Teaching Provenance to AI

An Annotation Scheme for Museum Data

Fabio Mariani, Lynn Rother, Max Koss[1]

With the advent of new digital tools, museums are being presented with ever-expanding possibilities not only to explore their role and function in society, but also to deliver transparency and accountability regarding the origins of their collections. These origins can, in turn, be traced through provenances, which typically record the chains of events of ownership and socioeconomic custody changes of an object (fig. 1). And it is provenance records in museums that are particularly well suited to the application of computational methods such as artificial intelligence.

Figure 1: Provenance text for Paul Cézanne's Houses in Provence: The Riaux Valley near L'Estaque. *Source: National Gallery of Art website (https://www.nga.gov/collection/art-object-page.54129.html, accessed in August 2023).*

Probably acquired through (Ambroise Vollard [1867-1939], Paris) by Egisto Fabbri [1866-1933], Florence, by 1920;[1] by whom sold c. 1928 to (Paul Rosenberg et Cie., Paris).[2] Marius de Zayas [1880-1961], and his wife Virginia Harrison, New York, by c. 1930; by inheritance to his wife; (Zayas sale, Parke-Bernet Galleries, New York, 14 October 1965, no. 92); Mr. Paul Mellon, Upperville, VA; gift 1973 to NGA.

[1]Published in article on Fabbri collection in *Daedalo*, 1920.

[2]See John Rewald, *The Paintings of Paul Cézanne: a Catalogue Raisonné*, New York, 1996, no. 438, regarding the dispersal of the Fabbri collection.

Over the past two and a half decades, investigating provenance has become a full-fledged field of mainly archival-based research, resulting in complex and nuanced texts that brim with historical detail. Provenance research has indeed produced large quantities of information about artworks—not least on how, when, and where people and institutions were involved in, for example, their commissioning, selling, or looting. The insights gained from this mass of information nonetheless remain quite limited. This is mainly because detailed object histories continue to be

1 The authors would like to thank Liza Weber for her insightful editing of this paper.

recorded in museum collection management systems in, primarily, free text fields, thus making them inaccessible to computational analysis.

Lifting the historical information out of its data siloes and transforming it into linked open data would be a game changer for provenance research, decolonization efforts, and restitution. Large-scale analysis across museum collections would enable claimants and other parties to intelligently search for and efficiently identify objects looted or expropriated in contexts of injustice, such as during National Socialism or periods of colonial rule. It would also make it possible for researchers across disciplines to engage in historical network analysis, generating insights that can, in turn, inform curatorial, collecting, or outreach decisions.

Purposeful structuring is key to asking scientifically relevant questions about large-scale datasets in the humanities. This structuring process must, in turn, be guided by the potential queries that researchers may want to pose. In the field of provenance studies, such questions may relate, for example, to the relative impact of collectors, dealers, museums, or militaries on the looting, philanthropic giving, or sale of objects across time and space; such studies may also be aimed at mapping interconnections and comparing trends and patterns. Queries may be even narrower and examine the role of specific individuals, organizations, and objects. Lastly, purposeful structuring facilitates queries that can also be related to vague, incomplete, uncertain, or even contradictory provenance information, whose mere identification can suggest avenues for further archival research.

In our paper, 'Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums' (Rother/Koss/Mariani 2022), we have proposed a conceptual framework for what data to transition into provenance linked open data (PLOD) and on what level of detail. Given its modular structure, the framework enables museum professionals to strategize provenance transformation and data production. Through the use of AI, we have shown how museums can make the process of automatically extracting knowledge from provenance texts speedy and efficient (Rother/Mariani/Koss 2023).

Key to the process of extracting knowledge from provenance texts is training AI models for specific tasks. As we will demonstrate, this necessitates designing and implementing an annotation scheme that applies specific categories to the various elements encountered in provenance texts, as well as their potential relationships to one another. As such, devising an annotation scheme is part of that first and fundamental step in transforming provenance texts into structured data: expert interpretation. With a provenance-specific annotation scheme, we introduce a set of categories to help museum professionals train a machine to operate much like a provenance expert: extracting knowledge from provenance texts based on expert-determined logic.

The Nature of Provenance Texts

The structuring and the publication of provenance as linked open data must build on the wealth of provenance information that institutions have gathered in recent decades. Indeed, given the large volume of provenance texts that have been compiled by museums, the most realistic and resource-efficient strategy involves extracting knowledge from them rather than creating structured data from scratch. In order to extract knowledge from pre-existing provenance texts, we must first understand past and present practices for writing provenance texts so as to identify the most appropriate computational techniques for extraction.

To guide museums in recording provenance, the American Alliance of Museums (AAM) and the International Foundation for Art Research (IFAR) have compiled guidelines on writing provenance texts (Yeide/Walsh/Akinsha 2001; IFAR 2023). These guidelines, with their allowances for variation, do not represent strict standards, nor do they anticipate machine readability. They do, however, introduce writing conventions that have found widespread adoption, especially in the English-speaking provenance world, for instance, organizing texts according to their chronology or using specific punctuation to convey meaning. We found this genre of provenance to be particularly suitable for automatic structuring.

According to the AAM and IFAR guidelines, the provenance of an object is presented in chronological order. Each period of ownership corresponds to a sentence in the provenance text. Each sentence is furthermore delimited by a specific punctuation mark, which brings a particular meaning to it. For example, if a sentence ends with a semicolon, we know that the change of ownership between the two parties was direct. In contrast, if a sentence ends with a period, we can infer that there was a gap in the ownership history. Indeed, a period indicates that we do not know what happened to the object at this juncture.

The first step in automatic knowledge extraction from provenance texts thus concerns separating individual sentences. The specific natural language processing (NLP) task that can help with this problem is sentence boundary disambiguation (or detection). Its purpose is to disambiguate the punctuation that ends a sentence from other uses, such as in an abbreviation. We can successfully address this task by training deep learning models, in other words, artificial intelligence models, to perform a task when given a set of output examples.

Thanks to the formulaic nature of provenances, once we have divided a provenance text into individual sentences, we have automatically dissected it according to its constitutive provenance events. But any resulting list of provenance events is insufficient for meaningful analysis, since the constitutive components of individual provenance events remain inaccessible. More granular structuring is thus needed in order to unlock the historical complexities contained in provenance texts.

We have identified span categorization as the most efficient NLP task for extracting the various components of provenance events (Rother/Mariani/Koss 2023). This is because span categorization identifies portions of text (or spans) belonging to specific, expert-determined categories (or tags). In addition, span categorization allows a portion of text to belong to more than one category. This enables us to categorize a portion of text as a specific event component and simultaneously assign to it other categories that can help convey additional information about it. It is, moreover, possible to identify different spans within portions of text already assigned to one or more categories (Finkel/Manning 2009). Indeed, given the density of the historical information found in each provenance event, this feature enables us to extract more detailed knowledge from individual event components. It also represents a necessary precondition for complex querying and large-scale analysis.

A deep learning model can successfully perform the task of span categorization. As defined above, this type of AI model learns from output examples annotated by experts. When training a deep learning model for span categorization, it is then necessary for an expert to first annotate provenance events by identifying the different portions of text and assigning appropriate categories to them. To address this challenge, we have developed a provenance-specific annotation scheme, that is, a set of categories with which to annotate provenance texts for span categorization. But developing such an annotation scheme first requires a preliminary analysis of how provenance texts function, from understanding which portions of text to categorize to choosing which categories to assign.

According to the AAM and IFAR guidelines, each provenance event may contain one or more of the following pieces of information: the owner of the object; any agent involved in the transfer; the method of transfer; the location; and the date. A provenance event may, however, also contain additional information concerning specific aspects of an event. Indeed, it is the heterogeneity of information that we encounter in provenance texts that informs our approach to developing the annotation scheme. For, such a scheme must be adaptable to each provenance text, regardless of its level of detail.

A Provenance-Specific Annotation Scheme

To help institutions structure their data and eventually transform their provenance texts into PLOD, we have designed the abovementioned framework, which conceptualizes the different types of information contained within provenance texts and their varying levels of detail in a modular structure. With respect to knowledge extraction from provenance events, this conceptual framework is implemented in practice in the provenance-specific annotation scheme. Both our framework and scheme have flexibility in modelling provenance information, particularly when it comes to combining semantic layers and thereby translating historical complexities into data.

The conceptual framework introduces a base layer of information to describe the fundamental elements of any given provenance, starting with its backbone, the individual provenance event. Each provenance event is, in turn, composed of and associated with: the parties involved, the transfer taking place, as well as its location and time of occurrence. Based on these four elements, we have devised four fundamental categories for the provenance-specific annotation scheme: 'party', 'method', 'location', and 'time'.

The first step in training a deep learning model involves annotating all identified participants in a provenance event with the category 'party'. Importantly, the 'party' portion of a text concerns not only the entity's name but also any additional biographical information that we may find in the text, such as dates of birth and death or places of residence. Two or more parties acting together should be regarded as a group and annotated as a single 'party' span, though the individual parties within a single span should also be annotated with the 'party' tag. This enables us to maintain both the group's collective identity and the unique identities of its members, thereby allowing us to analyse the group's collective actions as well as the actions of individuals. This does not apply, however, to groups where members' names are missing, such as in the case of married couples, where it is often impossible to tag female owners due to outdated and exclusionary recording conventions. In this case, we would annotate 'Mr. and Mrs. John Doe' as a single span in the 'party' category.

With the category 'method', we are able to annotate transfers that occurred in a provenance event, which are usually identified by verbs and expressions indicating a change in ownership or socioeconomic custody (for instance, 'purchased', 'by inheritance'). The category 'location' enables us to annotate geographical locations in the text. Such portions of text do not always stand alone, but may also be found within another span, such as 'party', in which case the location is associated with the party, for example, the person's place of birth. The last of the four fundamental categories, 'time', applies to all temporal indicators in the text. Portions of text categorized as 'time' may be present again within a 'party' span, for instance, the person's date of birth.

Since researchers are producing ever-more provenance information, the PLOD conceptual framework proposes four types of descriptive bricks, so to speak, from which to build a set of relevant facts that have not already been recorded in the base layer. These bricks concern biographical, geographical, economic, and contextual information. Such information can also be taken into account when annotating categories.

The biographical brick provides further information about parties, which we can, for example, extract from any span categorized as 'party'. For instance, with the categories of 'person' and 'group', we can differentiate between an individual and a group of individuals, such as a couple, family, or organization. These categories may, of course, overlap, and thus help us to distinguish, as already mentioned, individual behaviours from group actions, should they be of concern to researchers or claimants.

In extracting knowledge from a provenance event, we must furthermore identify the role of each party, so as to: 1) represent the chain of ownership accurately and 2) make perfectly clear who did what in a given transaction. To achieve this, we apply the categories of 'sender', 'receiver', and 'agent'. Here again, the possibility of layering various tags proves to be crucial in being faithful to historical complexities. With the 'sender' category, we can annotate parties that parted, voluntarily or involuntarily, with their objects, while with the 'receiver' category, we can annotate parties that obtain objects, whether ethically, legally, or not. Finally, with the 'agent' category, we can annotate parties that act as intermediaries in events, such as auction houses.

Having recognized that women are not only misrepresented in provenances but are often even ignored altogether, we have concluded that a provenance-specific annotation scheme should also be a tool for identifying, measuring, and rectifying biases. We have therefore introduced a gender classification task. Due to the limitations of historical recording conventions linked to the gender binary and the fact that women were often specifically identified through married titles and maiden names, we have introduced only one category: 'female party'. This category can be assigned to any party whose name suggests specifically this. The annotation of such a category assists not only in identifying any gender biases in the text, but also finally amending them. For example, a party represented as 'Mrs. John Doe' may be annotated as 'female party', even though no party name technically exists.

As indicated, span categorization makes it possible for multiple spans to be layered on top of one another, thus providing more complex information about individual provenance event components. Within a 'party' span, for example, we can annotate the portion of text that coincides with the party's name with the category 'name'. As previously discussed, spans categorized as 'party' can also include biographical information such as date of birth and death, which we can correspondingly annotate with the 'birth' and 'death' categories. In turn, both the 'birth' and 'death' spans can include text portions belonging to the categories of 'time' and 'location' (for instance, the date and place of birth). Finally, with the 'description' category, we can annotate portions of text within the 'party' span that describe the family or professional role of the party. A 'description' of a party can be, for example, the text portion 'his daughter', thus describing a relationship with the previous owner, who, in this case, is a daughter receiving an object from her father.

The geographical brick expands on location information in the base layer of provenance. When a location appears in a provenance text with its geographical hierarchy, for example, 'Upperville, VA', it is crucial to accurately portray that 'Upperville' is a location within the location 'Virginia'. Combining spans enables us to do this without introducing additional categories. We can assign the category 'location' to the entire span of 'Upperville, VA', but also to the span 'VA'. This makes it possible for us to unambiguously identify Upperville as the unincorporated town of that name in Virginia and to analyse all provenance events that have occurred in the state of Virginia.

Provenance events represent economic activities, such as buying, selling, or auctioning objects. In our conceptual framework, any additional information concerning these activities, such as identificatory numbers or specific monetary values, is part of the economic brick. For span categorization, we have devised the categories 'inventory' and 'money' in order to extract such information from provenance texts. With the category 'inventory', we can annotate the various inventory numbers assigned to an object during its long history, whether they were assigned by a collector, an institution, or an auction house (for instance, a lot number).

Extracting additional economic information is crucial for large-scale provenance data analysis, which, to return to our introduction, is one of the ultimate goals of transforming provenances into PLOD. With an inventory number alone, for instance, it is possible, based on the archival records, to distinguish between two untitled paintings by the same artist that were sold in the same auction, as well as to identify who purchased each piece. The outcome of such archival research could include determining the buyer's price. Indeed, as provenance research gathers momentum and produces ever more detailed information on the fate of artworks, provenances increasingly include the prices paid by buyers and insurance evaluations from export papers. In order to annotate such monetary amounts, we have thus introduced the category 'money'.

The contextual brick is the fourth and final descriptive brick in the PLOD conceptual framework. Provenance texts can describe the larger historical contexts in which individual provenance events occurred. With the category 'context', we can annotate portions of text describing the historical context in which an event occurred. This means we can trace objects associated with the same historical contexts in subsequent analysis. For example, we could track all the objects sold in a given auction by extracting the auction title as 'context'. Similarly, we might trace all objects linked to the 'context' of the 'British military occupation of Benin', to name but one example where providing context through annotation may prove useful for questions of restitution.

Finally, the PLOD conceptual framework introduces four interpretive tools to help address the interpretative challenges that researchers face when structuring provenance data: vagueness, incompleteness, subjectivity, and uncertainty. Span categorization makes it possible to categorize all four challenges. Take, for example, the span 'circa 1945'. We can assign it both the 'time' and 'vagueness' categories, given that it is only an approximate period of time. In cases where information is incomplete or even missing entirely, we can annotate expressions of missing information by assigning the category of 'incompleteness' (for instance, to the span 'unknown owner', we can assign the categories 'party', 'name', and 'incompleteness'). Subjectivity may refer to the presence of two (or more) contradictory hypotheses about historical facts in a given provenance. For example, we can annotate the span '1935 or 1937' by assigning the tag 'subjectivity' and individually categorizing both '1935' and '1937' as 'time'. Lastly, historical hypotheses in provenance texts are often met with uncertainty, which is characterized by expressions such as 'possibly' and 'probably'. These terms can indicate different degrees of confidence when formulating a hypothesis about the occurrence of a provenance event. And we can annotate them with the category 'uncertainty'.

Figure 2: Conceptual example of span categorization applied to a provenance event extracted from the provenance text of Paul Cézanne's Houses in Provence: The Riaux Valley near L'Estaque. *Source: National Gallery of Art (https://www.nga.gov/collection/art-object-page.54129.html, accessed in August 2023).*

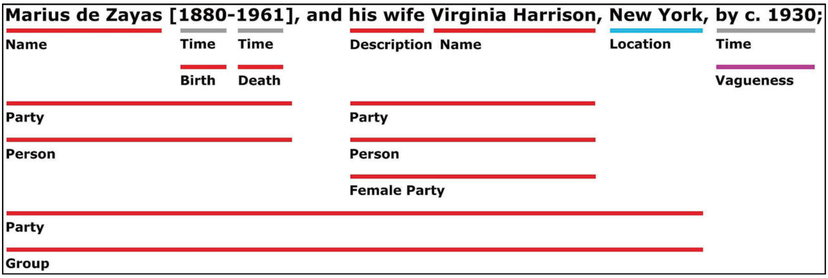

Figure 2 shows a conceptual example of annotation for span categorization that was applied to an event extracted from the provenance text for Paul Cézanne's painting *Houses in Provence*. At first glance, it is clear how the information in the text corresponds, for the most part, to the biographical brick in the PLOD conceptual framework. In fact, from the perspective of the base layer, we have a party containing, in turn, two parties, as well as the time of the event. Moreover, the time of the event is vague; based on the span 'by c. 1930', we know that the event occurred before 1930 or circa 1930. For this reason, we also categorize the portion of text 'by c. 1930' with the 'vagueness' tag. As for the parties involved, we annotated the individual persons according to single 'party' spans, to which we also added the tag 'group'. In addition to the two parties identified in the event, we annotated the 'location' span, here 'New York', since it is the location of the whole group.

The group's first party corresponds to the span 'Marius de Zayas [1880–1961]'. To this span, we can assign the categories 'party' and 'person'. We can also annotate additional information within the span. First comes the 'name', which corresponds to the 'Marius de Zayas' portion of the text. Then comes the individual's life span: 'birth' and 'time' ('1880') and 'death' and 'time' ('1961'). We can then annotate the span 'his wife Virginia Harrison' with the tag 'party' as the second group member. Here again, we can assign the category 'person', since she is also an individual. From the context and name, we can also assume the span concerns a 'female party' and annotate it as such. Moreover, within the span, we can tag additional information: from the 'name' of the party, 'Virginia Harrison', to the description 'his wife'.

Conclusion

Museums write provenance texts following similar principles. In light of this, we have developed a provenance-specific annotation scheme that can be adopted for similarly written provenances across institutions. Moreover, our scheme, based on the PLOD conceptual model, is intended to cover both the diverse content found in provenance texts and its varying levels of detail. AI is able to not only identify the main components of a provenance event (that is, its base layer), but also to recognize more complex and specific layers of additional information (that is, the bricks and interpretive tools). By annotating provenance texts with our scheme, we can address the NLP task of span categorization. This annotation process, which is ultimately undertaken by experts, aims to train AI to automatically replicate the same work performed by humans and follow the same logic, albeit on a much larger scale.

References

Finkel, Jenny Rose/Manning, Christopher D. (2009). Nested Named Entity Recognition. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, Association for Computational Linguistics, 141–50. https://doi.org/10.3115/1699510.1699529 (all URLs here accessed in August 2023).

IFAR (2023). International Foundation for Art Research (IFAR) Provenance Guide. Available online at https://www.ifar.org/Provenance_Guide.pdf.

Rother, Lynn/Koss, Max/Mariani, Fabio (2022). Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums. In: Emily Lew Fry/Erin Canning (Eds.). Perspectives on Data. Chicago, The Art Institute of Chicago. https://doi.org/10.53269/9780865593152/06.

Rother, Lynn/Mariani, Fabio/Koss, Max (2023). Hidden Value: Provenance as a Source for Economic and Social History. Economic History Yearbook, Special Issue: Digital History 64 (1), 111–42. https://doi.org/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.