alj

44/4 2019

Old metadata in a new world: Standardizing the Getty Provenance Index for linked data

Kelly Davis

What challenges does metadata managers face when embarking on a Linked Data transformation? How can legacy projects be better understood to ensure such data remodels stay true to their original content? Current efforts made by the Getty Provenance Index elucidate how this task can be much more than simple data re-structuring, and how much manual and specialized labor goes into updating a 35 year old database.

The Getty Provenance Index ® (GPI)[1] databases, an essential resource in the history of collecting and provenance, has a long and celebrated history. Since the 1980s, the GPI has made available difficult-to-access archival materials by pairing pioneering database technology with art historical content. Inputting content was completed by in-house staff, compiled from materials sent from a network of international collaborators[2] who transcribed documents and shipped photocopies of those documents to Santa Monica, California from across Western Europe. The content in the GPI is sourced from unpublished or uncommon materials (for example, inventories of family collections, auction catalogues dating to the 17th century, and business records of art dealers). The network of editors allowed the Index to acquire data quickly from a variety of archives and repositories. Over the years, the editorial styles and standards applied to the GPI have varied, in no small part due to the wide-ranging structure of the project. The result is abundant metadata that is a product of its history. When the Getty Research Institute launched a technical and conceptual remodel of the Provenance Index in 2016,[3] it opened a Pandora's Box of over 35 years of collected data that generated a significant amount of 21st-century concerns.

One of the many benefits of the GPI's history is the stability offered by the database system – the same one in use since its inception. It is powerful and has adapted to the changing needs of the Index via expansion and flexibility. In the early 1990s, it became possible for contributors to connect with the Index's databases remotely, entering data directly instead of sending documents to the central collection point in California. The databases were originally designed to support publication of the Index in book form,[4] but were later adapted for publishing on CD-ROM. As technology changed, the databases were made available on the web and have been available in that format since 2001. However, while robust and flexible, the system has a number of challenges for emerging users interested in analyzing large sets of data or for traditional provenance researchers seeking to understand complex events.

Rather than one database, the GPI is actually comprised of a family of related databases. For sales catalogs and the archival inventories, there are pairs of databases, one with item-level metadata and one with document-level metadata, that work together. But, records across distinct types of content databases do not link. A Rubens that appears in a Parisian auction in 1733 may be the same Rubens recorded on an archival inventory in the GPI, but a user would not learn this from accessing those records in their current form. The art historical nature of the content and history of the project also present challenges in standardization and fielding data. Some fields, particularly note fields where information that perhaps belonged elsewhere was recorded, have unstructured metadata. Long text strings may contain information about the object itself or add details of the

1. Provenance Index Databases (Getty Research Institute), accessed May 21, 2019, http://www.getty.edu/research/tools/provenance/search.html.

2. Provenance Collaborators (Getty Research Institute), accessed May 23, 2019, http://www.getty.edu/research/tools/provenance/collaborators/index.html.

3. Provenance Index Remodel (Getty Research Institute), accessed May 21, 2019, http://www.getty.edu/research/tools/provenance/provenance_remodel/index.html.

4. Books published by the GPI are discoverable in the Getty Research Institute's online library catalog. See https://library.getty.edu for access.

© ARLIS, 2019 doi:10.1017/alj.2019.24

https://doi.org/10.1017/alj.2019.24 Published online by Cambridge University Press

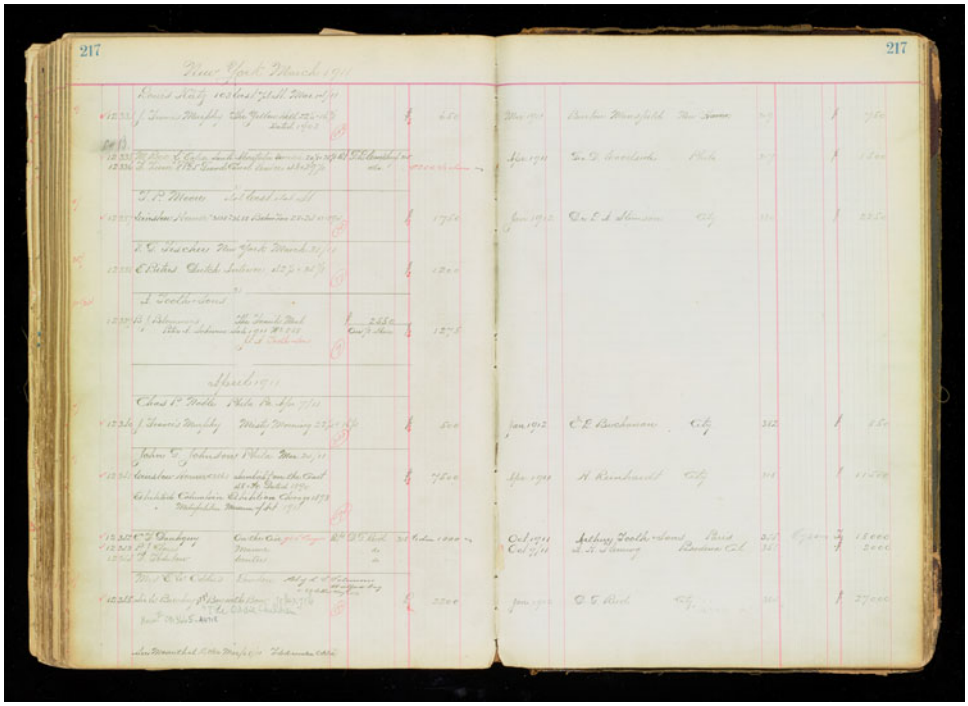


Fig. 1. Painting stock book 5: 8800-12652, 1899 April-1911 December. M. Knoedler & Co. records. Getty Research Institute, Los Angeles (2012.M.54)

provenance event with which it is associated. Or, perhaps these strings contains further background information about the artist culled from an editor's expert research. Recording metadata in this manner causes the information to lose its value. Standards about what should be recorded where, or how information should be recorded, differ from database to database, and a holistic view of the Index is only now coming to light.

The Getty Provenance Index remodel project has been a perfect and necessary opportunity to improve metadata structure and quality, and to rethink how to provide access to the relationships and lively histories contained within our data. It is a conceptual paradigm shift, as the events in the lives of these objects will move from flat data to relationships between the individual elements of data. This will allow us to make connections explicit in the way scholars and other specialists already do in their research. To achieve this shift, the data will be technically remodeled to Linked Open Data (LOD) using the Linked.art graph,[5] based on the CIDOC-CRM data model.[6] The LOD model has flexibility that the current system does not and can be used to express complex relationships inherent in our information. This means moving metadata to a graph-based model that expresses information as statements, or triples. However, in order for linked data to work effectively, we not only need a data model that suits the content we have, but the metadata itself needs to be standardized and tidy. Although the Index has been curated and edited frequently, the content variety and changes in editorial guidelines, staff, and preferences of the field have produced legacy data that needs standardization and, in some cases, must be restructured to fit with the CRM-based model. Without this necessary, iterative step of data cleaning, linked data will not work effectively.

While there are many components to the GPI remodel, data cleaning is one of the most essential and time-consuming. With clean, standardized data, enhanced with controlled vocabularies used widely in the field, one can analyze it more accurately and make informed assumptions about its meaning. When information is fielded correctly and metadata is consistently used, a user can more easily find answers to their queries and feel confident they discovered everything they sought. While some of this cleaning work can and has been automated, there is a large amount that requires a human with subject expertise. And, it's not always as straightforward as it may seem when planning a project. Often, what we see at the onset of a task is only the tip of the challenge. What will be necessary to

5. Linked Art, accessed May 21, 2019, https://linked.art/.

6. CIDOC-CRM, accessed May 21, 2019, http://www.cidoc-crm.org/.

https://doi.org/10.1017/alj.2019.24 Published online by Cambridge University Press



Fig. 2. From flat records to semantic triples, from a hierarchy to a web.

complete a task gets more complicated as soon as we start digging into what we have. For example, the remodeled GPI will be enhanced using external authorities, particularly the Getty Vocabularies. We thought it would be a simple task to reconcile nationalities of people in the Index to the Styles and Periods facet of the Art and Architecture Thesaurus (AAT) ®. Theoretically, this work could be automated, as unique values from the GPI such as 'Flemish' or 'British' would reconcile to their AAT value. But nationalities in the GPI suffer from various problems, including typos (e.g. Flemish might show up as 'Flenish', 'Flenusj', 'Fleisch', etc). There is also a fair amount of uncertainty in the values, represented by question marks in the data and either-or situations, that is unavoidable from verbatim transcription of texts. In addition, there may be data in the nationality field that is specifically not a nationality, like the word 'illegible'. Alternately, there may be legitimate data, but it is not always enough to define the nationality. Since the AAT is highly specific and disambiguated, a value like 'Roman' in the GPI, without any dates to provide context, could be either Roman Antique or Italian. Our data cannot always be matched one-to-one to the AAT without further research to achieve clarity. However, it is not always possible with 35 years of history to reconstruct what led an editor to enter 'Roman' as a nationality. We may not have access to the original document they viewed, or have data provenance. What appeared like a simple, reconciliation task took considerably longer than anticipated because of the necessary research involved. For these reasons, assessing standardization tasks can be difficult to put on a timeline or to estimate what skills will be needed to complete them. As this remodel project is unique to the GPI, there is no precedent or benchmarking to rely on. While this example regarding nationalities was not the most difficult to standardize, it illustrates how much variety there is in the data and how difficult it is to rely on an automated process.

A further example of standardization needed in the GPI's legacy metadata comes from the attribution modifier field. Attributing and re-attributing objects to different creators is common in art history, and the Provenance Index has approximately 200,000 individual records across the databases that contain attribution modifiers. Those records may have as many as five occurrences of the field, each corresponding to an artist, so the 200,000 must be multiplied to estimate how many modifiers exist. As with the other values in the GPI, most of the attributions were recorded verbatim from source material. They retain their original vernacular, so in addition to English, one may find French, Dutch or German. Beyond the language variations, the content and meaning of the modifiers varies greatly across the databases and, historically, the GPI did not follow a unified control list. In addition, a long-standing practice has been to include the referred-to artist within the text of the field, so that one may see confusing or repetitive values. While this seems strange at first glance, it served an important formatting purpose when book versions of the Provenance Index were published in the 1980s and 1990s. Given that the Linked.art model relies on the attribution modifier value as a pivot point to express relationships between creators, we needed to create and abide to a strictly controlled list.

The first step of this task was to export the existing values and use the web-based platform OpenRefine to strip excess name strings from the modifier itself. It was time-consuming, but by leveraging GREL (Google Refine Expression Language) and regular expressions, not impossible. Once the names were removed and unique values were isolated and identified, the next step was for the Provenance Index staff to discuss how the terms and phrases have traditionally

https://doi.org/10.1017/alj.2019.24 Published online by Cambridge University Press

	A	B	C	D	F
1	FRENCH	ENGLISH	GERMAN	DUTCH	AAT Term
2	attribué a, que l'on donne à, tient à	attributed/attributed to, given as	zugeschrieben	toegeschreven aan	attributed to
3	probablement	probably			attributed to
4		studio of			studio of
5		workshop of	werkstatt		workshop of
6	atelier de	atelier of			atelier of
7	d'après	after			after
8	copie	copy			after
9	copie d'apres	copy after	Kopie nach		after
10	copie par	copy by	Kopie von, Kopie		copy by
11	dans la manière de, dans le genre de, dans le goût de, gout de, dans le style de, de la manière de, à la maniere de, genre, genre de, imitation de, maniere, maniere de, style de, style, sur la mode de	manner of, imitation of, style of, in the manner of, style	Geschmack von, imitation nach, stil	Manier	manner of; style of
12	changé de/ au lieu de	changed from	geändert von		formerly attributed to
13	changé à / remplacé par	re-attributed/changed to	geändert in, geändert zu		attributed to
14	ecole de, dans l'école de, de l'Ecole de, ecole	school of	schule der		school of
15	élève de/ disciple de	student of/ disciple of		discipel van	pupil of
16		follower of		volger van	follower of
17	cercle de	circle of	kreis		circle of
18	peut-être le	possibly by			possibly by OR attributed to depending on context
19					

Fig. 3. In progress standardization work shows unique attribution modifier values mapped across languages and to their AAT term.

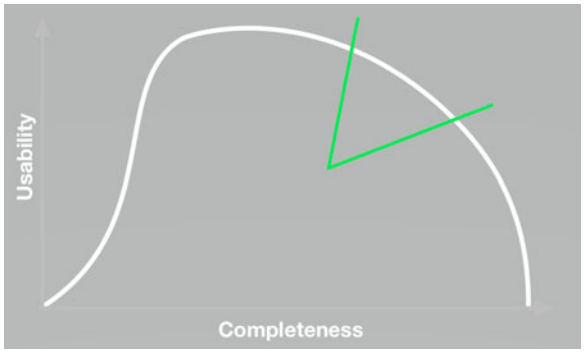


Fig. 4. The sweet spot of usability versus completeness. Image courtesy Robert Sanderson.

been used, and what they believed the true conceptual meanings to be. As the GPI remodel is a conceptual shift as much as a technical shift, standardizing the modifiers meant defining the modifiers in ways that did not exist previously. The goal was to match the meanings across languages and identify similar phrases so that we could map modifiers to corresponding AAT values. For example, the German word *bezeichnet*, which translates to 'signed', could indicate the object is identified as by a specific artist. We also had a challenge with data from sources that were contradictory or corrections made by hand on a printed auction catalog. This was recorded in the Index as 'changed to' and 'changed from', and required some thought before we could settle on describing this as 'formerly attributed to'. Another challenge was 'manner of' versus 'style of'. The AAT defines 'style of' as 'a work by an unknown creator whose style is strongly under the influence of the style of the named master',[7] and 'manner of' as 'a work by an unknown creator whose style or elements of whose style are somewhat close to the style of the named artist'.[8] The editors did not think the usage of those terms across the GPI aligned with the AAT, so we decided to use both of them for objects tagged with either. Users may not make a distinction between the two similar phrases, but using both as tags ensures they will still find what they are looking for. A significant lesson that came out of this cleaning work is that it is important to convey

7. "Style of (attribution Qualifier),", accessed May 21, 2019, http://vocab.getty.edu/aat/300404285.

8. "Manner of (attribution Qualifier).", accessed May 21, 2019, http://vocab.getty.edu/aat/300404288.

https://doi.org/10.1017/alj.2019.24 Published online by Cambridge University Press

the conceptual meaning, but not always important to retain specific legacy language. Respect for the original order does not need to extend to respect for original chaos.

By the end of the attribution modifier work, the extensive variety was trimmed down and mapped to only a handful of AAT values. The verbatim fields were left as is and an additional authority field was added to contain newly mapped terms, which are key to the transformation to linked data. It is not a perfect fix, as compound modifiers continue, e.g. 'copy of the school of Rembrandt'. Standardization work touches on art historical practice, i.e. who made the decision for this attribution? What motivation did they have? Is 'copy after the school of' a useful designation for historians? The standardization is necessary for transformation, but also must be useable within the linked data model. It is risky to attempt to standardize and model every detail, as there is a point when the context is lost and the data model becomes unusable. Finally, metadata cleaning has ramifications to every part of the remodeled Index, from discoverability of information to the design of the user interface. It is not always possible to see where this impact will fall. The high-level decisions made during the course of the project, such as reconciling to the Getty Vocabularies, will have continuing influence on metadata standardization throughout the life of the remodel and beyond. Finding the sweet spot in terms of data quality is our goal: to clean the data enough that is useful. What useful means may change over time, but having it within your sights can help to answer the many questions that arise during this challenging process.

Kelly Davis
Metadata Specialist I
Getty Research Institute
1200 Getty Center Drive Suite 1100
Los Angeles, 90068
California
USA
Email: kkdavis@getty.edu

https://doi.org/10.1017/alj.2019.24 Published online by Cambridge University Press