Devarenne 2014). The auto-generated labels are processed by the Europeana Data Model (EDM), which seamlessly integrates this information into the system's semantic layer for enhanced search and retrieval. The Europeana metadata record for Johannes Vermeer's oil painting *The Milkmaid (Het melkmeisje)* at the Rijksmuseum[3] includes "auto-generated" tags from the *AAT*, which were populated from URIs present in the Type and Format fields. The object type "easel painting" displays the multilingual translations as human-readable labels: [Staffeleibild] (de); [peintures de chevalet] (fr); [easel paintings (paintings by form)] (en); [pinturas de caballete] (es); [schilderijen] (nl). Enriching collection metadata with machine-readable URIs from a multilingual thesaurus such as the *AAT* produces comprehensive search results across different languages, independent of the object record's native language. The Dutch object record is retrieved whether searching with the German descriptor "Staffeleibild" or the French descriptor "peintures de chevalet," therefore facilitating greater resource recall. The *AAT*'s encoded IANA language values are also utilized to benefit metadata display. For example, when selecting "Français" for the web display language setting, the system will read the encoded language labels and automatically generate the French terms "peinture de chevalet" under Type, and "peinture à l'huile" under Format. Europeana's usage of the Linked Open Data version of the *AAT* in their search illustrates how multilingual and interoperable NKOS can be incorporated into complex data repositories and utilized for enhanced search and retrieval. Although this application is limited to existing *AAT* URIs in contributor data, this example illustrates the potential for developing comprehensive multilingual knowledge management services.

7.0 Conclusion

The Semantic Web offers the promise of universal, unfettered access to a vast array of information in different formats and languages, and from different cultures. Metadata is an important component in the deployment and success of the Semantic Web (Greenberg, Sutton, and Campbell 2003, 7). At the time of this writing, the adoption of Linked Data technologies by libraries, archives, and museums is still in the preliminary stages of development. A critical mass of information is necessary to exploit the full potential of cross-domain semantic search, and libraries, archives, and museums have only relatively recently begun to transform and release their collection metadata as LOD. Furthermore, applications for interpreting and displaying linked data as human-readable information are still needed for users to fully benefit from semantic technologies. Several cultural heritage institutions and consortia, including the Mellon-funded Research Space Project, the American Art Collective, and the Getty, are moving forward to explore conceptualizing and building such tools. Now that the Getty vocabularies are available as LOD, major purveyors of cultural information such as OCLC and the Google Cul-

tural Institute have approached the Getty to inquire about how these rich multilingual thesauri can be incorporated with other datasets and technologies. It is clear that multilingual, semantically structured thesauri and authorities are needed now more than ever. We believe that controlled vocabularies such as the VIAF and the thesauri produced by the Getty Research Institute will play a pivotal role in the emerging universe of semantically linked information resources.

Notes

1. See ARTstor, http://www.artstor.org/; Google Cultural Institute, https://www.google.com/culturalinstitute/home; Digital Public Library of America, http://dp.la/; and Europeana http://www.europeana.eu/portal/.
2. See http://www.getty.edu/about/opencontent.html. See also Murtha Baca. 2014. "Open Content: A Concept Whose Time Has Come" *Visual Resources* 30, no. 1: 1-4.
3. See http://www.europeana.eu/portal/record/90402/SK_A_2344.html.

References

Baca, Murtha. 2014. "On Language." *Visual Resources: An International Journal of Documentation* 30, no. 2: 121-4.

Bates, Marcia J. 1998. "Indexing and Access for Digital Libraries and the Internet: Human, Database and Domain Factors". *Journal of the American Society for Information Science* 49, no. 13: 1185-205.

Bermes, Emmanuelle. 2011. "Convergence and Interoperability: a Linked Data perspective." In *Proceeding of the 77th IFLA World Library and Information Congress 13-18 August 2011 San Juan, Puerto Rico.*

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5: 28-37.

Bower, James M. and Murtha Baca. 1994. *Union List of Artist Names*. New York : G.K. Hall.

Charles, Valentine and Cecile Dévarenne. 2014. "Europeana Enriches Its Data with the Art and Architecture Thesaurus." *Europeana Professional*. http://pro.europeana.eu/europeana-aat.

Charles, Valentine, Nuno Freire and Antoine Isaac. 2014. "Links, Languages and Semantics: Linked Data Approaches in the European Library and Europeana." In *Proceeding of the 80th IFLA World Library and Information Congress 16-22 August 2014, Lyon, France.*

Dijkshoorn, Chris, Wesley ter Weele, Lizzy Jongma and Lora Aroyol. 2014. "The Rijksmuseum Collection as Linked Data." *Semantic Web* 0 IOS Press: 1-6.

Greenberg, Jane, Stuart Sutton, and D. Grant Campbell. 2003. "Metadata: A Fundamental Component of the Semantic Web." *Bulletin of the American Society for Information Science and Technology* 29, no. 4: 16-18.

Harpring, Patricia. 2013. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and other Cultural Works*. Los Angeles: Getty Research Institute.

Hooland, Seth van and Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish your Metadata.* Chicago: Neal-Shuman.

Hyvönen, Eero. 2010. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. San Rafael, California: Morgan & Claypool Publishers.

Long, Matthew and Roger C. Schonfeld. 2014. *Supporting the Changing Research Practices of Art Historians.* New York: Ithaka S+R. http://sr.ithaka.org/sites/default/files/reports/SR_Support-Changing-Research-ArtHist_20140429.pdf.

National Information Standards Organization. 2010. *ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Multilingual Controlled Vocabularies*. Baltimore: National Information Standards Organization.

Petersen, Toni. 1990. *Art & Architecture Thesaurus*. New York: Oxford University Press.

Petersen, Toni. 1994. *Art & Architecture Thesaurus. 2nd ed.* New York: Oxford University Press.

Stiller, Juliane, Maria Gade and Vivien Petras. 2012. *Midterm Report on Innovative Multilingual Information Access.* Europeana. http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/Readings/D7.7%20Midterm%20Report%20on%20Innovative%20Multilingual%20Information%20Access.pdf.

Zeng, Marcia Lei and Lois Mai Chan. 2004. "Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems." *Journal of the American Society for Information Science and Technology* 55, no. 5: 377-95.

Glossary

Concept: a discrete entity or idea; in an ISO- and NISO-compliant thesaurus, each record represents a single concept or "subject." In the Getty thesauri, a concept may represent an agent (an individual person or corporate body), a place, an object type, an abstract concept, and so on.

Linked data: semantically structured datasets that are machine-readable and processible.

Linked open data (LOD): linked data published on the web with an open license for use, re-use and re-distribution.

Ontology: a formal machine-readable specification in which entities, attributes, and their interrelationships are explicitly defined and represent a particular domain of knowledge or discourse. Identifying an existing ontology or ontologies, or developing an appropriate ontol-

ogy, is a necessary first step when transforming datasets into LOD.

Term: a word or group of words denoting a single concept in a controlled vocabulary.

Thesaurus: a monolingual or multilingual controlled vocabulary that is explicitly structured to encode the equivalence, hierarchical, and associative relationships between concepts.

Value vocabulary: defined within the Semantic Web domain as semantically structured and machine-readable data value standards (authority files, thesauri, subject headings, and controlled lists) that are used to populate metadata elements.