Preparing Legacy Data for Publication in a Linked and Open World

Joan Cobb
Getty Digital, the J. Paul Getty Trust

The Getty's Open Content Program is an ongoing commitment to make the knowledge resources produced by the Getty's various programs (Getty Museum, Getty Research Institute, Getty Conservation Institute) freely available to all. These resources include content from across the different Getty programs, comprising digital images, a rich and wide variety of textual content, descriptive and structural metadata, provenance information in the form of "big data," and identity information (often called "authorities") relating to people, places, and art-historical concepts. The first step in this direction was taken in August 2013 with the release of more than 4000 high-resolution images of the Getty Museum's collection. Currently, there are more than 100,000 images from the J. Paul Getty Museum and the Getty Research Institute (GRI) available through the Getty's Open Content Program, including more than 30,000 images from the Getty Museum's collection using the International Image Interoperability Framework (IIF), a state-of-the-art standard for publishing and sharing high-resolution digital images.

The first Getty resources released as Linked Open Data (LOD) were the Getty vocabularies. The decision to start LOD work with these large and complex electronic thesauri was based on the fact that they are broadly used resources in the library, museum, archive, and cultural heritage communities. The *Art & Architecture Thesaurus*® (AAT) was published as LOD in February 2014, followed by the *Getty Thesaurus of Geographic Names*® (TGN) in August 2014, and the *Union List of Artist Names*® (ULAN) in April 2015. Currently, the Getty is actively engaged in preparing additional resources to be released as Linked Open Data, and helping other organizations to do the same. The types of resources include museum, provenance, conservation, library, and archival datasets.

One of the things learned in preparing the Getty vocabularies for release as Linked Open Data was that preparing legacy data for publication as LOD is neither simple nor easy. In addition, the complexity is dramatically increased when the datasets are still active and growing, and include a broad range of orthographic, historical, and multilingual terms and names.

This presentation will provide an overview of some of the issues encountered and solutions found during the process of preparing a variety of resources at the Getty for release as Linked Open Data. Topics will range from the importance of choosing the right ontology to reconciling resources with the Getty vocabularies. I will also share strategies for dealing with issues such as as uncertain and ambiguous dates, historical forms of terms and names, currencies, revision history, and synchronization.

Keywords: LOD, legacy data, reconciliation, Getty vocabularies, multilingual terminology, open content