Linking (In)Completeness: A Collaborative Approach to Representing People in Art Provenance Data

Rother, Lynn
lynn.rother@leuphana.de
Leuphana University Lüneburg, Germany

Mariani, Fabio
fabio.mariani@leuphana.de
Leuphana University Lüneburg, Germany

Koss, Max
max.koss@leuphana.de
Leuphana University Lüneburg, Germany

Provenance records, as maintained by museums, capture the historical sequence of ownership and socio-economic custody changes of an artwork, the parties engaged in them, and their role in the transactions. These records usually include names, dates, places, and other transactional details (Yeide et al., 2001). With museums increasingly engaging in a structured data environment, such information allows us to observe and analyze human activity relating to artworks across time and geography and at scale, providing insights into the displacement and circulation of artworks in times of war and peace.

To support museums in their cataloging in a structured data environment, the Getty Research Institute hosts and edits the Getty Union List of Artist Names (ULAN), a controlled thesaurus of biographical information about artists and other parties, both individuals and groups (Harpring, 2018). As of 2015, ULAN is published as linked open data, in line with the FAIR principles of findability, accessibility, interoperability, and reusability, making it an authoritative repository for museums to link their data to. [1]

To address the unevenness of biographical details in provenance records within and across museums, they can enrich their data by linking it to ULAN. At the same time, museums often create detailed records of locally important but generally lesser-known individuals that are missing from ULAN. This expertise on locally known actors, in turn, can complement the records of other institutions.

Using collection data from the Art Institute of Chicago, our paper shows how the issue of uneven knowledge about people can be addressed by linking entities to ULAN. Given the incomplete nature of controlled vocabularies, we advocate that museums not only use such data but recognize their role as providers of expert knowledge. Because of their expertise, museums can become crucial stakeholders in a structured provenance data environment. By collaborating with ULAN, museums can counteract biases our analysis has uncovered.

Analysis

Although digitizing provenance is still a work in progress, some museums allow access to provenance texts through APIs and data dumps. Among these pioneering museums is the Art Institute of Chicago, which makes data available for 122,317 artworks. [2] Of those, 11,392 have provenance texts. In Rother et al., 2023, we introduced the use of deep learning to extract knowledge from these texts automatically. We divided the process, attributable to an event extraction problem, into two tasks. First, Sentence Boundary Disambiguation divided the text into discrete events according to punctuation. The second task concerns Span Categorization. This task extracted from each provenance event numerous pieces of information regarding a party, such as type (group or person), gender, and role in the event, but also biographical information such as name(s), birth and death dates, and locations (figure 1).

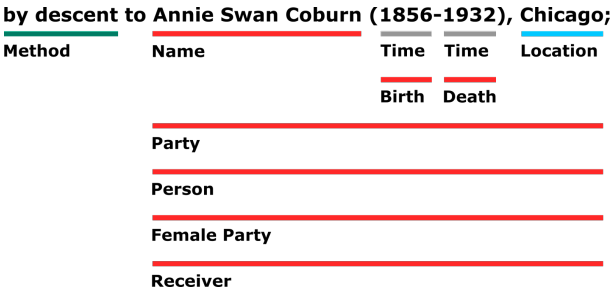

Figure 1. Example of Span Categorization based on a single provenance event of the Art Institute of Chicago's drawing *Two Women*, c. 1840, by Constantin Guys (https://www.artic.edu/artworks/14825/two-women).

In total, we extracted 8,817 unique parties. Of these, 5,941 were identified as persons and 2,304 as groups. [3] Among the parties recognized as persons, 25.2% (1,498) were female parties. In analyzing the additional biographical information extracted through Span Categorization, we observed degrees of completeness of the parties' biographies. For life dates, 14.9% of the parties had a birth date (or formation, in the case of groups), while 23.8% had a death date (or dissolution). Moreover, 68.1% of the parties were associated with a location (6,007 parties). Only 11% (972 parties), however, can be considered complete, as they include life dates and at least one location.

Given the high number of incomplete parties in Art Institute data, we tried to link those parties to matching entities in ULAN. As a result, we found the matching entity for 1,296 parties in Art Institute provenances (14.7% of 8,817 unique parties). This surprisingly low number notwithstanding, the parties that did match often appeared across multiple Art Institute provenances. In fact, the 1,296 Art Institute parties represented in ULAN were involved in 11,876 provenance events (9.2 events per party), while the 6,587 Art Institute parties not yet included in ULAN were involved in 20,909 events (3.2 events per party). [4]

Finally, we also found a gender gap. Of the Art Institute/ULAN matches, 75.4% involved individuals (977), of which only 14% can be identified as women (137). The results were in line with the current gender distribution in ULAN. In fact, of the 333,977 ULAN entities identifiable as persons, 44,526 (13.3%) are identified as female. [5] However, this contrasts with the gender distribution in the Art Institute data, where the number stands at 25.2%– almost double the ratio of the dominant controlled vocabulary in the cultural field.

Discussion

While an institution such as the Art Institute can enrich its data on parties through ULAN, many individuals recorded in its provenances are not yet represented in this controlled vocabulary. Given the often idiosyncratic development of collections shaped by locally significant individuals, it is the museums that can best contribute their knowledge on these missing figures, however fragmentary, to such thesauri, in turn helping to complete the records of other museums.

The higher percentage of female parties in the Art Institute data indicates that institutions can address endemic biases in thesauri through structured provenance data. This requires that museums not only use ULAN but act as trusted experts providing vetted information. The digital future of provenance relies on sustained collaboration between the various stakeholders that goes beyond mere entity linking and involves the active knowledge exchange of researchers.

Notes

1. https://www.getty.edu/research/tools/vocabularies/ulan/.
2. The dataset was downloaded on April 7, 2022 (https://github.com/art-institute-of-chicago/api-data).
3. The 6.5% of parties (572) are ambiguous since they cannot be classified as either individuals or groups.
4. Of the 8,817 unique parties in Art Institute data, multiple valid candidates were found for 362 of them in ULAN (4.1%). Since they would require disambiguation, we excluded them. These parties were involved in 7,364 events in the data (20.3 events per party). Indeed, these parties include highly active auction houses such as Sotheby's and Christie's.
5. Queries were made on the Getty SPARQL endpoint (http://vocab.getty.edu/sparql) on April 21, 2023.

Bibliography

Harpring, Patricia (2018): "Linking the Getty Vocabularies: The Content Perspective, Including an Update on CONA", in: *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)* , San Francisco, CA, 2018: 1–8. DOI: 10.23919/PNC.2018.8579460 .

Rother, Lynn / **Mariani, Fabio** / **Koss, Max** (2023): "Hidden Value: Provenance as a Source for Economic and Social History", in: *Economic History Yearbook* , 64, 1: 111–142. DOI: 10.1515/jbwg-2023-0005 .

Yeide, Nancy H. / **Walsh, Amy L.** / **Akinsha, Konstantin** (2001): *The AAM Guide to Provenance Research* . Washington, DC: American Association of Museums.