People Information in Provenance Data: Biographical Entity Linking with Wikidata and ULAN

Informacije o osobama u podacima o provenijenciji: povezivanje biografskih entiteta s bazama Wikidata i ULAN

IZVORNI ZNANSTVENI RAD
Primljen: 13. veljače 2024.
Prihvaćen: 28. svibnja 2024.
DOI: https://doi.org/10.31664/zu.2024.114.07

ORIGINAL SCIENTIFIC PAPER
Received: February 13, 2024
Accepted: May 28, 2024
DOI: https://doi.org/10.31664/zu.2024.114.07

SAŽETAK

Integriranje podataka o provenijenciji u okvir povezanih otvorenih podataka (engl. *linked open data,* LOD) predstavlja znatnu priliku za muzeje da unaprijede transparentnost, olakšaju istraživanje i pridonesu širem digitalnom ekosustavu informacija iz povijesti umjetnosti. U ovom se radu istražuje proces povezivanja biografskih entiteta, s fokusom na pojedince povezane s umjetničkim djelima, koristeći se evidencijom o provenijenciji Instituta za umjetnost u Chicagu kao studijom slučaja. Primjenom standarda LOD-a i resursa kao što su baze Wikidata i Union List of Artist Names (ULAN), rad istražuje kako muzeji mogu iskoristiti digitalne platforme za otključavanje znanja koje je prethodno bilo pohranjeno u muzejskim bazama podataka. Rad započinje isticanjem važnosti podataka o provenijenciji, kojima se dokumentiraju promjene vlasništva i skrbništva nad umjetničkim djelima tijekom vremena. Usvajanje pristupa povezanih otvorenih podataka o provenijenciji (engl. *provenance linked open data,* PLOD) omogućuje institucijama da postanu transparentnije u pogledu podrijetla svojih zbirki, podržavajući napore koji se ulažu u osiguravanje povijesne pravde i restituciju. Nadalje, prihvaćanjem standarda LOD-a muzejima se omogućuje sudjelovanje u rastućem digitalnom ekosustavu informacija iz povijesti umjetnosti, potičući suradnju i razmjenu znanja između institucija.

→

ABSTRACT

This paper discusses how provenance data can be integrated into a linked open data (LOD) framework. It focuses on the biographical information of people recorded in provenance texts of museums. The Art Institute of Chicago's provenance records serve as a case study to examine the process of entity linking. This process helps to connect individuals mentioned in provenances with entries in LOD repositories like Wikidata and the Getty's Union List of Artist Names (ULAN). The paper evaluates the effectiveness of entity linking through quantitative and qualitative analyses and discusses the role of museums as both a user and a contributor to LOD repositories. The findings emphasize the importance of accurate data representation, particularly regarding underrepresented groups like women, and highlight the potential for museums to enrich LOD platforms with authoritative biographical information.

KEYWORDS
provenance, provenance data, entity linking, biographical data, Wikidata, Union List of Artist Names, provenance linked open data

Fabio Mariani
Max Koss
Lynn Rother

Institut za filozofiju i znanost o umjetnosti, Sveučilište Leuphana u Lüneburgu
/ Institute of Philosophy and Art History, Leuphana University Lüneburg

U fokusu je rada ispitivanje biografskih informacija o pojedincima koji se spominju u podacima o provenijenciji. Ti pojedinci mogu uključivati povijesne vlasnike, skrbnike, trgovce, članove obitelji i druge relevantne strane povezane s umjetničkim djelima. Analiziranjem evidencije podataka o provenijenciji Instituta za umjetnost u Chicagu, rad identificira više od 5000 različitih strana, naglašavajući raznolikost i složenost biografskih podataka unutar muzejskih zbirki.

Proces povezivanja entiteta uključuje povezivanje pojedinaca spomenutih u podacima o provenijenciji s unosima u repozitorijima LOD-a kao što su Wikidata i ULAN. Wikidata, kolaborativna baza znanja, pruža strukturirane podatke o različitim temama, uključujući biografske informacije. Istodobno, ULAN služi kao repozitorij specifičan za pojedinu domenu koji vodi Gettyjev istraživački institut. Komparativnom analizom rad ocjenjuje uloge i strateške doprinose obiju platformi u kontekstu PLOD-a.

Provođenjem kvantitativnih i kvalitativnih analiza procjenjuje se učinkovitost povezivanja entiteta. Rad otkriva da, premda Wikidata omogućuje znatan broj podudaranja, generalistički opseg te baze može dovesti do dvosmislenosti. Nasuprot tome, ULAN pokazuje manju dvosmislenost, ali pruža manje podudaranja zbog svojeg specijaliziranog fokusa. Validacija putem biografskih datuma pomaže razjasniti podudaranja i osigurava točnost u povezivanju entiteta. Nadalje, rad istražuje funkcije povezivanja entiteta, uključujući normativnu kontrolu i obogaćivanje podataka. LOD olakšava dosljednost i interoperabilnost podataka dodjeljivanjem jedinstvenih identifikatora entitetima i njihovim povezivanjem kroz različite repozitorije. Osim toga, platforme temeljene na LOD-u nude mogućnosti za obogaćivanje podataka, omogućujući istraživačima istraživanje odnosa među pojedincima i rekonstrukciju društvenih mreža.

Analiza također razjašnjava ulogu muzeja kao korisnika i pružatelja biografskih informacija unutar ekosustava LOD-a. Iako muzeji imaju koristi od vanjskih repozitorija, također igraju ključnu ulogu u obogaćivanju LOD platformi vjerodostojnim podacima. Rad naglašava važnost rješavanja problema podzastupljenosti, posebno kada je riječ o ženskim osobama, i ističe potencijal muzeja da pridonesu vrijednim uvidima u repozitorije temeljene na LOD-u.

Zaključno, rad pokazuje potencijal povezivanja biografskih entiteta u poboljšanju dostupnosti i korisnosti podataka o provenijenciji unutar digitalnog krajolika. Prihvaćanjem standarda LOD-a i suradnjom s vanjskim repozitorijima muzeji mogu unaprijediti istraživanja, promicati transparentnost i obogatiti sektor kulturne baštine.

KLJUČNE RIJEČI
provenijencija, podaci o provenijenciji, povezivanje entiteta, biografski podaci, Wikidata, Union List of Artist Names, povezani otvoreni podaci o provenijenciji

INTRODUCTION

A linked open data (LOD) strategy for the digital transformation of museum records helps institutions respond to the cultural, social, and technological changes they are facing. With individually identified online resources linked to other such resources, LOD promises to unlock knowledge hitherto siloed in museum databases, not least provenances, the records of ownership and socio-economic custody changes of artworks, and the focus of this paper. Indeed, a provenance linked open data (PLOD) approach helps institutions become more transparent about the origins of their collections, facilitating efforts at redressing historical injustices and restitution.[1]

Adopting LOD standards also allows museums to benefit from, participate in, and help shape a burgeoning digital ecosystem of art historical information produced by experts across institutions from around the globe. The benefits of a web-based knowledge infrastructure range from synergies in research efforts (i.e., all objects that once belonged to a collector that are now dispersed could be easily identified with a single query) to eliminating research redundancies through sharing knowledge produced by one institution with the wider museum and research community. Furthermore, pursuing a PLOD strategy creates research opportunities for disciplines further afield, such as economic and social history, for example.[2]

This paper addresses the most salient type of information in provenance and the potential of its transformation into LOD: facts about people. Most often, these people are the historical owners or custodians of a work. The facts about them may include their names, honorifics and titles, life dates, and location. Biographical facts in provenances can extend to information about dealers, family members, gallerists, government officials, intermediaries, mentors and teachers, military personnel, or any other person who may have been recorded as relating to the object in an ownership or custody role in the course of its life, whether they occupied these roles lawfully or not.

Using the provenances published online by the Art Institute of Chicago as a case study, we examine not only the data about people in them but also the existing LOD ecosystem within which they can be linked. Entity linking is the process of connecting identical facts recorded in different locations on the web, such as websites, data repositories, etc. It connects a specific dataset with the LOD ecosystem, contributing to integrating and navigating diverse information sources from around the globe.

In this paper, we explore and chart the potential of entity linking of individuals recorded in the provenances of the Art Institute of Chicago with entries from Wikidata and the Union List of Artist Names (ULAN). Wikidata, operated by the Wikimedia Foundation, is a collaborative knowledge base that crowdsources structured data on a wide array of topics,

[1] Currently, museums record provenance as free text, making it arduous to analyze historical information automatically. For an in-depth analysis of the current state of provenance records in museums, the problem of data siloing, and the opportunities of provenance linked open data, see: Rother, Koss, Mariani, "Taking Care of History."

[2] A preliminary study of structured provenance data showed the potential for economic and social history analysis. We found, for example, that women who inherit an artwork are more likely to gift or donate it than men, who are more likely to sell an inherited artwork. For more insights into the method and analysis, see: Rother, Mariani, Koss, "Hidden Value."

including biographical information.[3] ULAN, for its part, is a shared expert resource curated and maintained by the Getty Research Institute as part of their Getty Vocabulary Program that includes other LOD-based resources. ULAN stands as an authoritative repository specifically designed for information about individuals associated with the art world, containing detailed biographical records.[4]

In the following, we first recapitulate briefly the steps required to digitally analyze biographical information in provenances. This is a necessary step to, secondly, identify the quantity and quality of biographical information available for linking. Thirdly, we elucidate the distinctive roles and strategic contributions of Wikidata as a community-driven database and of ULAN as a domain-specific repository in the context of PLOD through a comparative analysis of both.[5] Lastly, we emphasize the role museums occupy in the LOD ecosystem as both users and valuable contributors to shared repositories, especially regarding information stemming from provenance research.

PEOPLE RECORDS: AN ANALYSIS OF PROVENANCES AT THE ART INSTITUTE OF CHICAGO

The Art Institute of Chicago, founded in 1879 and one of the largest museums in the United States, is a pioneering institution in sharing provenances on its collection website and making them available for download. It also adheres to the provenance guidelines of the American Alliance of Museums (AAM), outlined in 2001. Published in response to the watershed 1998 *Washington Conference Principles on Nazi-Confiscated Art* that codified in a legally non-binding way the measures for sustained provenance practice and increased transparency, the *AAM Guide to Provenance Research* provides a set of rules on how to write provenance records.[6] They should be structured as a chronological list of sentences, each documenting a specific provenance event. Each event encompasses information about parties, methods of transfer, dates, and locations. The AAM guidelines also recommend including life dates in parentheses when recording parties.

While the AAM guidelines provide a set of rules for humans to record provenance information, they were not written with machine readability in mind, a prerequisite for LOD. They are both too flexible in their implementation by individual institutions and too human-reader-oriented for automatically extracting and analyzing information. However, they provide a systematic baseline of structure through their emphasis on sentences, a set of required elements in a provenance event, and specific punctuation rules.

The Art Institute provenance dataset that we were thus able to build contains 11,392 provenance texts divided into 35,554 distinct provenance events.[6] Because they are AAM-compliant provenances, we can extract information from them with the help of two natural language processing tasks performed by deep learning models.[8] The first task, sentence boundary disambiguation, divides provenance texts into discrete provenance events. The second task, span categorization, identifies and classifies text segments using a set of tags described by a domain-specific annotation scheme.[9] In particular, it allows us to extract information about the parties mentioned in each provenance event.

Each party is automatically categorized as either a person or a group by applying the "group" or "person" tags. If classified as a "person," explicative details such as "female party" are also registered. The "party" text segment contains additional biographical information. Thus, the entire name of the party is annotated with the "name" tag. The annotation scheme also enables the extraction of life dates.

Given the variability of biographical details recorded in Art Institute provenances, we must first establish what to include and what to exclude from the people data in the dataset. Indeed, when multiple individuals act together, achieving consistent disambiguation of each person, if at all possible, is a challenge. For instance, one of the parties might be documented in a way that the machine cannot comprehend, identifying it by its first name only (e.g., "Mary and Leigh Block"). To complicate matters further, it is impossible to determine a priori whether the two individuals are a couple, siblings, or business partners. Moreover, a couple may be recorded using only honorifics (i.e., "Mr. and Mrs. Harry L. Winston"). While in this case, it is clear that the group represents a couple, such conventional recording is associated with heterosexual marriages and consistently conceals the identity of the female partner. It also belies and reinscribes a strictly binary understanding of gender.

Overcoming such ambiguity in recording necessitates human intellectual intervention to ensure accurate representation of such information as data, particularly in cases such as wives within couples that may be misrepresented. Given these recording issues, our analysis focuses on people who are recorded as having acted alone.

Focusing solely on such individuals, we have identified 5,147 distinct parties for analysis. The term "distinct" points to the preliminary reconciliation process we implemented to merge those extracted parties referring to the same individual. For two parties to be considered identical, we decided they must share at least one name and have the exact birth and death years (if available). Through span categorization, we identified 1,188 distinct female parties, constituting approximately 23.1% of the total individuals in the dataset.

As we have noted, the AAM guidelines recommend recording the life dates of individuals. This information is valuable for at least two reasons. Firstly, in provenance research, a person's life dates can help establish periods of ownership as they mark the temporal limits within which such ownership is possible. For instance, if there is no clear ownership period, we know that the owner either separated from an

[3] Vrandečić, Krötzsch, "Wikidata: A Free Collaborative Knowledgebase."

[4] Harpring, "Development of the Getty Vocabularies."

[5] Comparisons between cultural heritage LOD repositories have been examined in several studies. Sugimoto, "Instance Level Analysis on Linked Open Data Connectivity" compares platform connectivity, including Wikidata and ULAN, across different aspects, including people. A comparison between platforms focused on individuals is found in Freire, Manguinhas, Isaac, "An Observational Study of Equivalence Links" and Goldfarb, Merkl, "Visualizing Art Historical Developments." Context-specific comparisons based on the analysis of a dataset in relation to Wikidata and ULAN have been conducted, for example, in Faraj, Micsik, "Persons, GLAM institutes and collections," in the context of the COURAGE registry.

[6] Yeide, Akinsha, Walsh, *The AAM Guide to Provenance Research*.

[7] Rother, Mariani, Koss, "Hidden Value."

[8] The experiment was carried out using Art Institute of Chicago data downloaded from the museum repository on April 7, 2022. The sentence boundary disambiguation model achieved an F1 score of 0.99, while the span categorization model achieved an F1 score of 0.94. See: *Ibid.*

[9] Mariani, Rother, Koss, "Teaching Provenance to AI."

object before their date of death or passed it on to heirs after death. Additionally, as mentioned, life dates are valuable elements for disambiguating individuals in the data extraction process. The date of birth or death is available for 36% of individuals in our dataset. Among them, 19.9% have information on both birth and death date, while only the death date is recorded for 15%, and in rare cases, only the birth date is available (1.1%).[10]

Besides excluding groups from our entity-linking experiment and using life dates to disambiguate individuals, a third element to assess is the names of individuals. Span categorization enables the extraction of one or more names recorded for the same individual (e.g., "Jean Baptiste Théophile, also known as Théophile Bascle"). In our dataset, 9.7% of individuals (497 entities) are documented with more than one name.

Notably, out of these, 245 (49.3%) are female parties. Various reasons can account for a person being recorded with multiple names, including holding names of nobility (e.g., "Lord Francis Egerton, 1st Earl of Ellesmere (1800–1857)") or religious names (e.g., "Fabio Chigi, later Pope Alexander VII (died 1667)"). However, the high prevalence of female individuals with multiple names when their overall share of individuals is significantly lower can be explained by the differentiation between maiden and married names (e.g., "Mrs. John Alden Carpenter (née Ellen Waller Borden)"). While this recording practice expresses bias and conventions, having multiple names for the same person facilitates entity disambiguation and reconciliation.

In light of this, it is crucial to consider how female individuals are named in provenance texts. A total of 449 female individuals (37.8% of all female parties) are recorded with at least one name containing an honorific (e.g., "Mrs.," "Ms.," or "Madame"). For 306 female individuals (25.8%), the name with the honorific is the only recorded name. In these cases, the honorific likely includes the husband's name (e.g., "Mrs. H. Harris Jonas"), compromising the accurate representation of the woman.

FINDING THE RIGHT MATCH: A QUANTITATIVE AND QUALITATIVE APPROACH

Having identified the individual parties that may be potentially linked, we can now investigate the potential of entity linking with online resources such as Wikidata and ULAN. Due to the ambiguous recording of names and the limited biographical information in provenance records, following a two-step match discovery process involving quantitative and qualitative approaches became necessary.

In the quantitative stage, we automatically selected matching candidates in Wikidata and ULAN for the 5,147 distinct entities in our dataset. Our criteria for identifying potential matches involved selecting entities from each repository that shared at least one exact name match with those we extracted from the dataset.[11] We did not consider biographical dates at this stage due to their unavailability for all entities. For the 5,147 distinct individuals extracted from the provenance records, Wikidata provided a potential match for 2,239 (43.5%). Within these, 1,461 involved a single candidate (65.3%), while 778 were ambiguous as they included multiple candidates.

The scenario differed starkly for potential matches with ULAN. In this case, we identified at least one potential match for 1,064 individuals (20.7%). Despite ULAN providing far fewer potential matches, the results were less ambiguous. Of the potential matches, 940 involved only one candidate (88.3%), and 124 involved multiple candidates, a significantly lower number than obtained for Wikidata.

When focusing the comparison on female parties only, the results were significantly poorer. While 23% of the individuals in our dataset were identifiable as female, only 15.1% of unambiguous matches with Wikidata involved a female party. Conversely, for ULAN, this percentage dropped to 12.1% of unambiguous matches. Both figures are stark expressions of the underrepresentation of female parties in the crowdsourced, as well as in the expert-sourced repository.

Acknowledging the unreliability of names as a means to establish definitive matches, the second phase of our analysis involved validating the potential matches of individuals using biographical dates. We validated the potential match between two entities with the same full name if there was at least one coinciding biographical date (birth or death).[12] While this approach reduces the number of entities under analysis, it enables a qualitative evaluation of the experiment.

Of the 1,461 unambiguous Wikidata matches, 698 entities (47.8%) are documented with birth or death dates in both Wikidata and Art Institute records. In this case, comparing birth or death years confirmed the matches for 624 entities (89.4%). Match validation through biographical dates allows the assessment of ambiguous cases involving multiple potential matches. Out of the 778 Art Institute entities that matched with more than one entity in Wikidata, it was possible to disambiguate the proper match for 210 individuals, accounting for circa 27% of ambiguous cases.

We applied a similar approach to the 940 unambiguous ULAN matches. Here, we obtained 500 matches (53.2%) for which the date of birth or death can be found in Art Institute records. Out of these, the comparison with life dates confirmed 432 matches (86.4%). Biographical dates contributed to disambiguating 38 of the 124 ambiguous matches (30.6%), a low number reflecting the relatively low overall occurrence of ambiguous ULAN matches.

The match discovery process outlined in this section highlighted the capabilities of Wikidata and ULAN in a PLOD context. While Wikidata facilitates the matching of a substantial

[10] Provenance records occasionally exhibited discrepancies in life dates, with 16 individuals having multiple birth dates and 25 individuals having multiple death dates. These variations can be attributed to disagreements among different authors of provenance records. All recorded dates were considered during the analysis.

[11] The selection process was conducted using OpenRefine reconciliation API services and SPARQL queries to the respective platform endpoints on February 5, 2024. To be selected as a candidate, an entity needed to have a label or an alternative label identical to one of the names extracted for the entity in the provenance records (including titles and abbreviations). Similarity was calculated by considering word order variations.

[12] Every piece of information analyzed, including years of birth and death, was acquired through SPARQL queries to the respective Wikidata and ULAN endpoints on February 6, 2024. On these platforms, disagreements related to birth and death dates can be found. Therefore, multiple dates were taken into account when necessary.

number of entities, its generalist encyclopedic scope makes it prone to ambiguity. Addressing this limitation would require museums to consistently record individuals with their biographical dates, providing a means for disambiguating homonymous entities. In contrast, ULAN exhibits lower ambiguity but also fewer matches due to its smaller size. In the end, we successfully established entity linking for 890 entities in total, using life dates for validation. Of these, 834 entities were linked with entries in Wikidata, and 470 entities were linked with entries in ULAN (Fig. 1).

ENTITY LINKING: AUTHORITY CONTROL AND DATA ENRICHMENT

In the context of a LOD framework, linking entities from a museum's provenance records with those of external repositories serves two distinct functions: authority control and data enrichment. Authority control functions by assigning unique Uniform Resource Identifiers (URIs) to each entity. When entities are linked to Wikidata or ULAN entries, the relevant URI from these platforms is allocated to the entity in question. This practice ensures data consistency and fosters interoperability by assigning identical URIs to identical entities from different repositories.

By the same mechanism, Wikidata and ULAN achieve interoperability by sharing URIs, as their respective entities are linked to one another. This aspect helps refine the entity linking when an individual is linked to an entity in only one of the two repositories. Consequently, 41 individuals linked to Wikidata entries gained entity links to ULAN, and 19 individuals linked to ULAN entries gained entity links to Wikidata.

Furthermore, authority control enabled a new reconciliation process for entities within the Art Institute dataset. This process identified and reconciled 33 pairs of entities that, although referring to the same person, were recorded differently. This reconciliation was feasible due to the pairs being linked to the same entity in Wikidata or ULAN. In light of this, the entities under consideration for entity linking dropped from 890 to 857.

The second function of entity linking involves data enrichment, i.e., acquiring new information from linked platforms. Each platform's contribution criteria, whether open to crowdsourcing or limited to authoritative contributors, significantly shape the available information.

From a data enrichment standpoint, both Wikidata and ULAN enable the exploration of individuals' relationships, facilitating the reconstruction of social networks. In the context of provenance, this approach proves valuable for comprehending personal relationships, such as understanding inheritances within a family, and scrutinizing professional relationships, like uncovering market networks between dealers and collectors.[13]

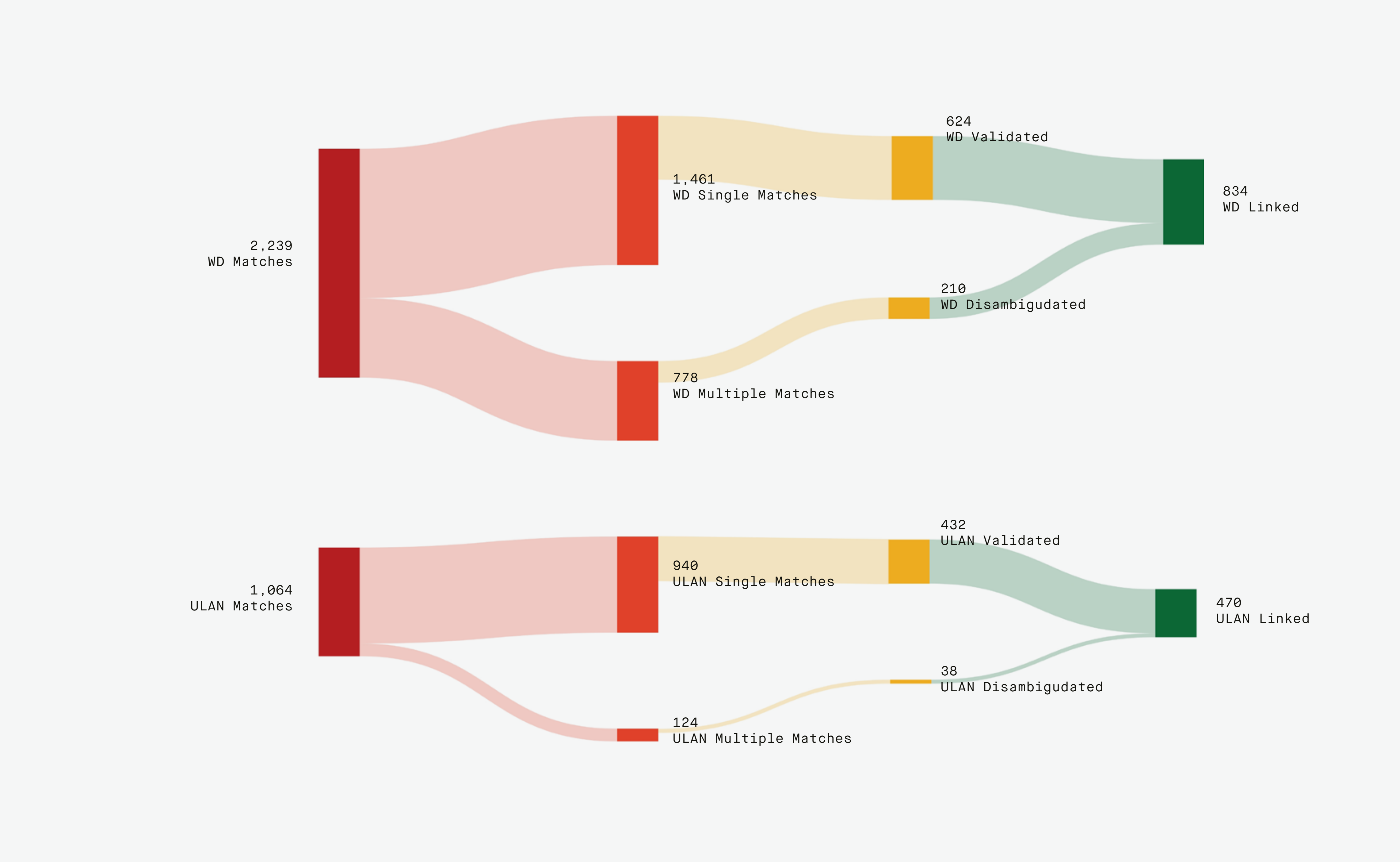

Fig. / Sl. 1 Sankey diagram summarizing the process of match finding, validation, disambiguation, and entity linking for both Wikidata (WD) and ULAN entities. / Sankeyjev dijagram sa sažetim prikazom procesa pronalaženja podudaranja, provjere valjanosti, razjašnjenja i povezivanja entiteta za entitete u bazama Wikidata (WD) i ULAN.
↑

[13] An example of network analysis applied to the study of the art market can be found in Schich et al., "Network Dimensions in the Getty Provenance Index."

Wikidata records relationships for 537 entities, averaging 4.5 relationships per individual. In ULAN, 201 entities have at least one relationship, with an average of 6.8 relationships per individual.

By categorizing relationship types, a distinct contrast emerges between the two repositories. These relationships can be categorized into three main groups: personal (such as family ties, friendships, and romantic engagements), educational (including master-student relationships), and professional (encompassing roles like client, collaborator, patron, or associate).[14] An examination reveals a significant disparity between Wikidata and ULAN regarding personal and educational relationships. Within Wikidata entities, a significant majority (79.9%) of recorded relationships pertain to personal ties, with only a minority (13.8%) involving educational relationships. Conversely, ULAN exhibits a higher emphasis on educational relationships (65.8%) and less focus on personal relationships (19.7%).[15]

The occupational type of the entities under analysis may offer an explanation of this trend. According to ULAN, 231 linked entities are classified as "visual artists" (47.3%). This suggests that entity linking is notably biased toward individuals known in the art world as artists themselves. This pattern becomes even more pronounced when considering the 55 linked female parties, among which 31 (56.4%) are recorded as "visual artists." ULAN, at least in the context under analysis, remains predominantly focused on entries related to artists. Despite its designation as the "Union List of Artist Names," ULAN's scope encompasses any individual associated with the art world, potentially including those present in the provenance records of institutions like the Art Institute.

From a social network standpoint, the information available in Wikidata and ULAN reflects their respective contributors. Wikidata's wider, essentially public user base exhibits a tendency to record personal relationships, which are often easily available and less controversial. Conversely, ULAN, with its institutional, specialized, and, above all, purposefully selected user base, displays a keen interest in academic aspects such as the relationships between individuals, predominantly artists, and their teachers and students.

LINKING INSTITUTIONS: THE MUSEUM AS PROVIDER OF BIOGRAPHICAL INFORMATION

Given the networked structure of LOD, anyone participating occupies a dual role as a provider and user of information. Museums, therefore, not only rely on external repositories, but they also serve as an expert source of reliable information related to their collection.

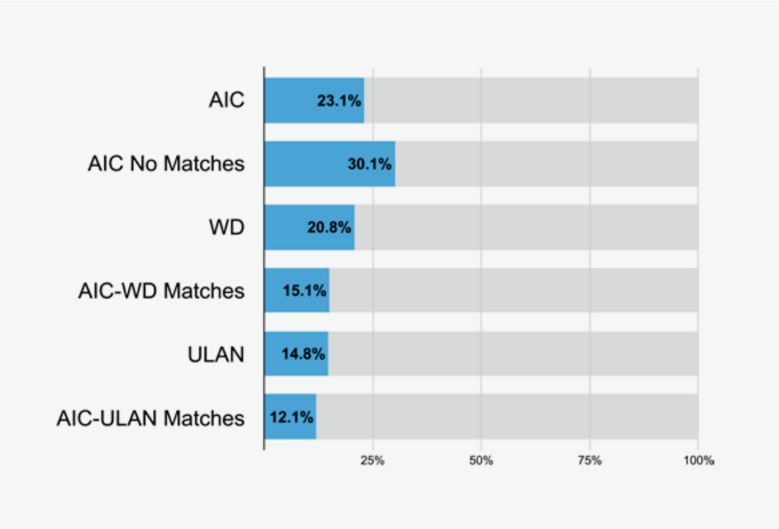

Fig. / Sl. 2 Female parties representation across Art Institute of Chicago (AIC), Wikidata (WD), and ULAN. / Zastupljenost ženskih osoba u bazi Instituta za umjetnost u Chicagu (AIC) te bazama Wikidata (WD) i ULAN.
↑

INDIVIDUAL	NUMBER OF EVENTS	WIKIDATA	ULAN
Eduard Gaffron (1861–1931)	892	✓	
William F. Dunham (1857–1936)	720		
Reverend Chauncey Murch (1859–1907)	406	(✓)	
Nathan Cummings (1896–1985)	293	✓	
B. J. Wassermann (Bruno John)	282		
Martin A. Ryerson (d. 1932)	272	✓	✓
Mrs. William Nelson (Helen T.) Pelouze (1866–1953)	271		
Dorothy Braude Edinburg (1920–2015)	248	✓	
William F. E. Gurley (1854–1943)	195	✓	
Francis H. Bacon (1856–1940)	147	✓	✓
Émile Brugsch (1842–1930)	140	✓	
Charles Deering (1852–1927)	130	✓	✓
E. M. (Pete) Bakwin	122		

Tab. / Tab. 1 Table of the 13 individuals documented in the provenance records of the Art Institute of Chicago who participated in more than 100 provenance events. / Table of the 13 individuals documented in the provenance records of the Art Institute of Chicago who participated in more than 100 provenance events.
↑

Reverting to the initial, quantitative phase of entity linking, it becomes apparent that 2,794 individuals (54.3%) yielded no match in either Wikidata or ULAN. The quantitative analysis additionally exposed the underrepresentation of female parties on both platforms. Specifically, 841 unmatched female parties constitute 30.1% of all such individuals, in contrast to female parties representing 23.1% of all individuals recorded by the Art Institute. Examining Wikidata, female parties represent 20.8% of recorded individuals, while in the case of ULAN, this percentage drops to 14.8% (Fig. 2).[16] This emphasizes the valuable role that institutions like the Art Institute can play in mitigating the systemic underrepresentation evident in repositories like Wikidata and ULAN. It is crucial to highlight that female parties whose identity is veiled within the married titles of couples were not included in the statistical count. If an institution like the Art Institute were to address and modify this recording practice, appropriately documenting the members of a couple individually, the potential impact on the representation of female parties would undoubtedly become even more substantial.

When considering a museum's perspective, it is crucial to acknowledge the recording priorities that an institution may have, particularly concerning biographical information. These priorities might be influenced by the frequency of an individual's appearance in recorded events. When evaluating the number of events associated with each individual, it is apparent that the 5,147 individuals display a long-tail distribution.

Among the 5,147 individuals, 3,848 (74.8%) were involved in a single recorded event, while 13 took part in over 100 events (0.3%). Notably, among the individuals registered in only one event, 925 are female parties, constituting 24%. In the prominent group of the top 13 individuals, only 2 are female parties.

Table 1 compares the 13 individuals engaged in more than 100 provenance events, indicating an elevated status for the Art Institute, as they participated in 11.6% of recorded events (4,118 out of 35,554 events). It is worth noting that, through manual verification, we can identify a potential candidate for "Reverend Chauncey Murch" in Wikidata. We can attribute the absence of an automatic match to the Wikidata entity lacking a name that is written in the exact same way. Furthermore, the birth year recorded on Wikidata is 1856, whereas the provenance texts document it as 1859. Given the historical importance of this individual in the Art Institute's collection, the institution is in a position of authority to enrich and potentially rectify information related to him.

Of the 13 most active individuals, nine are represented in Wikidata, and only three in ULAN. Four individuals are not represented in any of the repositories under analysis. This highlights that, despite the high representation of the most active individuals from Art Institute provenances in crowdsourced Wikidata, there is a notable relative absence of contributions from authoritative institutions to ULAN concerning provenance biographical data. In such instances, the museum's role as a data provider for its key parties comes to

[14] A comparable classification approach was also introduced in Goldfarb, Merkl, "Visualizing Art Historical Developments."

[15] This trend was also noted in a broader analysis of ULAN entities. See: *Ibid.*

[16] SPARQL queries were executed at the respective platform endpoints on February 6, 2024. Among the 11,049,161 instances of humans in Wikidata, 2,300,413 are associated with the female sex or gender. In comparison, of the 348,794 instances of humans in ULAN, 51,666 are associated with the female gender.

the fore. Such contribution not only streamlines the museum's data management processes, avoiding information redundancy and ambiguity, but also benefits other institutions. An individual highly involved in the provenance records of one museum might have also taken part in events at another museum and vice versa.

CONCLUSION

Exploring a provenance linked open data strategy applied to biographical information has illustrated the challenges and opportunities inherent in transforming museum provenance records within the digital environment. By examining the Art Institute of Chicago's provenance records, this paper has demonstrated how repositories such as Wikidata and the Getty's Union List of Artist Names embody two distinct approaches within the LOD ecosystem.

These contrasting visions — one generalist and open, the other more specialized and authoritative — complement each other and offer different types of support for museums on both quantitative and qualitative levels.

In this scenario, museums possess the wealth of information and the institutional authority to play a significant role as information providers on both platforms. However, this role necessitates an effort towards digitization, facilitated by computational methods, that not only enhances data accessibility and interoperability but also prompts a reevaluation of how art history conceptualizes key players in the art world.

This paradigm shift entails expanding the narrative beyond artists to encompass individuals involved in various facets of the art ecosystem: collectors, owners, and even those associated with illicit activities such as looting. By embracing this holistic perspective and leveraging digital tools for data enrichment and collaboration, museums can contribute to a more comprehensive understanding of cultural heritage and facilitate broader engagement across institutions in reconstructing the history of their collections.

.

BIBLIOGRAPHY / POPIS LITERATURE

Faraj, Ghazal; Micsik, András. "Persons, GLAM Institutes and Collections: an Analysis of Entity Linking Based on the COURAGE Registry." *International Journal of Metadata, Semantics and Ontologies* 15, 1 (2021), 39–49. https://doi.org/10.1504/ijmso.2021.117105.

Freire, Nuno; Manguinhas, Hugo; Isaac, Antoine. "An Observational Study of Equivalence Links in Cultural Heritage Linked Data for Agents." In: *Digital Libraries for Open Knowledge*, ed. Mark Hall, Tanja Mercun, Thomas Risse, Fabien Duchateau. Berlin, Heidelberg: Springer-Verlag, 2020: 62–70. https://dl.acm.org/doi/10.1007/978-3-030-54956-5_5.

Goldfarb, Doron; Merkl, Dieter. "Visualizing Art Historical Developments Using the Getty ULAN, Wikipedia and Wikidata." In: *2018 22nd International Conference Information Visualisation (IV)*. IEEE, 2018: 459–466. https://doi.org/10.1109/IV.2018.00086.

Harpring, Patricia. "Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA." *Art Documentation: Journal of the Art Libraries Society of North America* 29 (2010): 67–72. https://doi.org/10.1086/adx.29.1.27949541.

Mariani, Fabio; Rother, Lynn; Koss, Max. "Teaching Provenance to AI. An Annotation Scheme for Museum Data." In: *AI in Museums: Reflections, Perspectives and Applications,* ed. Sonja Thiel, Johannes Bernhardt. Bielefeld: transcript Verlag, 2023: 163–172. https://doi.org/10.14361/9783839467107-014.

Rother, Lynn; Koss, Max; Mariani, Fabio. "Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums." In: *Perspectives on Data,* ed. Emily Lew Fry and Erin Canning. Chicago: Art Institute of Chicago, 2022. https://doi.org/10.1177/0308518X15594899.

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

Schich, Maximilian; Huemer, Christian; Adamczyk, Piotr; Manovich, Lev; Liu, Yang-Yu. "Network Dimensions in the Getty Provenance Index." *arXiv* (2017). https://doi.org/10.48550/arXiv.1706.02804.

Sugimoto, Go. "Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data Integration." *Semantic Web* 14, 1 (2023): 55–100. https://doi.org/10.3233/SW-223026.

Vrandečić, Denny; Krötzsch, Markus. "Wikidata: A Free Collaborative Knowledgebase." *Association for Computing Machinery* 57, 10 (2014): 78–85. https://doi.org/10.1145/2629489.

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

ONLINE SOURCES / MREŽNI IZVORI

Wikidata. https://www.wikidata.org/ (date of access February 5, 2024).

Union List of Artist Names (ULAN), Getty Research Institute. https://www.getty.edu/research/tools/vocabularies/ulan/ (date of access February 5, 2024).