Best Practice Poster: Making Vendor-Generated Metadata Work for Archival Collections Using VRA and Python

Carolyn Hansen
University of Cincinnati
United States
carolyn.hansen@uc.edu

Sean Crowe
University of Cincinnati
United States
sean.crowe@uc.edu

Keywords: vendor-generated metadata; metadata mapping; archival description; Dublin Core; VRA; Python

1. Introduction

Although cataloging cultural resources requires a greater level of descriptive granularity than standard library materials, metadata for digital collections is often generated by non-specialists. This can lead to significant problems with metadata accuracy and consistency, causing breakdown of authority control, high incidence of false positives in searching, and impeded access to materials. The purpose of this poster is to illustrate a successful workflow for improving vendor-generated metadata for a large digital collection of archival materials by converting the metadata from the Dublin Core standard to the VRA standard using the scripting language Python.

2. Background

The University of Cincinnati Libraries (UCL) contracted with a vendor to scan and generate metadata for the Cincinnati Subway and Street Improvements Collection. Consisting of photographs and documents related to the construction of the unfinished Cincinnati Subway system and street improvements throughout the city, the collection is a unique resource documenting early 20th century transportation, urban planning, and social history. Following the initial load of approximately 9,000 scanned images and associated Dublin Core metadata records into the shared OhioLINK Digital Repository Center, librarians Sean Crowe and Carolyn Hansen were charged with converting the metadata to the VRA standard, improving metadata quality, and loading the collection into the University's Luna image repository. Carolyn Hansen brought metadata standard expertise and Sean Crowe provided technical and scripting skills to the project.

3. Implementation

The planning and specifications for the contract scanning project were conducted by UCL's Digital Projects Repositories Department, and did not include input from UCL's Content Services Division, in which the authors work. As a result, the project workflow began with an assessment phase, which involved researching the initial scanning project, assessing the vendor-generated metadata, and gathering domain-specific information about the original physical format of the materials. A metadata map was created to record decisions about field equivalents between Dublin Core and VRA, controlled vocabulary usage, improvement of vendor-generated metadata, and addition of VRA-specific fields to describe original materials and digital surrogates.

These decisions were then encoded into a Python script. The Python script incorporated a custom class to parse and process the metadata in CSV format. In addition to coding the field conversions and formatting field contents based on the metadata map, the script ran several validation processes on the input and output metadata files. Finally, a function was added to the

script to link records to image files by unique identifier. Coding the script comprised a considerable portion of the project timeline though the script run-time was negligible.

4. Challenges

Project implementation involved a number of challenges. In terms of metadata mapping, moving from a less robust standard like Dublin Core to a very robust standard like VRA required strategic decisions. Since VRA provides the opportunity for highly-detailed descriptive metadata, it is necessary to look at the metadata with a strong editorial eye in order to balance detailed description with project time constraints and vendor-created metadata of varying quality. In order to accomplish this, a baseline for acceptable metadata was created, detailing changes to vendor-created metadata as well as who would be responsible for metadata enrichment. For example, errors in access points from controlled vocabularies such as LCNAF or LCSH headings would be corrected by Content Services faculty, but additional subject analysis would be provided by curators at a later stage in the project. The metadata quality baseline was also applied to controlled vocabulary usage. For example, when working with detailed vocabularies like the Getty Research Institutes' Art & Architecture Thesaurus, it was important to balance the level of descriptive granularity with vocabulary that was understandable to users and applicable to a wide range of materials.

Additionally, local practices regarding archival materials presented unique challenges to the project. Specifically, university archivists at UCL preferred that the structure of the digital collection should replicate the physical archive, including record order and collection level titles for item records. As a result, titles without description of the image content such as "Rapid Transit Photographs -- Box 17, Folder 22 (September 21, 1922 - October 24, 1922) -- negative, 1922-09-28, 9:42 A.M." were used. These titles offer little descriptive content and create greater reliance on subject searching. Further work needs to be done to make the collection searchable based on the content of the image. Lastly, geographic coordinates, included in some of the records, enrich the collection and should be added where possible.

5. Conclusions/Results

Since the collection was posted in Fall 2013, it has received over 17,000 unique page-views in the Luna Repository. This project serves as a template for future shared, interdepartmental projects. Further collaboration is certain as traditional Library Technical Services operations evolve to support local and unique digital content, including research data, archival material, and beyond.

Acknowledgements

Linda Newman, Head, University of Cincinnati Libraries Digital Content & Repositories Dept. and Elna Saxton, Head, University of Cincinnati Libraries Content Services Division.

References

Crowe, Sean and Carolyn Hansen (2014). DC_to_VRA. In GitHub. Retrieved from https://github.com/crowesn/DC_to_VRA.

University of Cincinnati Libraries (2014). Cincinnati Subway and Street Improvements, 1916-1955. Retrieved from http://digital.libraries.uc.edu/subway/.

University of Cincinnati Libraries (n.d.). LUNA Digital Repository. Retrieved from http://digproj.libraries.uc.edu:8180/luna/servlet/univcincin~42~42>.