Vague, Incomplete, Subjective, and Uncertain Information in Digital History

By the School of Culture and Society of Leuphana University
Lüneburg for the award of the degree of

Doctor of Philosophy

Dr. phil.

Approved Dissertation by Fabio Mariani
born on January 19, 1993 in Ascoli Piceno (Italy)

Submitted on: July 7, 2025
Oral defence (disputation) on: October 16, 2025
Year of publication: 2026
First supervisor: Prof. Dr. Lynn Rother, Leuphana University Lüneburg
Second supervisor: Prof. Dr. Ricardo Usbeck, Leuphana University Lüneburg
External reviewer: Dr. Marilena Daquino, University of Bologna, Italy

The cumulative dissertation comprises the following published contributions:

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

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

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

- Mariani, Fabio, Max Koss, and Lynn Rother. 2024. "People Information in Prove- nance Data: Biographical Entity Linking with Wikidata and ULAN." *Život umjetnosti*, no. 114, 148–161. https://doi.org/10.31664/zu.2024.114.07.

- Mariani, Fabio. 2023. "Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data." In *Proceedings of the Workshop on Computational Methods in the Humanities 2022. Lausanne, Switzerland.* https://ceur-ws.org/Vol-3602/paper5.pdf.

- Rother, Lynn, Fabio Mariani, and Max Koss. 2024. "Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI." In *Sammlungsforschung im digitalen Zeitalter: Chancen, Herausforderungen und Grenzen*, edited by Katharina Günther and Stefan Alschner, 93–103. Göttingen: Wallstein. https://doi.org/10.1515/jbwg-2023-0005.

- Mariani, Fabio. 2025. "PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data." *Zeitschrift für digitale Geisteswissenschaften*, no. 10. https://doi.org/10.17175/2025_012.

Use of Third-Party Software

In the course of writing this thesis, I made use of the following third-party tools to support the linguistic quality and clarity of the text:

- **DeepL** – to assist in expressing ideas more clearly in English, as a non-native speaker.
- **Grammarly** – for grammar correction and style refinement.
- **ChatGPT-4o** – for language polishing and proofreading support.

The use of these tools was strictly limited to improving the thesis's language and presentation. Their application is consistent with the methodological commitments of this research, which emphasise the use of automation as a means of supporting human authorship, not replacing it.

Abstract

This cumulative dissertation investigates how digital infrastructures can accommodate the epistemic complexity of historical knowledge, focusing on the specific challenges posed by vague, incomplete, subjective, and uncertain (VISU) information. Using art provenance as a focused domain of inquiry, the research addresses how such information can be identified, structured, and preserved in digital formats without flattening its interpretative depth.

The central hypothesis underpinning this work is that VISU information can be effectively preserved and meaningfully operationalised in digital history only through workflows that integrate computational methods with interpretive oversight. Specifically, this requires combining automated extraction, adequate data modelling, and expert validation to ensure that the epistemic complexity of historical knowledge is not lost in the process of digitisation.

This guiding hypothesis gives rise to three interrelated research questions. First, how can automated extraction processes be implemented to identify and flag VISU information for preservation in historical texts? Second, how can VISU information be formally represented in structured data models without sacrificing interpretive complexity? Third, how can expert validation be operationalised to safeguard VISU information during the transformation of historical data?

To answer the first research question, the thesis identifies key natural language processing tasks such as sentence boundary detection and span categorisation, and develops a tailored annotation scheme to capture VISU features. This scheme is used in training and evaluation of models for automatic extraction of structured knowledge from provenance records. Addressing the second question, the research explores modelling strategies that extend CIDOC CRM by aligning CRMinf with the Historical Context Ontology (HiCO) and structuring data using nanopublications. This approach supports the formal representation of historical claims, interpretative assertions, and metadata from the digitisation process. Finally, in response to the third question, the thesis introduces PROV-A, a web based tool that integrates automated extraction with expert validation. It allows historians to refine extracted data, annotate epistemic qualifiers, and publish structured provenance information as linked open data, preserving interpretative depth within scalable workflows.

These contributions are presented across seven publications that form the basis of this cumulative dissertation. Together, they establish a methodological approach for preserving the historiographical richness of provenance records as they are transformed into digital formats.

Contents

List of Figures

List of Tables

Listings

1. Introduction

1.1 Motivation

Historical knowledge is shaped not only by what is known, but by the conditions under which it is known—and just as critically, by what remains uncertain, contested, or only partially discernible. Across the discipline, historians contend with information that is *vague*, *incomplete*, *subjective*, or *uncertain*. These features are not anomalies to be corrected, but epistemic conditions of working with the fragmentary, mediated, and interpretive nature of the past. In this doctoral research, these conditions are collectively conceptualised under the acronym VISU. Inspired by the Latin *de visu* ("with your own eyes"), the VISU framework foregrounds the epistemological demands of engaging with such forms of historical information, underscoring the need for expert interpretation when confronting ambiguity and absence.

Historiography has long developed methods for engaging with the ambiguity, fragmentation, and interpretive complexity of historical sources, using techniques such as critical source analysis, contextual inference, and narrative synthesis. As historical records are increasingly digitised, structured, and subjected to computational analysis, these established practices encounter new methodological tensions. The promise of precision, reproducibility, and scale offered by digital methods introduces what Rieder and Röhle (2012) describe as the "lure of objectivity": a tendency to privilege formal clarity and computational legibility at the expense of epistemic nuance. This dynamic risks reducing historically situated interpretations to ostensibly neutral data representations. Zundert (2015) calls for a hermeneutics that not only persists alongside digital methods but critically engages with the implicit interpretive frameworks embedded in algorithms and models. He argues that digital tools carry their own hermeneutics, which humanities scholars must confront through closer dialogue with computer science. Such an approach acknowledges that interpretation begins not after, but within, the design and application of computational methods. Ter Braake et al. (2016) describe this as "tool criticism", a critical stance that evaluates digital tools in light of the epistemological commitments of historical scholarship, ensuring that digital history retains its interpretive and reflexive core.

It is within this tension—between the epistemic character of historical knowledge and the formalising imperatives of digital infrastructures—that this research is situated. My inquiry is grounded in the domain of provenance, where historical reconstruction intersects with institutional responsibilities such as transparency, restitution, and accountability. In the GLAM context (galleries, libraries, archives,

and museums), provenance refers to the documentation of an object's ownership and custody history, often reconstructed from heterogeneous archival traces and material evidence. It offers a particularly rich terrain for examining how ambiguity, incompleteness, and interpretative plurality can be critically structured within computational systems. This perspective took shape during my work as an associate researcher at the *Provenance Lab*, where I contributed to the inaugural project *Modern Migrants: Paintings from Europe in US Museums*. The project mapped how European paintings entered US museums from 1860 to 1945, shaped by historical events, art market trends, and collecting policies. Within this context, provenance records frequently exhibited forms of epistemic complexity, ranging from vague dates and uncertain attributions to transactional gaps and inconsistencies in the identification of actors. Far from being incidental, these features reflect the archival, political, and institutional conditions under which provenance information was historically produced and transmitted.

As museums and cultural heritage institutions increasingly seek to digitise such records—in response to public expectations of transparency, accountability, and restitution—these epistemic complexities pose methodological challenges. Converting provenance information into structured, machine-readable formats raises critical concerns about how historical knowledge is affected: what aspects are faithfully preserved, which ones are modified in translation, and what elements risk being omitted altogether. Ambiguities that may be acknowledged or even foregrounded in narrative form can become difficult to represent within formal ontologies and data models. The risk is that epistemic nuance may be reduced in favour of interoperability and computational legibility. This tension is particularly significant in contexts where provenance is not merely descriptive but implicated in ethical, legal, or political claims. Against this backdrop, the question arises as to how digital methods might be designed or adapted to accommodate the interpretive character of provenance data, rather than presuppose its resolution.

As the digitisation of historical records accelerates, cultural heritage institutions face increasing pressure to process large volumes of legacy documentation. Manual transcription and structuring are time-consuming and resource-intensive, prompting interest in automated solutions to address the challenge of scale. Yet despite a wider turn toward natural language processing (NLP) in historical research, no studies have systematically examined the distinctive linguistic and semantic characteristics of provenance records. These texts are shaped by domain-specific conventions that diverge significantly from those found in commonly analysed historical corpora. As a result, there is limited understanding of how automated extraction workflows can be adapted to preserve the interpretive nuances that are central to provenance research.

Within NLP, tasks such as Named Entity Recognition (NER) are well established

for automatically identifying and classifying mentions of people, places, dates, and other key entities in text, and it has demonstrated robust performance across a wide range of historical and contemporary sources (Ehrmann et al. 2023). Yet these approaches are oriented toward extracting unambiguous factual entities and offer limited means of representing the epistemic nuances found in historical texts. As a result, research on vagueness and uncertainty in historical writing has tended to focus on the automatic identification of discrete linguistic cues such as vague quantifiers, modal adverbs, and epistemic verbs, treating these cues as a separate processing task (Vertan 2019). Provenance records, however, present methodological demands that these approaches are not designed to address. The distinctive writing conventions of provenance records, which are non-standard yet highly codified, yield telegraphic and elliptical expressions that fall outside the assumptions of NLP methods developed for fluent, grammatically coherent text exhibiting regular syntactic and lexical patterns. Nonetheless, this same structure opens possibilities for domain-specific models that can extract entities and semantic relations, as well as the epistemic nuances that shape provenance knowledge.

The composition of provenance records adds an additional layer of complexity for NLP. Much of the intellectual work undertaken by their authors is expressed in accompanying notes, where clarifications, bibliographic references, tentative identifications, and even conflicting hypotheses are recorded. These notes function as a space for reasoning rather than for conveying straightforward factual statements, and their varied and discursive form resists the uniformity required by automated extraction methods. For this reason, the study undertaken in this dissertation considers not only automatic extraction techniques but also strategies for embedding human interpretation within the modelling process.

These limitations in automatic text processing extend to the modelling layer, where provenance information must ultimately be expressed within structured semantic frameworks. Within the cultural heritage domain, the primary reference ontology is the CIDOC Conceptual Reference Model (CIDOC-CRM), an ISO standard (ISO 21127) developed by the International Documentation Committee (CIDOC) of the International Council of Museums (ICOM). CIDOC-CRM offers a formal, event-centric ontology for representing cultural heritage information, structuring relationships among entities such as objects, people, places, and activities. Its event-based design is particularly well suited to provenance modelling, as it allows for the reconstruction of an object's historical trajectory through sequences of documented events. However, while CIDOC-CRM excels in modelling factual relations and ensuring semantic interoperability, it lacks elements to express interpretive uncertainty, contested claims, or the epistemic status of historical assertions. Its conceptual architecture presupposes a level of completeness and coherence that does not align

with the fragmentary and interpretive nature of many provenance records. Application profiles such as Linked Art adapt implementation within institutional contexts, but inherit these structural constraints. This dissertation critically examines how CIDOC-CRM and its application profiles could be adapted to represent provenance in a manner that retains both semantic rigour and historiographical complexity.

These methodological challenges are central to the motivation for this dissertation. Tackling the scale of provenance digitisation requires more than technical efficiency; it calls for methods that respect the complexity of historical knowledge. If digital systems are to handle provenance data without stripping away its interpretive depth, then new approaches are needed across the entire workflow: information extraction, structured modelling, and expert review. The next section outlines the hypothesis and research questions that shape this investigation.

1.2 Hypothesis and Research Questions

This section delineates the central hypothesis and research questions that guide the investigation into how digital methods can accommodate the epistemic complexity of historical knowledge. The aim is to explore how VISU information can be retained and made computationally meaningful across the digital workflow, encompassing automated extraction, formal modelling, and expert validation.

> **Hypothesis:** VISU information can be effectively preserved and meaningfully operationalised in digital history only through workflows that integrate computational methods with interpretive oversight. Specifically, this requires combining automated extraction, semantically expressive data modelling, and expert validation to ensure that the epistemic complexity of historical knowledge is not lost in the process of digitisation.

This hypothesis gives rise to three interdependent research questions. Each addresses a key phase in the digital handling of historical data, with a focus on how VISU information can be identified, structured, and critically maintained without flattening its interpretive nuance.

RQ1: How can automated extraction processes be implemented to identify and flag VISU information for preservation in historical texts?

This question explores the computational preconditions for working with VISU information. While automated extraction cannot resolve the interpretative complexity of VISU data, it plays a critical preparatory role: identifying and capturing linguistic markers such as “probably”, “circa”, or “unknown” that signal ambiguity or incompleteness. These markers are essential for flagging interpretative challenges in

subsequent processing stages. Particular attention is given to person-related data, where ambiguity in names, roles, and identities demands cautious extraction and careful disambiguation across sources.

RQ2: How can VISU information be formally represented in structured data models without sacrificing interpretive complexity?

This question focuses on the formal representation of VISU information in machine-readable formats. It explores the adequacy of existing ontological standards and the potential of extending or adapting them to express epistemic uncertainty, alternative interpretations, and evidentiary gaps. Particular attention is given to how the production, mediation, and reinterpretation of historical information can be modelled to reflect the historiographical nature of the data.

RQ3: How can expert validation be operationalised to safeguard VISU information during the transformation of historical data?

This final question turns to the role of human supervision in the digital workflow. It investigates how historians and domain experts can be supported in supervising and refining machine-generated data, particularly where automated methods fall short in recognising or contextualising VISU information. The emphasis lies on developing human-in-the-loop systems and user interfaces that allow for granular correction, interpretation, and scholarly curation, ensuring that expert judgement remains central to the encoding of historical complexity.

Together, these research questions articulate a methodology for working with VISU information in digital history that balances automation with interpretation. The proposed framework advances a critical approach to computational historiography, by preserving the situated, uncertain, and contested nature of historical knowledge within formalised digital environments.

1.3 Contributions and Publications

The dissertation contributes to the development of computational historiography by articulating and operationalising VISU information, that is, vague, incomplete, subjective, and uncertain elements that characterise historical knowledge. Before addressing the three central research questions, the project establishes a conceptual and institutional foundation for the study of provenance as a case study through which VISU information is examined in digital contexts. This groundwork situates the research within broader debates around data ethics, curatorial responsibility,

and the epistemic status of historical records, providing both a justification for the VISU framework and a critical orientation for subsequent technical developments. Building on this foundation, the dissertation then addresses its core research questions, engaging with the methodological and representational challenges posed by VISU data: its automatic extraction from text, its formal modelling within structured ontologies, and its validation through expert oversight and human-in-the-loop approaches. These components, while distinct in focus, are mutually reinforcing. Together, they delineate a coherent methodological trajectory from theoretical framing to practical implementation, aimed at preserving historical complexity within digital infrastructures.

Foundational Work and Conceptual Orientation

A foundational contribution to this dissertation is articulated in the article **"Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums"** (Rother, Koss, and Mariani 2022), which lays the theoretical and strategic groundwork for the research agenda pursued throughout the doctoral research. Rather than addressing a single research question, this piece situates provenance digitisation within broader curatorial, political, and epistemological debates, thereby providing the necessary backdrop against which the VISU framework is developed.

The article introduces the Provenance Linked Open Data (PLOD) Conceptual Framework, a modular and layered strategy designed to support museums in transitioning from analogue records to interoperable, ethically attuned digital formats. My specific contribution focused on the methodological design and visual modelling of this framework. It responds to four primary challenges in provenance digitisation: inconsistency in legacy records, the prevalence of incomplete or uncertain information, resource constraints, and the growing demand for institutional accountability in contexts such as restitution and decolonisation.

Rather than proposing a universal solution, the framework advocates for strategies tailored to individual institutions, while remaining compatible with broader linked data initiatives. Central to this approach is the commitment to preserving, rather than neutralising, the ambiguity and partiality intrinsic to historical records. This methodological stance forms the foundation of the dissertation's normative orientation, clarifying the ethical and epistemic stakes of digital provenance and motivating the focus on the complex nature of VISU information in digital history.

Contributions to RQ1: How can automated extraction processes be implemented to identify and flag VISU information for preservation in historical texts?

The first set of contributions focuses on developing methods to recognise and flag VISU information in provenance texts shaped by institutional recording practices. To address this, the research identified span categorisation as a suitable NLP task. Unlike traditional entity recognition, span categorisation allows for overlapping and multi-labelled segments, making it especially well-suited to capturing the layered nature of VISU elements.

Building on this insight, I developed an annotation scheme tailored to the specific linguistic and epistemic features of provenance narratives, and grounded in the PLOD Conceptual Framework. The scheme is presented in the article **"Teaching Provenance to AI: An Annotation Scheme for Museum Data"** (Mariani, Rother, and Koss 2023b), where it formalises an approach for systematically encoding epistemically significant elements, such as vague temporal markers, uncertain attributions, and omitted intermediaries, within textual corpora. This work establishes a foundation for training models capable of recognising and retaining the ambiguity and complexity inherent in historical data. It thereby advances a hybrid workflow in which automated systems support, but do not replace, expert human interpretation.

Building on this foundation, a second contribution emerged from an interdisciplinary experiment that evaluated the viability of automated extraction from provenance records for the purposes of socio-economic and network analysis, as detailed in **"Hidden Value: Provenance as a Source for Economic and Social History"** (Rother, Mariani, and Koss 2023). In this project, I was responsible for curating and annotating the training corpus, implementing NLP models, and assessing extraction performance. Before span categorisation, provenance notes were excluded. These notes, which often provide valuable contextual information such as source references, interpretive comments, and indicators of uncertainty, were structurally inconsistent and incompatible with automated extraction. Their exclusion, though methodologically necessary, highlights a fundamental limitation of automated approaches and reinforces the critical role of expert oversight in capturing the full epistemic complexity of historical data.

Consequently, the study focused on dimensions of VISU information that were more reliably retrievable at scale, particularly vagueness and incompleteness. Despite these constraints, trained models successfully extracted structured event data from heterogeneous textual sources, demonstrating that certain forms of epistemic ambiguity can be computationally flagged without being flattened. While the approach remains tailored to the specificities of museum provenance, it points toward

the potential for adapting similar techniques to other historical contexts.

A final contribution in this area centres on the extraction and disambiguation of biographical entities, the most frequently occurring and semantically dense components of provenance records. The article **"People Information in Provenance Data: Biographical Entity Linking with Wikidata and ULAN"** (Mariani, Koss, and Rother 2024) outlines the development of a pipeline for linking extracted person names to external authority files, specifically Wikidata and the Getty Union List of Artist Names (ULAN). This task posed significant interpretive challenges: individuals often appear under inconsistent or institution-specific naming conventions, and compound identifiers such as "Mr. and Mrs." obscure personal identity, particularly in terms of gender representation. In doing so, the project opened new avenues for critical historical analysis, particularly in areas such as gender representation and the sociology of collecting, where VISU elements are not incidental anomalies but fundamental aspects of how the archival record is shaped and preserved.

Contributions to RQ2: How can VISU information be formally represented in structured data models without sacrificing interpretive complexity?

A key contribution of this dissertation lies in developing a data modelling strategy capable of accommodating the epistemic complexity of provenance records within structured digital environments. This challenge is addressed in the article **"Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data"** (Mariani 2023), which introduces the VISU framework as both a conceptual vocabulary and a practical design principle for modelling cultural heritage data.

The article begins by identifying common forms of interpretive ambiguity in provenance records, such as vague dates ("circa 1945"), subjective attributions ("possibly sold to…"), and contestable claims. It then proposes ways to record these ambiguities within structured datasets. A central concern is the representation of incompleteness, not merely as missing values, but as historically meaningful absences shaped by archival silences and gaps in the record. To address this, the article introduces querying and documentation strategies that allow such absences to be surfaced and interpreted, even if they cannot be formally encoded as data entities.

The representational work draws on CIDOC-CRM as a semantic foundation and evaluates Linked Art as its application profile. While CIDOC-CRM supports fine-grained modelling of historical entities and events, it does not natively accommodate conflicting interpretations, uncertainty, or justification metadata. Linked Art, meanwhile, provides an implementable structure aligned with museum prac-

tices, but lacks mechanisms to express the interpretive and evidentiary dimensions of historical data.

To address these limitations, the article proposes an alignment between CIDOC-CRM and the Historical Context Ontology (HiCO), which is specifically designed to model interpretive acts and their attribution. This alignment is extended through the use of CRMinf, a CIDOC-CRM module developed to represent reasoning and inference processes, enabling the documentation of how particular interpretations are reached and justified. The proposed approach culminates in the use of nanopublications: compact data structures that capture individual assertions alongside their evidentiary sources, authorship, and contextual metadata. Rather than reducing complex historical claims to single, authoritative statements, the model supports a digital historiography in which uncertainty and interpretive plurality remain visible, traceable, and open to revision.

Contributions to RQ3: How can expert validation be operationalised to safeguard VISU information during the transformation of historical data?

The third research question explores how human expertise can be integrated into computational workflows to preserve the epistemic integrity of historical data. As digital infrastructures scale, there is an increasing risk that machine learning systems will obscure or simplify the interpretive complexity embedded in provenance records, documents often shaped by fragmented sources, multi-authored perspectives, and institution-specific conventions. To mitigate this risk, the research advances a human-in-the-loop approach in which curators and historians iteratively engage with automated processes, not as external validators but as integral participants in the production of structured data.

The conceptual foundations of this approach are outlined in **"Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI"** (Rother, Mariani, and Koss 2024), which addresses the limitations of relying solely on automation in contexts marked by ambiguity and heterogeneity. The article argues for a collaborative workflow in which expert users refine, contextualise, and, where necessary, reject machine-generated outputs, preserving the layered interpretations that make provenance records historically meaningful.

This collaborative paradigm is implemented in **"PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data"** (Mariani 2025), which offers a practical interface for embedding expert judgement within automated workflows. PROV-A enables users to structure provenance information as Linked Open Data (LOD) while preserving the epistemic complexity associated with VISU characteristics. The tool is designed to interface with externally automated processed outputs, offering an interactive environment in which curators and researchers can

refine machine-generated statements, annotate features such as approximation and uncertainty, and associate claims with sources and interpretive commentary.

As a research output, PROV-A translates the conceptual findings of RQ2 into a working application that models provenance information as nanopublications. The interface supports structured querying of VISU dimensions, for example, enabling users to detect and interrogate gaps, approximations, or epistemic qualifiers within a dataset. The article presents a case study based on a subset of provenance records extracted in RQ1, illustrating how the tool completes the digitisation pipeline from automated extraction to structured publication through expert supervision. In doing so, PROV-A not only operationalises the methodological strategies developed in this dissertation, but also exemplifies how digital tools can expose, rather than flatten, the epistemic contours of historical knowledge, thereby enabling their critical engagement through computational means.

1.4 Research Outputs

This dissertation is complemented by research outputs that embody its theoretical and technical work. Developed in response to the challenges posed by VISU information, these tools and datasets support automated and semi-automated approaches to provenance research. Each resource is released under an open licence to promote critical engagement, reuse, and further development within digital scholarship.

- **NLP Models for Extracting Knowledge from Museum Provenance Texts**
  Natural Language Processing models implemented with *spaCy*, trained to perform sentence boundary detection and span categorisation on museum provenance texts. The models were trained using annotated data derived from the Art Institute of Chicago.
  DOI: https://doi.org/10.5281/zenodo.13987656

- **Structured Provenance Events Dataset from the Art Institute of Chicago**
  A dataset comprising 35,554 structured provenance events, automatically extracted from 11,392 records published by the Art Institute of Chicago. The data was structured using the NLP models developed in this project, without further manual refinement.
  DOI: https://doi.org/10.5281/zenodo.15655528

- **PROV-A: The Provenance App**
  A web-based application for the curation and publication of provenance as Linked Open Data. PROV-A is designed to integrate automated extraction

workflows with manual refinement, enabling the representation of VISU information within structured provenance data.
Access: https://prov-a.github.io
Codebase: https://github.com/prov-a/prov-a.github.io

- **Case Study: Department of Modern Art, Art Institute of Chicago**
  A dataset produced as a case study for PROV-A, comprising the provenance records of 235 paintings from the Department of Modern Art at the Art Institute of Chicago. The data is available both as PROV-A project files (in JSON format) and as RDF Nanopublications serialised in N-Quads.
  Assets: https://github.com/prov-a/prov-a.github.io/releases/AIC_CaseStudy

1.5 Editorial Notes

This cumulative dissertation consists of a series of full-length articles, reproduced here with only minor editorial revisions. Where appropriate, structural elements, layout, tables, and code listings have been standardized to ensure a consistent appearance across the document. The bibliography and citation style have also been unified throughout the dissertation, conforming to the guidelines of the Chicago Manual of Style. To maintain consistency, all citations have been converted to footnotes, including those originally presented as endnotes or as in-text citations.

Each article is prefaced by a brief statement outlining the doctoral author's specific contributions, following the CRediT (Contributor Roles Taxonomy) system to clarify roles in co-authored work.

While the main body of the dissertation is written in British English, some variations in spelling and usage occur in the included articles, reflecting their diverse publication venues. In particular, "Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums" (Rother, Koss, and Mariani 2022), published by the Art Institute of Chicago follows American English conventions. Additionally, Figure 2.3 has been reformatted from its original digital layout to suit the print format of the thesis, enhancing legibility and integration with the surrounding text.

2. Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums

Bibliographic Information
Rother, Lynn, Max Koss, and Fabio Mariani. 2022. "Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums." In *Perspectives on Data*, edited by Emily Lew Fry and Erin Canning. Chicago: The Art Institute of Chicago. https://doi.org/10.53269/9780865593152/06.
CRediT Roles: Methodology, Investigation, Writing - Original Draft, and Visualization

2.1 Introduction

The histories of artworks are linked. After a work is created, its path may be marked by stints in galleries, private homes, storage facilities, exhibitions, and museums. At each of these stops, its life intersects with the lives of other works. We can think of these intersections as events involving people buying, selling, inheriting, looting, or otherwise transferring objects in specific places, at specific moments in time. When artworks end up in museums, information about the object's journey is recorded in its provenance, but these intersections with the lives of other works usually are not. Moreover, until recently, these shared histories of museum objects were only made visible through scholarly studies or exhibitions focusing on themes such as the influence of a particular collector or art dealer.

With museums increasingly recording and publishing their collection information digitally, provenance data is calling out to be connected through the use of technology. Provenance records typically consist of information, with varying levels of detail, on individuals, organizations, locations, transfers, and time periods related to ownership and custody changes. Transforming such records into linked open data (LOD), a web standard that defines how to publish resources online, would allow for large-scale analysis of the patterns, trends, and networks of the circulation and dislocation of objects relevant to disciplines such as art history, economic history, and sociology. Moreover, an LOD approach to provenance would help museums make better use of limited resources: identifying objects with the same owners or similar paths within and across museum collections would produce synergistic effects in the

research of specific object histories within and across institutions. The distributed forms of knowledge production that an LOD-driven approach facilitates would also enable multiple stakeholders to share more efficiently the work of collecting and recording critical data on, for example, a specific art dealer or a historical event that influenced the circulation of many objects across many institutions.

As object records are a form of writing history, provenance data also offers museums an opportunity to fulfill their social responsibilities of transparency, accountability, and inclusivity in light of twenty-first-century efforts around restitution and decolonization. While the relationship between provenance records and social responsibility may seem insignificant, this is by no means the case. For instance, if museums applied LOD standards to provenance data, third parties could research, query, and link this data. Among other effects, this would provide an opportunity for the reparative extraction of data, especially for objects whose histories are of greatest concern. In our view, it is in the interest of these institutions' missions to facilitate such endeavors, as this would allow any researcher to learn not only about the objects but also about museums' entanglement with the aftermath of injustice, thereby contributing precisely the kinds of knowledge and perspectives that may be missing from mainstream museum data.

With Linked Art, moreover, the museum and cultural heritage community has developed a data model that explicitly addresses the needs of art museums in creating LOD, with a particular focus on provenance.[1] Additionally, the Getty Research Institute, with its Getty Provenance Index, has embarked on a multiyear project to transform information about art dealers and collectors compiled over more than thirty-five years into LOD.[2] With the imminent release of millions of auction and art dealer data, information from this unique repository can be linked to, and often enhance, provenance records of individual museum objects, providing new insights on both the micro and macro scales—particularly about the art market. These initiatives are, in turn, echoed in the fledgling field of digital provenance studies.[3] In light

1. Chaired by Robert Sanderson of Yale University and Emmanuelle Delmas-Glass of the Yale Center for British Art, the Linked Art community is a consortium of people working with cultural heritage data. The community is currently comprised of twenty-four institutions based primarily in North America and Europe. Lynn Rother has served on the editorial board of Linked Art since January 2019. See https://linked.art/. The authors want to thank the editorial team, the anonymous peer reviewer, and Duane Degler for generously sharing their experience and valuable feedback. Margaret Doyle from the National Gallery of Art has graciously helped with procuring a crucial illustration. Thanks also go to Amy R. Peltz for editing this essay and to Liza Weber for her editing.

2. See Davis (2019) and Schich et al. (2017).

3. Grana-Behrens (2021); Claassen et al. (2020); Huemer (2020); Lincoln and Ginhoven (2018); Cranston (2020); Luther (2020); Newbury and Lippincott (2019); Smith (2018); and Kuhnen et al. (2018).

of these recent developments, we believe that this is the moment for museums to conceptualize the process of transforming analog museum records or their digitized offspring into machine-readable, searchable, and linkable provenance data.[4]

As we see it, this process of transformation must take into consideration four interdependent challenges. The first concerns how museums deal with the legacy of diverse and heterogeneous provenance practices and information, which we address in the first section of this essay. Until recently, provenances were often recorded inconsistently both across and within museums, and so we face biases and divergent levels of detail in and across provenance records with regards to which events and parties are documented, and to what standards. As museums are being asked to take stock of their collections, the consistency and transparency of their provenance records is a pressing matter.

In the effort to make their records consistent and transparent, museums are faced with a second challenge: the absence of provenance information, given that much historical knowledge is still missing. Based on our own experience, we have identified four aspects of provenance records that result from gaps in historical knowledge and ultimately pose challenges for these records' representation in a digital format: incompleteness, vagueness, uncertainty, and subjectivity.[5] As museums consider moving from analog to digital records, they must base any complex data modeling on the data they actually have, not on the model they want.

Of course, transformation from analog to digital is labor intensive, which raises the question of how museums allocate their resources—the third challenge they face. Not only is provenance research itself demanding work, but so too is the process of mapping provenance records into linked data. Furthermore, funding to support basic research and infrastructure remains limited. In order to transform provenance records into linked open data, museums thus face a resource-intensive process that requires not only commitment but also specialist knowledge from data and provenance experts—at least while data literacy remains beyond the standard skillset of humanities scholars.[6]

4. Machine readability refers to information's ability to be read and analyzed by a computer automatically. Non-digital material, such as printouts or handwritten letters, are not machine-readable. But digital material, such as a JPG image file showing text, can also be non-machine-readable. To the computer, such digital material constitutes an image that it cannot automatically read and process as a text. If the text is stored differently, e.g. in word-processing software, it is machine readable. While all machine-readable texts have some structure, structured data refers to data where the relation between textual elements is explicit in the way the data is stored. This means that the logic that is embedded in texts and understable by humans, is made explicit to the computer by creating a structured and machine-readable representation, e.g., a table, of the text's logic.

5. Mariani (2022a).

6. On data literacy, see Klinke (2020).

Finally, if we zoom out and consider museums as part of wider society, we see that they constantly face external demands. At the moment, for example, calls for restitution and the decolonization of museums require both accountability and action. As museums reconsider their collections in light of these evolving exigencies, they must consider where provenance sits within their wider institutional goals, missions, and priorities.

In light of these challenges, it is clear that if museums are to remain relevant to contemporary audiences, they must engage critically and self-reflexively with their own collections and, ultimately, the provenances of these collections. In so doing, they must position themselves strategically and also set priorities to determine how to distribute their already limited resources. As the transformation of provenance records from analog into digital formats becomes a priority for museums, a crucial question arises: Which data should be included, and which should be excluded?

In this essay, we lay out a conceptual framework that may help museums make informed decisions when they create machine-readable data from existing provenance texts, or when they begin to build provenance data from scratch. The framework offers a limited, resource-conscious intervention tailored to our present moment, which is a transitional one toward a more fully digitally engaged museum. The framework can guide museums in developing strategies for what data to model (and to what level of detail) when transforming their provenance records into LOD. In its adaptability, our framework allows for depth of description, where needed, through a layered approach to building provenance—a thickness that is in itself networked and collaborative, and potentially inclusive of a multiplicity of voices, thus addressing the changing role of museums today and the diverse expectations they face. At its core, this conceptual framework allows for institution-specific strategies while arguing for the use of collaboratively developed resources.

2.2 Legacies of Provenance

In its basic definition, provenance (derived from the Latin *provenire*, or "to originate") refers to an origin, to where a thing comes from; in art history, a provenance is considered a record of ownership changes of a cultural artifact.[7] However, as Gail Feigenbaum and Inge Reist have noted, "Provenance, firmly entrenched though it may be as a standard part of art historical research today, is neither stable as a concept nor constant as an instrument."[8] We build on this insight here to highlight three interrelated dimensions of legacy information that need ongoing critical engagement as the digitization of provenances continues apace: that this information

7. Yeide, Akinsha, and Walsh (2001).
8. Feigenbaum and Reist (2013).

has served specific and sometimes competing interests, that, by convention, it merely approximates historical complexity, and that it has been shaped by cultural biases, implicit and explicit.

The goals of the different practitioners of provenance have not necessarily aligned and may have even been opposed. Provenance has been and continues to be produced in varying contexts, such as museums, the academy, and the art market. Curators, art historians, art dealers—all have brought their agendas to the writing of provenance, whether driven by scholarly or commercial interests, or both. Before museums began to emerge as independent institutions in the eighteenth century, provenances had a practical use: to substantiate value. Namely, they served as tools to verify (or, in the hands of bad actors, fake) the authenticity and attribution of works.[9] (Provenance still performs this function today, of course.) In a similar vein, provenances containing illustrious names—such as those of European aristocracy—have been used to market works as having "pedigree," turning ownership history into cultural capital (to reference Pierre Bourdieu), by prioritizing important names and omitting lesser-known ones from the record.[10]

Both now and in the past, provenances are the outcome of historical research obtained from a variety of sources, including but not limited to documentary evidence (catalogues, contracts, deeds, photographs, or wills) and material information that can be gleaned from the object itself (inscriptions, labels, or stamps). Yet the richness of the historical complexity that may live in the documentary evidence or the material traces has often not been reflected in provenance records themselves. These are governed by recording conventions that have favored and continue to favor lists of names, creating a kind of provenance shorthand: a highly conventionalized style that was shaped by practical necessities and the demands of the art market. (This is in marked contrast to the space granted discussions of attribution and subject matter on, for example, museum labels and catalogue pages.) In other words, provenances, as conventionally written, have always constituted a reduction or simplification of the information available in the documentary or material evidence—with or without digital transformation.

Provenance records now in the files of institutions and collectors are palimpsests of prior uses and the concomitant historical entanglements and biases of those uses.[11] This means these legacy records may not align with today's expectations for historical accuracy and detail. One significant broad cultural and/or historical bias is gender discrimination. This has resulted in the under- and misrepresentation of women in provenances, still visible in old-fashioned naming conventions, such as

9. See Huemer (2020).
10. See Pergam (2013); and Bourdieu (1984).
11. Feigenbaum and Reist (2013).

"Mrs. John Doe."[12] Similarly, usually only objects that could be attributed to a single maker were given provenances that start with this artist, expressing a cultural bias prioritizing individual authorship.[13] By contrast, provenances of ethnographic objects rarely included any information predating the object's extraction.[14] Sometimes, bias is expressed implicitly in the way transfers of ownership are described: an object that was looted during colonial terror, such as the 1897 sack of Benin City, may be described significantly more neutrally as having been "remov[ed] from the Royal Palace."[15]

These interrelated dimensions of legacy information are still present in today's data but are being corrected by new approaches to the use of provenance records. Inspired by anthropologist Arjun Appadurai's concept of the social life of things, provenances have recently been used as a means of narrating biographies of works and have become an invaluable source for histories of taste, collecting, and art markets. Moreover, they have become the main source for identifying unlawfully appropriated artworks.[16] These newer political and legal demands regarding objects from contexts of injustice have put pressure on museums and changed their practices with regard to provenance, as we will discuss in the next section.

2.3 Provenance and the Museum

As the ultimate repositories for many artworks, museums have become, of all stakeholders, the most intimately engaged with the practice of provenance. Their role has become that of a clearinghouse for the diverse historical and historiographical tendencies that have shaped the production of provenance and the varying ways of studying, compiling, and recording information that underlie them. Today museums face a responsibility to account for the histories of the objects they own. When museums encounter provenance gaps—especially for historically problematic contexts, such as World War II or the colonial era—it falls to them to actively fill in the missing information, which often requires resource-intensive research in multiple archives spread over several countries. Regardless of what provenance information museums

12. Niederacher (2012). In 2021 Stanford University's Archeological Collections published an online exhibition titled Women in Provenance, available at https://storymaps.arcgis.com/stories/5bf914cc05164cd2a7758457567f7c33#ref-n-4YiOcR, accessed December 12, 2022.

13. Gagliardi (2020). Gagliardi highlights the problematic character of attributing authorship of an African bronze statue to an entire ethnic group given that a European bronze statue is often easily attributed to a single creator such as Picasso.

14. Higonnet (2013).

15. See, for example, the provenance for the Edo Queen Mother Pendant Mask at the Metropolitan Museum of Art, 1978.412.323, https://www.metmuseum.org/art/collection/search/318622, accessed on Oct 14, 2022.

16. See Amineddoleh (2020); and Rother and Schmeisser (2020).

may receive, they carry the burden of ensuring that the provenances they publish and produce are in line with contemporary standards. This section describes the context in which, relatively recently, museums' standards for recording provenance emerged. Any move toward provenance LOD will have to grapple with these standards while also recognizing the problematic character of different forms of legacy information.

The end of the Cold War and the unification of the two German states constitute a historical watershed for museums and their engagement with provenances. This development coincided with the rise of memory studies, especially concerning the legacies of the Holocaust. While the geopolitical changes meant that artworks and archives that had been difficult to access since the end of World War II were now available to study, a new awareness, especially in unified Germany, arose around the need to address the injustices related to National Socialism. The 1998 Washington Conference Principles on Nazi-Confiscated Art constituted a culmination of these developments and established a scientifically rigorous provenance practice, especially in museums. With their commitment to transparency, the so-called Washington Principles, endorsed by delegates from forty-four countries and non-governmental organizations, codified in a legally nonbinding way measures to make establishing provenance easier as a step toward restitution and historical justice.[17] More recently, provenances have also become essential in researching unlawfully appropriated objects in the context of European colonialism and the extractivist policies that exploited both natural and cultural resources.[18] Following the example set by the Washington Principles, Germany (to cite just one country with a colonial legacy) has established its own guidelines on the documentation and publication of collections from colonial contexts, the so-called 3-Road Strategy.[19]

In light of these developments, in recent years many museums have devoted resources to examine works possibly affected by National Socialism in particular, and this field has produced the best examples of detailed documentation of provenance.[20] With the available information, we can describe in detail specific events, such as the removal of what the Nazis labeled "degenerate art" from museums in 1937, for example, or confiscations during World War II undertaken by the Einsatzstab Reichsleiter

17. U.S. Department of State, Office of the Special Envoy for Holocaust Issues (1998).

18. See Grimme (2020); and Zuschlag (2019).

19. "3-Road Strategy on the Documentation and Digital Publication of Collections from Colonial Contexts Held in Germany, German Contact Point for Collections from Colonial Contexts, https://www.cp3c.org/3-road-strategy/.

20. Especially in the German-speaking countries, a vast amount of provenance scholarship related to National Socialism has been produced over the last twenty years. For a critical take on this development, see Fuhrmeister and Hopp (2019); and Masurovsky (2020). Specifically for the US context, see Milosch, Nicholas, and Fontanella (2014).

Rosenberg, or ERR (Reichsleiter Rosenberg Taskforce).[21] The provenances of affected objects may therefore be rich in information for a short period of the object's life but thin for much of the rest. Moreover, as provenance research is sometimes funded by third parties (such as the Samuel H. Kress Foundation), and this research often targets specific types of objects or specific historical moments or geographies, there can exist biases within collections around which objects have more detailed provenances—especially at institutions that have not engaged in systematic provenance work before.[22]

In response to the Washington Principles and the aforementioned discrepancies in producing provenances, the American Association of Museums (AAM) published their guide to provenance research in 2001. The guide covers a host of issues relating to provenance research broadly, and it also offers parameters for how to record provenance information, aiming to establish a modicum of standardization across institutions.[23] The AAM format, which developed out of an analog, textual provenance practice, uses syntax to create limited structure in the provenance. It relies on punctuation to convey meaning: a period records a gap between events, whereas a semicolon between events signals that the second event was directly subsequent to the first; brackets set off the life dates of parties; and parentheses mark the type of party in an event, such as a dealer or agent. Information that is not immediately relevant to the actual transfer of ownership or title—for example, knowledge about location changes, consignment status, or illegal transfers, say in the context of National Socialist expropriation—may be kept in separate notes.

In line with the efforts initiated by the Washington Principles to make provenance information public, museums, particularly in the United States, have begun to transfer their provenance texts from the analog originals, usually held in object files in the registrar's office or curatorial departments, to the digital domain. Some institutions, such as the Metropolitan Museum of Art in New York and the Art Institute of Chicago, have done so for tens, if not hundreds of thousands, of works. These provenances mostly live in free, unstructured text fields in the museums' col-

21. See, for example, the confiscation inventory database of "Entartete Kunst" (degenerate art), Freie Universität Berlin (https://www.geschkult.fu-berlin.de/e/khi/ressourcen/diathek/beschlagnahmeinventar/index.html) and "Cultural Plunder: Database of Art Objects at the Jeu de Paume" by the Einsatzstab Reichsleiter Rosenberg (https://www.errproject.org/jeudepaume/). See also Fleckner (2015); Barron (1991); Nicholas (1995).

22. For example, the Presidential Advisory Commission on Holocaust Assets, the American Alliance of Museums (former American Association of Museums), and the Association of Art Museum Directors established guidelines regarding objects misappropriated during the National Socialist Era in 1999 that recommended an initial focus on European paintings and Judaica. See https://www.aam-us.org/programs/ethics-standards-and-professional-practices/unlawful-appropriation-of-objects-during-the-nazi-era/, accessed Oct 14, 2022.

23. Yeide, Akinsha, and Walsh (2001).

lection management databases, which is often the source material that populates the museums' collection websites. Figure 2.1 shows one such example from the collection management system of the National Gallery of Art in Washington, DC: the provenance of a 1940 painting by Henri Matisse, *Woman Seated in an Armchair*. We will return to this example throughout this essay.[24]

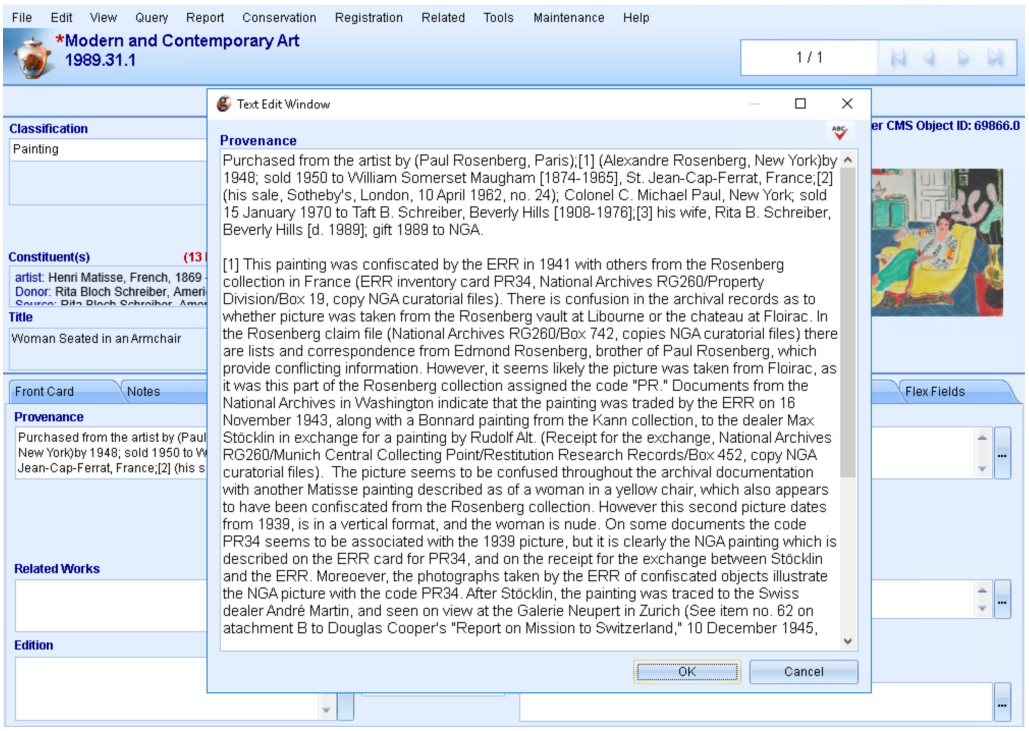

Figure 2.1: The provenance of Henri Matisse, *Woman Seated in an Armchair*, as it appears in the free, unstructured text field of the collection management system of the National Gallery of Art, Washington, DC.

These provenances mainly conform to the limited best practice format laid out by the AAM guide. However, even within these parameters, there exists a fair amount of variety in provenances. And while variations do exist among museums (due to what we might call "house style" for recording provenance), they also exist across disciplines and curatorial departments, as well as within departments—even within single records, where it is sometimes possible to pinpoint provenance styles belonging

24. The full tombstone information for this work, as given on the National Gallery website, is as follows: Henri Matisse, Woman Seated in an Armchair, 1940. Oil on canvas. National Gallery of Art, Washington DC, Given in loving memory of her husband, Taft Schreiber, by Rita Schreiber, 1989.31.1, https://www.nga.gov/collection/art-object-page.71071.html. The National Gallery is, in fact, one of the few museums that enters the names of former owners in not only its free text provenance field but also the "Constituent Assistant," provided by the collection management software The Museum System (by Gallery Systems), allowing searchability for names.

to different individuals.[25] With digital provenance records now freely available on museum websites, the heterogeneity of provenance is more visible than ever before.

The increasing online publishing of provenance provides a basic level of transparency, useful especially for people with claims on specific artworks. We must remember that having a published provenance, even if only in a free text field, is better than not having *any* published provenance—which continues to be the case for the majority of works across thousands of institutions. While artworks can be found with relative ease through basic tombstone information (creator, title, date of creation, medium, dimensions), this is only possible when claimants know precisely what they are looking for and, furthermore, when the attribution, title, dimensions, and so on have not changed over time. As re-attributions and other changes are common for cultural objects, the presence of a specific family name or a gap within a published provenance on a museum website may be precisely what would allow claimants to find potentially looted objects. As cumbersome as that process may be, given that claimants would have to look up objects individually, object by object, at every museum because provenance criteria remain unsearchable on the majority of museum websites, claimants would at least have the possibility to search for some criteria. Lastly, if a provenance is published, even if only in the basic format laid out by the AAM, when new information appears, it can be kept up to date by editing the underlying information in the collection management database.

We must acknowledge the considerable effort it took museums and other cultural heritage institutions to arrive at this basic level of provenance transparency and recognize it as the game-changing undertaking that it is and continues to be. Museum documentation is a herculean task with occasionally competing goals. Institutions have to balance researching, recording, and keeping records up to date with incorporating information from new sources, as well as responding to even more fundamental shifts such as the present moment's calls to decolonize the museum (and not only ethnographic museums).[26] At the same time—and this is the crucial part—given the ever-increasing amount of information available, research being done, and new archival sources accessible digitally, museums face legitimate questions as to whether their current provision of provenance information is adequate or not.

Considered together, the variation in how provenance texts are organized and written makes it clear that we are dealing with diverging concepts of what prove-

25. For example, at the Museum of Modern Art in New York, there exist different provenance formats for the same artist, some making use of the AAM format and some not. This is the product of different authors working at different times. Compare, for instance, the provenance for two works by Giorgio de Chirico: https://www.moma.org/collection/works/78738 and https://www.moma.org/collection/works/80588, accessed October 14, 2022.

26. Efforts at decolonizing the museum are not entirely new developments and can be traced back to the post–World War II era. See Wintle (2013).

nance is and what purpose it serves. Since we are concerned in this essay with the question of provenance *data*, we will need to keep these heterogeneous and sometimes ideologically unreconstructed provenance *texts* in mind as we consider the technical possibilities available for transforming existing texts into data.

2.4 Toward Provenance Data

In this third section, we address the difference between unstructured data in text-based provenance records and structured provenance data. We do so to draw out the kinds of considerations that need to be kept in mind when data is being structured—labor performed by humans, aided or not by artificial intelligence—and to point out the potentials and pitfalls of structured and unstructured data, respectively.[27]

Let us begin by looking at the limitations of unstructured data, i.e., information whose internal logic is not explicit to computers and hence cannot be processed automatically. First, recording provenances in free text fields in collection management systems (as shown in figure 2.1 above) not only silos information but also duplicates (or multiplies) it within the database: for example, each mention of a particular collector's name does not reference a single, digitally available biography record of that person but rather constitutes a unique appearance in each provenance in which it is referenced, without any connection to its other appearances. This, in turn, multiplies the work involved in the upkeep of this information and the addition of new research. For example, if research uncovers new, essential documents about a nineteenth-century Parisian art dealer that affect the provenance records of multiple artworks in a collection, each record would need to be updated individually. Free-text provenances, even those produced and shared in line with the AAM format, also make it difficult to perform complex queries because they are not machine readable, which means that the structure required for a computer to understand the textual logic of the provenance is missing. This is highly relevant for people with legal claims to specific works that may have been looted or otherwise unlawfully expropriated. Structured data can be queried by multiple parameters simultaneously, for example, by searching for objects that meet the criteria of being produced by Edo people and are known to have left the territory of Nigeria before its independence from the United Kingdom on October 1, 1960. Such a query would include only the objects meeting these criteria, and exclude those acquired legally after the end

27. The development of NLP (Natural Language Processing) techniques over the past fifty years allows for extracting meaning from unstructured data. In the context of the Provenance Lab at Leuphana University Lüneburg, the authors of this essay are developing statistical models to apply NLP techniques on unstructured provenance text to extract meaning and facilitate structuring. See Rother, Mariani, and Koss (2023).

of colonial rule, for example. Without machine-readable data, claimants have to go through databases object record by object record, one museum at a time, to find objects with provenance gaps or information that might fit their search criteria, as indicated earlier.[28]

Siloing information also prevents analysis within or across institutions, and creates unnecessary obstacles in linking to data, such as archival materials or digitized auction catalogues, provided by external sources. Continuing with the example of Matisse, we would ideally like to be able to use provenance in databases to answer a question such as, "For how many and for which Matisse paintings sold between 1939 and 1945 in Paris is the purchaser's name known?" But today, this is not a question we can answer through the available data, and thus we are unable to discern the links, patterns, and particular historical trends that would be made visible through the bigger picture—that is, with aggregate data.[29]

Finally, provenance records on museum websites are, currently, often not downloadable. This is a considerable impediment to networked research, which is a fundamental aspect of provenance research, in which researchers rely heavily on the work and findings of other researchers.[30] It is also exclusionary inasmuch as only a small number of individuals associated with the institution can access and edit the information. While it is understandable that museums feel a sense of responsibility around ensuring the accuracy of this information, it also may have an unintentional gatekeeping effect: it perpetuates professional, institutional, and disciplinary biases and hampers epistemological shifts toward a more inclusive, multi-perspectival approach.[31]

The shortcomings we have just described are particularly problematic in the context of looting, appropriation, and restitution, as well as in the context of accessibility. However, they are by no means limited to such cases. For example, authentication and identification of fakes could be much improved by cross-referencing and triangulating the data of objects, provenance records, and the records of collectors and dealers. Indeed, the history of collecting and art markets would benefit from identifying trends and patterns in large amounts of data, decentering consideration

28. It is noteworthy that in 2003 the AAM launched the Nazi-Era Provenance Internet Portal to facilitate the search for objects potentially affected by National Socialist looting across US museum collections. However, while the portal provides object information and links to the museum websites, it does not provide provenance information and museums have not updated their registered objects. See https://nepip.org. In 2022 the German government launched a database for holdings from colonial contexts in German museums, recording 6,636 objects. See https://ccc.deutsche-digitale-bibliothek.de/?lang=en, accessed October 14, 2022.

29. With regard to exhibition data and how the computational analysis can provide new insights to art history, see Greenwald (2021); and Joyeux-Prunel and Marcel (2015).

30. Weber-Sinn and Ivanov (2020); and Fuhrmeister and Hopp (2019).

31. See Bell, Christen, and Turin (2013).

of individual objects in isolation and turning the focus of research to potentially unaddressed historical phenomena involving many objects.[32] The list of areas in which the analysis of provenance data could be useful is long: identifying tax fraud, money laundering, black market movements, and the distribution (or lack thereof) of wealth and capital across time and geography, to name a few.

Provenance

Purchased from the artist by (Paul Rosenberg, Paris);[1] (Alexandre Rosenberg, New York)by 1948; sold 1950 to William Somerset Maugham [1874-1965], St. Jean-Cap-Ferrat, France;[2] (his sale, Sotheby's, London, 10 April 1962, no. 24); Colonel C. Michael Paul, New York; sold 15 January 1970 to Taft B. Schreiber, Beverly Hills [1908-1976];[3] his wife, Rita B. Schreiber, Beverly Hills [d. 1989]; gift 1989 to NGA.

[1] This painting was confiscated by the ERR in 1941 with others from the Rosenberg collection in France (ERR inventory card PR34, National Archives RG260/Property Division/Box 19, copy NGA curatorial files). There is confusion in the archival records as to whether picture was taken from the Rosenberg vault at Libourne or the chateau at Floirac. In the Rosenberg claim file (National Archives RG260/Box 742, copies NGA curatorial files) there are lists and correspondence from Edmond Rosenberg, brother of Paul Rosenberg, which provide conflicting information. However, it seems likely the picture was taken from Floirac, as it was this part of the Rosenberg collection assigned the code "PR." Documents from the National Archives in Washington indicate that the painting was traded by the ERR on 16 November 1943, along with a Bonnard painting from the Kann collection, to the dealer Max Stöcklin in exchange for a painting by Rudolf Alt. (Receipt for the exchange, National Archives RG260/Munich Central Collecting Point/Restitution Research Records/Box 452, copy NGA curatorial files). The picture seems to be confused throughout the archival documentation with another Matisse painting described as of a woman in a yellow chair, which also appears to have been confiscated from the Rosenberg collection. However this second picture dates from 1939, is in a vertical format, and the woman is nude. On some documents the code PR34 seems to be associated with the 1939 picture, but it is clearly the NGA painting which is described on the ERR card for PR34, and on the receipt for the exchange between Stöcklin and the ERR. Moreoever, the photographs taken by the ERR of confiscated objects illustrate the NGA picture with the code PR34. After Stöcklin, the painting was traced to the Swiss dealer André Martin, and seen on view at the Galerie Neupert in Zurich (See item no. 62 on atachment B to Douglas Cooper's "Report on Mission to Switzerland," 10 December 1945, National Archives RG239/Entry 73/Box 82, copy NGA curatorial files).

The NGA picture was returned to the Rosenbergs by 1948, according to the records of the gallery, which sold it to Somerset Maugham in 1950.

[2]According to Rosenberg gallery records.

[3]Correspondence between Paul and Schreiber in NGA curatorial files.

Associated Names

ERR
Martin, André
Maugham, William Somerset
Paul, C. Michael
Rosenberg & Co., Paul P.
Schreiber, Rita Bloch
Schreiber, Taft B.
Sotheby's
Stöcklin, Max

Figure 2.2: The provenance, with notes, of Henri Matisse's *Woman Seated in an Armchair*, as published on the website of the National Gallery of Art, Washington, DC, as it appeared on November 22, 2022.

32. See, e.g. Rother, Mariani, and Koss (2023).

The limitations of unstructured provenance records notwithstanding, such records, like the one provided by the National Gallery for its Matisse painting (fig. 2.2), do currently allow for full-text searching: finding a particular string of alphanumeric characters (any combination of letters, numbers, and special characters such as exclamation marks) within their texts. The computer can, for example, retrieve all provenance records containing a particular string of characters such as "Paul Rosenberg." With such a search, the computer will find all records containing "Paul Rosenberg," opening up possibilities for analyzing the art dealer's importance for the collection. However, as the example from the National Gallery shows, we could not reliably find this very record by performing a search for the string "Matisse" in the museum's provenance fields. While Matisse was the first owner of this work, it cannot be found by searching for "Matisse" because he is recorded as "the artist" and not explicitly named. Instead, this information is only captured as so-called tombstone data and thus lives in a separate field in the database. Similarly, searching for "Paul Rosenberg" in unstructured provenance texts will return all entries containing "Paul Rosenberg" without any further specification. This means the search will bring up all objects in which his name appears with that exact spelling, regardless of his role in a particular provenance—whether he was buying or selling an object, or even just published a book referenced in the notes of the provenance text. For searches to return the objects that meet a set of criteria, such as a particular individual playing a particular role, we need structuring.

Table 2.1 shows an example of structured provenance data for *Woman Seated in an Armchair*, in which, for reasons of clarity, we have replaced "artist" with "Henri Matisse." Because we do not know the exact date when the object passed from Paul to Alexander Rosenberg, only that it happened before or in 1948, we have structured this information according to the Extended Date/Time Format (EDTF), created by the Library of Congress to address data fuzziness in date and time formats.[33] In this case, we chose a tabular data structure: each row of the table represents a provenance event, and each column represents an attribute associated with it, such as the parties involved (the sender, the receiver, and any intermediary agents), its location, its time, and the method by which the transfer was carried out.

If we were to structure provenances in a standardized way on a large scale—say, all the provenance records from the National Gallery—it would be possible to query the data by analyzing the events (the rows of the tables) through the characteristics expressed by each column. For example, it would be possible to formulate queries such as "Give me all the events in which Paul Rosenberg bought an object from Henri Matisse." To answer this query, the machine can count all rows of all the tables in which "Paul Rosenberg" is the value in the "Receiver" column and "Henri

33. See https://www.loc.gov/standards/datetime/.

Sender	Receiver	Agent	Location	Time	Method of Transfer
Henri Matisse	Paul Rosenberg				purchase
Paul Rosenberg	Alexandre Rosenberg			..1984	
Alexandre Rosenberg	William Somerset Maugham			1950	sale
William Somerset Maugham	Colonel C. Michael Paul	Sotheby's	London	1962-04-10	auction
Colonel C. Michael Paul	Taft B. Schreiber			1970-01-15	sale
Taft B. Schreiber	Rita B. Schreiber				
Rita B. Schreiber	NGA			1989	gift

Table 2.1: The provenance of Henri Matisse, *Woman Seated in an Armchair*, structured in a table.

Matisse" is the value in the "Sender" column. The value of structuring data thus lies in how it allows for more complex quantitative analysis. One can store a tabular data structure in proprietary formats such as Microsoft Excel or open formats such as CSV (Comma Separated Values).

The information's semantics is not explicit in a table, however: the machine does not have a semantic understanding of any of the columns. It knows that "London," for example, is in the "Location" column but is oblivious to the fact that it represents a location in space with administrative or geographic value (e.g., a city) that is part of a larger area (e.g., a region or a country). For this reason, even though the machine can perform a search with a given query, it cannot use the implicit semantic value of the table and therefore cannot combine it with external knowledge (e.g., data from other museums or repositories) to infer alternative knowledge. For example, the provenance of *Woman Seated in an Armchair* involves a 1962 Sotheby's auction held in London, yet the machine has no understanding that London is in England. This information is not in the text, so it probably would not be included during the data structuring process. Nevertheless, associating the London entity in the provenance to the respective entity in the Getty Thesaurus of Geographical Names would add the notion that London is in England and is its current capital. With this additional data and the proper provenance knowledge modeling (which we will address in the next section), the machine could logically infer that the auction was held in England, more precisely in the capital, thereby increasing possibilities for analysis and research (e.g., on the art market in capital cities and the art market in England).

The museum could instead add this geographical information to its database once

the provenance is structured. However, even for elementary notions such as "London is the capital of England," such effort would require extra and unnecessary work to structure knowledge. Given that such information is already findable, accessible, interoperable, and reusable on the web in the form of LOD, it is best to simply make use of it.

2.5 Provenance Linked Open Data

As noted in the previous section, structured data can help provide museum professionals and researchers with quantitative insights into specific collections and collecting histories. However, when an individual museum structures its data in an idiosyncratic way that is not compatible with how other organizations have done it, its usefulness will be limited only to queries about its collection. To overcome this limitation, museums will have to structure their data according to a set of shared principles that makes the data findable, accessible, interoperable, and reusable—that is, according to a set of principles known as the FAIR principles, which the scientific community established in 2016.[34]

However, for historical research (and provenance research is just that), the FAIR principles alone are insufficient for an inclusive, multi-perspectival approach, as these principles do not deal with open data's ethical and moral implications. While *accessible* data can be potentially *open*, it is not necessarily so.[35] Open data by definition "can be freely used, modified, and shared by anyone for any purpose," and to produce open provenance data, therefore, we must apply a data standard that respects both the FAIR principles and the open principle.[36] That standard is LOD, which relies on structured data and can link data in a way that allows for complex and potentially valuable queries, especially across institutions.[37]

Publishing LOD, built on web standards, means publishing resources online and identifying them through URIs (universal resource identifiers, i.e., a unique name for a given resource), such as the URI of the Getty's Union List of Artist Names referencing Henri Matisse; http://vocab.getty.edu/ulan/500017300. In addition, the URIs used to identify LOD resources are HTTP URIs, that is, URIs associated with the hypertext transfer protocol (HTTP). This type of URI makes every LOD resource findable. The curation and preservation of URIs are two of the core re-

34. Wilkinson et al. (2016).

35. Mons et al. (2017).

36. "Open Data," Open Definition 2.1, Open Knowledge Foundation, https://opendefinition.org/.

37. The potential usefulness of LOD is perhaps best summarized by Tim Berners-Lee, the co-inventor of the world wide web: "With linked data, when you have some of it, you can find other, related data." Tim Berners-Lee, "Linked Data," W3.org, https://www.w3.org/DesignIssues/LinkedData.html.

sponsibilities of linked open-data producers: indeed, the stability of URIs and their maintenance are a prerequisite for their long-term usefulness. The LOD community established another standard, RDF (resource description framework), to describe relationships between resources identified by URIs.[38] This standard relies on the fact that every entity and every relationship between entities in a given dataset can be identified by a discrete URI. In the context of provenance records, such entities can be people, organizations, objects, or events, and the relationship is what binds two such entities together, usually expressed by a verb (i.e., "Paul Rosenberg has French nationality"). As we have seen already, such descriptions come in the form of syntactical statements (i.e., sentences) based on a triple structure: subject—predicate—object. Thus, in our example, Paul Rosenberg is the subject (http://vocab.getty.edu/ulan/500372940) with the predicate of having a nationality (http://schema.org/nationality) that is French, which is specified as the object (http://vocab.getty.edu/aat/300111188).

The advantage of a shared standard based on these so-called triples is that they are structured and hence machine-readable, thus allowing for queries. Because these triples are constructed with stable URIs, we can analyze this data quantitatively not only within one museum collection but also in the context of the entire world wide web, where other museums can also publish their information following the same standard. Furthermore, this syntax also allows us to make descriptions using URIs managed by multiple stakeholders, which means that the labor involved in producing provenance LOD is distributed across institutions because all producers can rely on the interoperability and reuse of each other's linked open data.[39]

To guarantee interoperability between one's own provenance LOD and that of other stakeholders, it is essential to build data according to not only LOD standards such as RDF but also to a shared community standard so that triples are built from a common set of URIs. This shared standard is CIDOC CRM (Conceptual Reference Model of the International Committee for Documentation). It is an ISO standard (ISO 21127) developed by museum professionals under the auspices of the International Council of Museums (ICOM) since 1996.[40] As a standardized ontology (a data model that defines what kinds of things make up a domain, and what kinds of relationships exist between them), it currently provides URIs for 160 properties (such as "P74 has current or former residence," useful to, e.g., describe that Paul Rosenberg has residence in Paris) and 81 classes, categories to which an entity can

38. "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3.org, February 10, 2004, https://www.w3.org/TR/rdf-concepts/.

39. *Interoperability* is the technical integration of data (the structure of other data is compatible with one's own), whereas *reusability* refers to the legal integration of data, i.e. the rights related to that data (what can and cannot be done with that data).

40. Bekiari et al. (2021).

belong (e.g., Paul Rosenberg belongs to the class: "E21 Person").[41]

One of the most critical aspects of CIDOC CRM is its event-oriented modeling. Whereas the AAM format introduced earlier proposes a list of owners—an object-centric method—from which a chronology of events can be deduced, but does not build the provenance on events, provenance LOD (built on the CIDOC CRM standard) does. In other words, building LOD provenance data from AAM-formatted provenance records involves transposing an object-centered structure to an event-oriented structure that links people or organizations to events that may involve one or more objects. With this potential for knowledge production in mind, the Linked Art community is currently actively developing a data model built on CIDOC CRM that will cater to the specific needs of art museums. With a pared-down version of CIDOC CRM, Linked Art aims to encourage LOD implementation for museums that want their collection data to "be part of the Web, and not just on the Web."[42]

Despite its potential and ease of use, the Linked Art Data Model has not yet been applied across institutions in a real-world scenario.[43] In fact, only two projects that structure and link provenance data have been developed and published so far. In 2000, the ethnographic collection at the Museum of Cultural History in Oslo pioneered provenance data modeling when it implemented an event-oriented database for curation and research purposes using CIDOC CRM.[44] Even though LOD standards were far from defined, the Museum of Cultural History modeled 50,000 object records with 2 million events and 3.6 million relationships between events, objects, parties, locations, and time. The second project is Art Tracks, developed at the Carnegie Museum of Art in Pittsburgh between 2014 and 2017.[45] The project's goal was to reconstruct the history of Old Master paintings from one collection, the Northbrook Collection of the Baring family, and then to visualize the collection's growth and later dispersal on a digital map and timeline for the benefit of gallery visitors and online users. Art Tracks was at the forefront of applying technologies of structuring and linking to the problem of generating well-formatted provenance data from free-text information based on the AAM format.[46] As of the time of writ-

41. For the example of a property such as "P74 has current or former residence," see http://cidoc-crm.org/cidoc-crm/7.1.1/P74_has_current_or_former_residence and for an example of a class such as "E21 Person", see http://cidoc-crm.org/cidoc-crm/7.1.1/E21_Person.

42. "Linked Art Profile of CIDOC-CRM" Linked Art, https://linked.art/model/profile/.

43. In 2019 the Georgia O'Keeffe Museum in Santa Fe became the first museum to apply the Linked Art Data Model across its collections under the auspices of Liz Neely from the Georgia O'Keeffe Museum and Duane Degler from Design for Context. See their Linked Data documentation here: http://gokm-docs.okeeffemuseum.org/. We must note that while provenance information was included, it was not structured.

44. Jordal, Uleberg, and Hauge (2012).

45. Newbury (2017).

46. Art Tracks operates with a format on their own, which they call a superset of the AAM format.

ing, this extension of the AAM format proposed by Art Tracks is the only existing provenance text standard that anticipates machine readability.[47]

Both the CIDOC CRM and Linked Art communities and the Oslo and Art Tracks projects have begun to address challenges specific to provenance data, such as how to model gaps and uncertainties and how to record subjective assertions. Because of their importance for documentation and potential further research, these aspects have been discussed in the recent literature on digital provenance.[48] However, while some extant modeling techniques are waiting to be applied and tested on a large set of real-world provenance data, others still have to be conceptualized and defined.

Although LOD theoretically allows for modeling all historical information that can be found in and around provenance records, such as in notes or supplemental documents, it has not been determined what kind of provenance information should be selected for structuring and linking—and which should *not* since it is already available on the internet and would thus not constitute a smart use of resources. As mentioned above, massed data projects can be resource intensive; to quote computer scientist Ian Foster, "the creation, curation, maintenance, and delivery of digital information are all expensive and time-consuming activities."[49] This is true as well for museum documentation and provenance data, within which complexity, biases, and subjectivity require human intervention when structuring and linking. This requisite human element makes a resource-conscious approach to producing provenance LOD even more crucial.

A strictly economical approach would limit the structuring and linking to data directly related to ownership changes, in a sense replicating the traditional concept of provenance as a list of different owners. On this approach, any additional information would not be part of the museum's linked provenance data. Such information includes custody changes of an object (as opposed to ownership changes) as well as biographical information about the people involved, including their birth and death dates, alternative spellings of their names, and the relationships among them. However, in the spirit of the LOD principles laid out above, such additional information could be sourced from external LOD repositories that depend either on crowdsourced knowledge, such as Wikidata, or on more scientifically reliable terminologies (or vocabularies) for the cultural heritage domain, especially those provided and edited by the Getty Research Institute such as ULAN (for parties), AAT (for

47. See "The CMOA Digital Provenance Standard," draft version 0.2, Art Tracks: A Project of the Carnegie Museum of Art, October 14, 2016, http://www.museumprovenance.org/reference/standard/.

48. See especially Lincoln and Ginhoven (2018).

49. Foster (2011).

methods of transfer), and TGN (for historical locations).[50] Such crowdsourced platforms and authoritative institutions produce and maintain data, which museums can link to and which machines would be able to find and analyze in conjunction with provenance data from across museum collections. Using LOD from external sources in this manner addresses one of the significant pitfalls of digitizing provenance records in a structured way: the potential loss of information that is part of the provenance record but has nowhere to "live" in a structured environment—for example, biographical information that is not directly relevant to the event's description. However, for the (art) historian, such "extra" information can be useful, and therefore museums might want to preserve it.

We now aim to show how digitized provenance can be structured based on LOD principles, once again using the provenance of Matisse's *Woman Seated in an Armchair*. In its current form, this provenance is recorded in the syntactical AAM format, which tells us (focusing for a moment only on the first provenance event listed) that Paul Rosenberg purchased the painting from Henri Matisse. However, the provenance specifies neither the location nor the time of this event. On the other hand, it gives biographical information for both parties involved. Matisse is called "the artist," and Rosenberg's location is given as Paris, while his occupation is listed as "art dealer." The museum gives that information in an institution-specific style using parentheses.[51]

Both parties (Matisse and Rosenberg) and the method of transfer (purchase) are already recorded in ULAN or the AAT. Happily, machine-learning methods can assist humans in structuring and linking provenance events and their core entities. Already, semi-automated data services, such as the Getty Vocabularies OpenRefine reconciliation tool, can help link the event's entities to ULAN and the ATT.[52] With such a resource-conscious approach, the structuring effort would be relatively small but deliver considerable results. In the case of this example, perhaps even more important than the cost-effectiveness would be the fact that Paul Rosenberg, art dealer, would be indexed by a stable URI that would not only tie the object to the "right" Paul Rosenberg but immediately put it in relation to all other objects tied to this Paul Rosenberg. While not error-proof (authority files are not 100% accurate all the time), if multiple museums would take even a limited LOD approach, it would, with little effort but high reliability, enable researchers to find objects purchased by

50. ULAN is the Getty Union List of Artist Names, AAT is the Getty Art & Architecture Thesaurus, and the TGN is the Getty Thesaurus of Geographical Names. See https://www.getty.edu/research/tools/vocabularies/.

51. "Reading Collection Information: Provenance Texts," National Gallery of Art, https://www.nga.gov/collection/collection-information.html.

52. OpenRefine is an open-source tool for data clean up and transformation. See https://www.getty.edu/research/tools/vocabularies/obtain/openrefine/.

Paul Rosenberg directly from Matisse (or more generally from artists), across these museum collections.

In light of these considerations, we can model the information of the first provenance event described by the National Gallery of Art—that is, Paul Rosenberg's purchase of the painting from Henri Matisse—and show it in a diagram (fig. 2.3).[53] Using this limited example for demonstration purposes, the information is structured using the Linked Art Data Model based on CIDOC CRM; we have reused and linked the ULAN entities for Henri Matisse and Paul Rosenberg; and the acquisition method, "purchased," comes from ATT. The resulting structured data is thus provenance LOD.

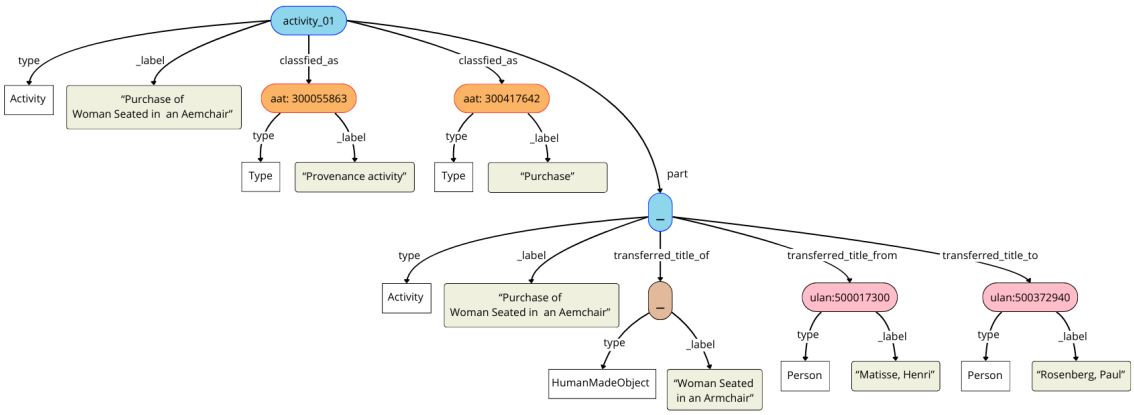

Figure 2.3: The purchase of *Woman Seated in an Armchair* by Paul Rosenberg from Henri Matisse, described in a diagram using the Linked Art data model.

The more provenance research scholars conduct, the more complicated the object histories become, which we can observe in the data. For example, figure 2.2 shows the complexity of Paul Rosenberg's ownership, recorded in a note longer than the entire provenance record. From the note, we learn that the painting was confiscated by the ERR in 1941, together with other paintings from two of Paul Rosenberg's storage locations, either in Libourne or at a chateau at Floirac. It was then traded together with another painting by another artist from another looted collection in France to a dealer and was later seen in Switzerland before eventually being returned to Paul Rosenberg. In a strict understanding of provenance as a sequence of mere ownership changes and by applying US legal standards according to which looted paintings do not change title (as is the provenance practice at the National Gallery of Art), this information would be omitted because Paul Rosenberg was the painting's only owner through all the cruel twists and turns of World War II. However, from a

53. We created the diagram using the Mermaid JS library (https://mermaid-js.github.io/), adapting Linked Art's layout and style (https://linked.art/model/intro/).

holistic perspective, secondary provenance such as the painting's changes in location provide vital insights. Just imagine, for example, the possibilities for provenance research related to National Socialism and the potential queries one could bring to the data: the objects that have been identified as being looted from one of Paul Rosenberg's storage spaces; the objects that were exchanged with and against each other by the ERR; the objects that have (or have not) been restituted to the Paul Rosenberg family.

Discarding well-documented and researched provenance information when moving or remodeling provenances in line with LOD principles would not constitute good scientific practice and would go against the AAM recommendations for publishing provenance information.[54] On the other hand, structuring these additional events—with their parties, transfers, locations, and times, plus the additional detail, some of it incomplete, vague, uncertain, or subjective—increases the overall effort needed to transition to structured provenance data and cannot be done without significant expert intervention. These hurdles notwithstanding (including those that have yet to be fully addressed from a modeling perspective, such as uncertainty), we believe that this extra effort is not only desirable but indeed necessary for an expanded notion of provenance that counters the risk of reduction implicit in the AAM format.

Provenance LOD is nowhere to be found today, as museums are still by and large merely digitizing provenance records without introducing *any* kind of machine-readable structure. Yet the kind of modeling that LOD allows for is, as we have just seen, not without its own pitfalls: it requires, above all, significant investment of resources in labor and data infrastructure, and it needs to account for the kinds of omissions, biases, and reductions (as well as complexities) that legacy provenance records contain.

2.6 Strategizing Provenance Data

To help facilitate museums' transitions to provenance LOD (PLOD), we will now propose a conceptual framework for what data to model, and to what level of detail, within a museum dataset. We think of this PLOD conceptual framework not as a strict roadmap but rather a blueprint for formulating a conscious, responsive, and sensitive data strategy, and, because it is conceptual in nature, it can be applied regardless of the data model. We designed it with the application profile of the

54. In fact, the AAM recommends including this data when museums provide information to the public about objects transferred in Europe during National Socialism. American Alliance of Museums, "Recommended Procedures for Providing Information to the Public about Objects Transferred in Europe during the Nazi Era," https://www.aam-us.org/wp-content/uploads/2018/01/nepip-recommended-procedures.pdf.

Linked Art Data Model in mind, as we consider it the current benchmark. The approach that we detail below begins, necessarily, with a base layer of information, capturing the provenance's core entities. A system of layers built from descriptive bricks complements the base layer to allow for a thicker description of the data, thus improving its quality and usefulness. This modularity addresses the need for a compromise between the resources to be invested in digitization and the problem of losing or flattening data. We again emphasize that the possibility of reusing LOD resources from external authorities such as ULAN can obviate the need to describe an entity from scratch.

A provenance record should, in theory, be understood as a sequence of provenance events in chronological order, each comprising one or more transfers, which occur between parties at a given location at a given time. Thus, five entities are required for the base layer: provenance event, parties, transfers, location, and time. The *provenance event* is a meta entity to which the other information is gradually associated. *Parties* are people, organizations, or any constituents involved in the provenance event, acting alone or in groups either as a sender (the one who loses ownership or custody), a receiver (the one who obtains ownership or custody), or as a mediating agent for one or the other. *Transfers* convey either ownership or custody changes by such methods as inheritance, gift, purchase, exchange of objects, looting, or restitution. A provenance event can have one transfer from a sender to a receiver, or it can have multiple transfers because multiple objects changed hands at the same location and the same time, such as an auction like the one held by Sotheby's in London in 1962. The *location* of this provenance event is London. In practice, such a location is usually only recorded and relevant for provenance events with multiple transfers (e.g., an auction). Locations needed for tracing an object's spatial movement tend to be recorded in provenances through the respective parties involved (e.g., through their places of residence). *Time* indicates when the provenance event took place, recorded in the internationally used, unambiguous calendar-and-clock format (ISO 8601). Building the base layer is easy enough for museum practitioners, but they could also apply crowdsourcing or machine-learning methods to existing provenance texts, although that would require a more extensive—but entirely realistic—implementation effort.[55]

In line with their ideals of inclusivity and in the spirit of decentering the individual object, museums should consider including additional information that has often been harvested through time-consuming provenance research and can be crucial for object documentation and data analysis. In our framework, four types of bricks are available for structuring advanced information and enhancing the data:

55. Oomen and Aroyo (2011). See also Newbury (2017), which proposes a similar approach for the base layer.

biographical, economic, geographic, and contextual (see fig. 2.4).

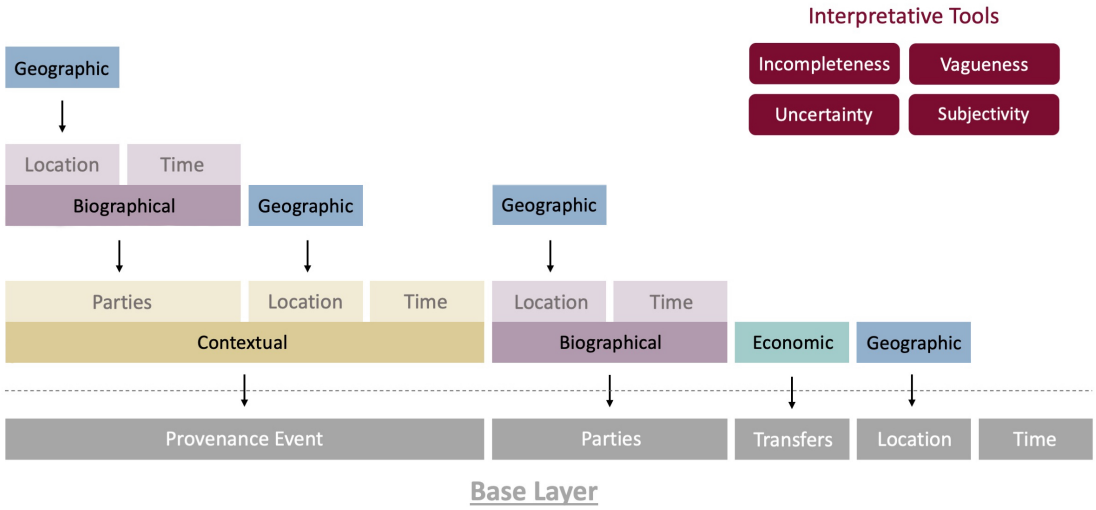

Figure 2.4: The PLOD conceptual framework with the various options for descriptive bricks and interpretative tools.

Provenance records often contain biographical information about the parties that may fall by the wayside during structuring. For example, if birth and death dates have served to disambiguate a person's identity in the provenance text, this data becomes unnecessary in PLOD. If these dates indicate the time of a transfer like an inheritance, the biographical data should be directly linked to the base layer party, either in the museum's dataset or from a trustworthy vocabulary like ULAN. Similar decisions must be made for other biographical aspects of individuals as well as groups and organizations, such as detailed onomastic descriptions, gender, nationality, religion, or life events and relationships between parties, which can, in turn, be modeled with a location and time. Such biographical data allows provenance aspects, like the buying of Matisse paintings, to be brought into dialogue with demographic analysis, which can be especially meaningful for the histories of collecting and taste.

The biographical brick can be used to correct the records of historically marginalized groups that are inadequately represented in LOD. Museums can and should actively contribute this data, becoming a de facto authority for a given digital resource. The Museum of Modern Art in New York, as the holder of the gallery archives of Paul Rosenberg, for example, might want to create its own "Paul Rosenberg" entity with its own URI linked to the one already in ULAN but described in more detail, thereby making this information available to the community.

All transfers of cultural objects have an economic dimension, whether prices or values are expressed explicitly or not. Some museums tend not to share prices

publicly. The AAM, however, recommends recording this information for objects transferred in Europe during National Socialism, as this may be relevant for assessing, for example, whether an artwork was a forced sale or not (particularly if the information includes details about currencies, discounts, and taxes).[56] Especially in the art market, transfers often go hand in hand with complex economic features such as joint ownership or bidding at auctions. Considering these details when structuring data lays the groundwork for financial analyses of provenances that would be meaningful for studies of the art market and economic trends, and potentially for identifying illegal acts (tax evasion, money laundering, black market deals).[57]

Paris, Texas is a film directed by Wim Wenders and, in a geographical sense, a city in the state of Texas named like the (current) capital of France. In a PLOD context, this presents a disambiguation challenge not unlike the one we encountered with parties. While this is often solvable through linking to shared vocabularies in the base layer, there are biases in the documentation of provenance research on this topic that must be addressed, for example with respect to under-recorded African towns and locations with complicated histories, such as Lviv in present-day Ukraine (but previously part of other countries and empires).[58] Geographic bricks can help carry that weight. Adding geospatial data is crucial for mapping and analysis, and data on administrative hierarchies and demographic aspects such as census and density can be meaningful for studying regional trends and patterns of collecting and taste.

Historical events such as Europe's colonial occupation of Africa, the Great Depression, and World War II, and the related German occupation of France affected transfers of ownership of artworks and are meaningful contextual information relevant to provenance records. At present, objects on museum websites are manually flagged as having changed ownership during National Socialism. By modeling the various territories occupied by the National Socialists between 1939 and 1945, the machine could precisely identify which objects changed ownership in this context and are therefore potentially relevant for potential claimants to this day. Other political, cultural, social, economic, and environmental situations of historical importance that have or might have influenced provenance events could also be meaningful to model for more extensive analysis—although, as we addressed earlier, modeling such historical events with hundreds if not thousands of subevents with parties, locations, and dates, though possible in this brick and its related layers, should ideally live in crowdsourced or authoritative vocabularies and be maintained by a larger community than simply cultural heritage institutions (see table 2.2).

56. American Alliance of Museums, "Recommended Procedures for Providing Information."

57. For an account of how provenance data can be made useful for social and economic historical analysis, see Rother, Mariani, and Koss (2023).

58. See Graham and Sabbata (2015).

Brick	Scope	Descriptive Range
Biographical	Parties	• Onomastic (e.g., title, first name, family name, maiden name) • Gender • Nationality • Religion • Biographical events (e.g., people's birth and death, group's formation and dissolution, change of location) • Relationships between people (e.g., family ties), between groups (e.g., a department of a museum), or between people and groups (e.g., a person's membership in a group)
Economic	Transfers	• Financial features (e.g., prices, currencies, sales discounts, profit margin, commission, sale tax) • Financial activities (e.g., payments, loans) • Auction activities (e.g., bidding, withdrawing) • Inventorying activities (e.g., insurance values, inventory numbers, depreciation) • Joint-ownerships
Geographic	Location, Provenance Event, Parties	• Geospatial features (e.g., geographical coordinates or shapes) • Administrative hierarchy (e.g., the country, region, province, and/or city of the location) • Historical Names • Demographic aspects (e.g., census, density)
Contextual	Provenance Event	Historical events of: • Social dimension (e.g., Olympic Games, World Fairs) • Political dimension (e.g., regimes, occupation) • Cultural dimension (e.g., epoches, fashions, ideology) • Economic dimension (e.g., depression, boom) • Environmental dimension (e.g., floods, fires, pandemics)

Table 2.2: Overview of the biographical, economic, geographic, and contextual bricks available in the PLOD conceptual framework.

History is rarely straightforward, and neither are our records of it—which, from a data perspective, is unfortunate. Historical knowledge in provenance is often incomplete ("unknown buyer," "private collection"), vague ("around," "circa"), uncertain to varying degrees ("probably," "possibly"), and subjective ("according to"). This dimension of knowledge is crucial for object documentation, interpretation, querying, and analysis of provenance. Hence, this incompleteness, vagueness, uncertainty, and subjectivity should be captured in PLOD, so it is available not only for humans but in a machine-readable way for use within and ideally across museum collections. Interpretative tools, as envisioned in our framework, should be applied at whatever scale they may prove useful in a brick-modeled provenance.

On the other hand, we understand the four bricks as options that museums can use based on their priorities according to their (provenance) mission and their (provenance) data. For example, a museum might begin the publication of PLOD by focusing on the core entities of the provenance and additional biographical information (base layer and biographical brick), leaving the description of economic details (economic brick) for a later stage—especially given that the Getty Provenance Index, currently transforming millions of auction and art dealer data to LOD, might soon provide these details in a structured and linkable format, ready to be employed by museums.

Indeed, our framework's modularity allows museums to address the inconsistency of their data. A move to PLOD will not fix this common issue, but it offers a chance to reframe it, and provenance along with it. In a field that has never been stable or constant, our framework aims to assist museums in making resource-conscious decisions so that perhaps their data strategies remain in line with their evolving missions.

2.7 Conclusion

Museums and their many internal and external stakeholders are faced today with innumerable and often competing demands, and they cannot address all such demands with the same care that they may warrant. Museums are also political actors because they tell stories about the past. These political pressures extend to museum data policy, including, but certainly not limited to, the question of provenance data: its production, long-term care, and accessibility. Through its web structure and commitment to openness, LOD can help begin to address concerns around these issues. The inevitable adoption of linked open-data standards in provenance is thus both a challenge and an opportunity. Its benefits outweigh the costs, for when done carefully and with a well-conceived data strategy, the move to PLOD can help museums pursue their goals of transparency, accountability, and inclusivity. It can also

help them address epistemic shifts and allow for a multi-perspectival but standardized and structured data practice. Finally, it is paramount that museums not only acknowledge but fully embrace the fact that recording and publishing provenance is a form of writing history. Whether museums rise to the challenge remains to be seen, but those that do will indeed write history.

3. Teaching Provenance to AI

Bibliographic Information
Mariani, Fabio, Lynn Rother, and Max Koss. 2023. "Teaching Provenance to AI: An Annotation Scheme for Museum Data." In *AI in Museums: Reflections, Perspectives and Applications*, edited by Sonja Thiel and Johannes Bernhardt, 163–172. Edition Museum. Bielefeld: transcript Verlag. https://doi.org/10.14361/9783839467107-014.
CRediT Roles: Conceptualization, Methodology, Writing – Original Draft, Visualization

3.1 Introduction

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. 3.1). And it is provenance records in museums that are particularly well suited to the application of computational methods such as artificial intelligence.

> 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.

Figure 3.1: 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).

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 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' [1], 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 [2].

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

1. Rother, Koss, and Mariani (2022).
2. Rother, Mariani, and Koss (2023).

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.

3.2 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 [3]. 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.

3. Yeide, Akinsha, and Walsh (2001); International Foundation for Art Research (IFAR) (2023).

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 [4]. 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 [5]. 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.

4. Rother, Mariani, and Koss (2023).
5. Finkel and Manning (2009).

3.3 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'.

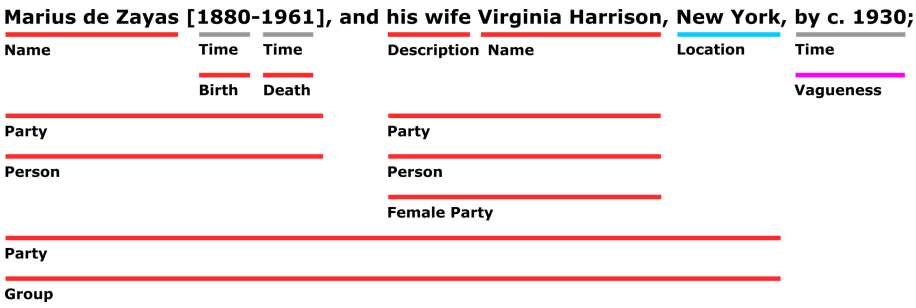

Figure 3.2: Conceptual example of span categorization applied to a provenance event extracted from the provenance text of Paul Cézanne's *Houses in Provence*: The Riaux Valley near L'Estaque. Source: National Gallery of Art (https://www.nga.gov/collection/art-object-page.54129.html, accessed in August 2023).

Figure 3.2 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'.

3.4 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.

4. Hidden Value: Provenance as a Source for Economic and Social History

Bibliographic Information
Rother, Lynn, Fabio Mariani, and Max Koss. 2023. "Hidden Value: Provenance as a Source for Economic and Social History." *Jahrbuch für Wirtschaftsgeschichte / Economic History Yearbook* 64 (1): 111–142. https://doi.org/10.1515/jbwg-2023-0005.
CRediT Roles: Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization

Abstract
Building on the extensive production of provenance data recently, this article explains how we can expand the purview of computational analysis in humanistic and social sciences by exploring how digital methods can be applied to provenances. Provenances document chains of events of ownership and socio-economic custody changes of artworks. They promise statistical and comparative insights into social and economic trends and networks. Such analyses, however, necessitate the transformation of provenances from their textual form into structured data. This article first explores some of the analytical avenues aggregate provenance data can offer for transdisciplinary historical research. It then explains in detail the use of deep learning to address natural language processing tasks for transforming provenance text into structured data, such as Sentence Boundary Detection and Span Categorization. To illustrate the potential of this pioneering approach, this article ends with two examples of preliminary analysis of structured provenance data.

4.1 Introduction

Since the end of the Cold War and the emergence of an international order focused on cooperation, cultural institutions have faced growing calls for transparency and accountability as to the origins of their collections. As a result, provenance research has benefited tremendously. And while the tradition of establishing ownership histories of artworks goes back centuries, born out of the wish to authenticate and establish the value of a work of art, the new context has brought with it new expec-

tations. In the past, it may have sufficed to provide a simple list of names without proof or documentation. Today, the standards with which the quality of provenances is gauged are different. They are scientific.

The production of provenance is generally carried out on behalf of commercial actors, such as auction houses (to do their due diligence), artist estates (to establish catalogues raisonnés), and museums – the single most important type of institution engaged in producing, publishing, and maintaining provenances. With the responsibility of identifying objects in their collections that were unlawfully appropriated during the National Socialist period or in other contexts of injustice, museums are reconstructing and publishing the provenances of their collections at an unprecedented rate and level of detail. Indeed, to the level of complicated transactions and varying valuations.

This vast and ever-growing corpus of increasingly detailed information on the ownership histories of artworks presents an untapped source for applying data-driven approaches. While such quantitative and statistical analyses have only recently been taken seriously by art historians, they have long belonged to the methodological toolkit of economic and social historians.[1] Considering provenances from both an art historical perspective and a social and economic one, as our paper does, is thus a logical proposition. This article presents what such an interdisciplinary historical perspective on aggregate provenance data may look like and what technological steps are required for its quantitative and statistical analysis.

Section one of this article provides a conceptual sketch of research avenues in economic and social history that – intersected with perspectives from art history – aggregate provenance data can open up and explore. Sections two and three describe how provenances, still essentially text-based sources of information, can be made machine-readable and enable historians to apply quantitative and statistical analysis methods. For this purpose, the article introduces deep learning models that address natural language processing tasks. The fourth and final section indicates how researchers can use structured provenance data for more complex analyses.

Why provenances? Provenances are ledgers containing the ownership and, to some extent, custody histories of artworks.[2] As Arjun Appadurai noted about the social lives of commodities:

> It is only through the analysis of these trajectories that we can interpret the human transactions and calculations that enliven things. Thus, even

1. As Diana Greenwald reminds us, some art historians looked at digital methods as early as the 1960s. See Greenwald (2021), pp. 1-2.

2. Provenances are usually defined as records of ownership changes. With auction houses prominently featured, for example, information on custody has always been present in provenances. Yet, such custody information remains partial and does not systematically include exhibition histories or other types of information on the geographical whereabouts of artworks.

> though from a *theoretical* point of view human actors encode things with significance, from a *methodological* point of view it is the things-in-motion that illuminate their human and social context.[3]

The potential analysis of provenances on a large scale thus allows us to address issues related to art ownership that are *social*, from the construction and reproduction of value to the distribution and reproduction of wealth. Such an analysis is possible because provenances record discrete economic decisions concerning art ownership by individuals or institutions. According to Appadurai, decisions made by individuals regarding an object become meaningful only in the context of an object's overall history. That is, in relation to the other decisions recorded in its given provenance.

In their most common form, provenances are lists of names, locations, and dates, identifying the consecutive owners or custodians of artworks, where and when they obtained them, and how they were transferred. To give an example, consider the provenance of Pablo Picasso's painting *Head of Young Boy* (1944), as published on the website of the Art Institute of Chicago, which owns the painting today:

> Sold by the artist to Paul Rosenberg & Co., New York, 1950 [invoice]; sold to Samuel Marx (1885–1964), Chicago, November 21, 1951 [invoice]; by descent to his wife, Florene May Schoenborn (1903–1995), New York and Chicago, from 1964; bequeathed by Florene M. Schoenborn to the Art Institute, 1997.[4]

We can see from this example that, beyond a mere index of ownership, provenances are matrixes of textual, historical information. By documenting a multiplicity of human activities relating to a specific class of goods, provenances are representations of complex social and economic relations and contexts.

The first line of provenance, culminating in a semicolon, indicates a single, direct purchase by the dealer Paul Rosenberg from the artist Pablo Picasso at a specific time in their careers and within a particular market. Artist-dealer relationships can take many forms, changing over the course of their careers and shifting according to wider trends in the circulation of art. Galleries successfully promoting living artists can influence market value and supply by enabling the artists to increase their output. The importance of such relationships for the marketing and valuation

3. Appadurai (1986), p. 5.

4. *Pablo Picasso*, Head of Young Boy, 1944, brush and black ink and grey wash, with scraping and touches of pen and black ink, on ivory wove paper, 500 × 288 mm, The Art Institute of Chicago, Chicago, Bequest of Florene May Schoenborn, 2012.569, https://www.artic.edu/artworks/158479/head-of-young-boy, 16.07.2022. The information contained in brackets in this provenance points to the existence of archival documents supporting the respective statement. As such, they are an idiosyncrasy of the museum's provenance style.

of certain artists, styles, media, and genres has been underlined in monographic studies.[5]

The above provenance also provides a sense of the geography traversed, not merely identifying but also locating the individuals involved. A large enough dataset of provenances can thus deliver insights into both the local and trans-regional, if not trans-continental, circulation of art, helping us to map economic phenomena in relation to cultural ones.[6] In this case, the life of the painting, and with it its provenance, began in the artist's studio in Paris, the world's cultural capital in the first half of the 20th century. The ownership of the painting then passed to Paul Rosenberg's gallery in New York, which took over Paris in importance in the second half of the 20th century. Subsequently, the work belonged to a collector in Chicago, underscoring the role of cultural accumulation in the industrial heartland of the United States.

The final line of provenance concerns a bequest made to a museum in 1997 by the artwork's last private owner, a widowed woman. When analysed against a large provenance dataset, such an event is telling for several reasons. Firstly, it raises the question of gender and the circulation of art – an issue highlighted in our paper. Secondly, it throws into sharp relief the role of donations and gifts, a particular form of exchanging value that can be investigated in and of itself. Lastly, and perhaps most importantly, it highlights not only the role of museums in the circulation of art but also in the creation and maintenance of the reputation of artists – another aspect addressed in our article. As it turns out, donating a work by Picasso to the Art Institute of Chicago in 1997 is much less consequential than it would have been decades earlier. This is because the artist's stature as a giant of modern art and the Art Institute of Chicago's global reputation as a major museum of significant modern art holdings were already firmly entrenched by 1997.

From a single provenance, then, we can already intuit how individual pieces of information can open up and are tied in with geo-temporal social networks and their historical dynamics and cycles. Aggregate provenance data and its analysis thus allow us, for the first time, to not only account for them quantitatively but also make them visible.

5. See for example, FitzGerald (1995).

6. For the potential of spatialisation of artistic phenomena based on numerical and statistical data and its analysis, see Joyeux-Prunel, Dossin, and Matei (2013).

4.2 Provenance as a Source for Economic and Social History

Based on the literature engaging with socio-economic questions related to art market mechanisms and behaviour, we can make out at least two areas of inquiry for which provenances provide unexplored source material: the construction of value and the dynamics of wealth (as embodied in cultural objects). When taking our cue from Appadurai and considering the commodity status of artworks, we are immediately confronted with a set of complications. In economic terms, i.e. in terms of market behaviour, artworks function according to a unique and nontransparent set of rules.[7] The art market and its construction of value are opaque. Despite an active research field tracing and studying the prices of artworks, explanations for how artworks obtain their economic value remain limited due to a lack of sufficiently detailed data on a large scale.

Indeed, purchasing art, a luxury commodity, is a form of conspicuous consumption and can be understood as a proxy for wealth.[8] Yet, it remains difficult to measure due to the relative discretion of the actors involved. The historical interest in the dynamics of wealth indicated by the concentration, circulation, and distribution of artworks is also an aspect of art market research that has been little explored. This is especially true on an international, cross-border level, not least because, for a long time, various national art markets were not integrated in the way they are in today's globalised world.[9] And while this set of problems in dealing with art market mechanisms and behaviour has been widely acknowledged, a long-term, historical perspective that can describe both the processes of value formation and the dynamics of wealth remains missing.

Economists and economic historians interested in the value of artworks have focused predominantly on investigating prices and their development over time, often with the view to establishing art's profitability as an investment. For this, they have long relied on Gerald Reitlinger's work around *The Economics of Taste*, whose first volume was published in 1961.[10] With his impressive but selective data, which covers mainly auction transactions that span over 200 years of art market activity, Reitlinger's insights are anecdotal at best. A generation of economic historians has nevertheless built on his work due to the lack of alternative data compilations. To

7. See various chapters in Towse and Navarrete Hernández (2020).

8. Mandel (2009), here: p. 1653: "As a luxury good, relative art demand is an increasing function of wealth." More recently, Kim Oosterlinck has studied art as wartime investment and has proposed that certain types of art may in exceptional circumstances also be purchased for reasons to hide wealth, indicating that motives for purchasing art over time may vary. See Oosterlinck (2016).

9. Challis (2021), pp. 6-7.

10. Reitlinger (1961); Reitlinger (1963); Reitlinger (1970).

date, auction data has been the primary source for analysing the art market. This means that research has been limited to a fraction of the activities in which artworks change hands.

Despite the lack of large representative datasets, scholars use price indices to understand historical developments of the art market.[11] Studying repeat sales, for example, they track the price of one artwork at different times, locations, and with different people involved. Repeat sales indices acknowledge the fact that artworks are unique, non-interchangeable goods. The underlying assumption is that an artwork remains unchanged over time. The history of the art market has proven otherwise, however. The condition of artworks can change significantly, and even small changes can be reflected in the price, although we do not know exactly to what extent. A further limitation of such a method is that only artworks bought more than once can be considered. Expensive works that only change hands once in their lifetime, such as Rembrandt's *Nightwatch*, cannot be included due to their short provenance. This means that repeat sales indices are even less representative than other methodologies of analysing prices to understand value construction in the art market.

In contrast, the more commonly used hedonic regression can rely on more data points in any given price dataset. Here scholars have increasingly tried to identify the deciding factors in price formation. Characteristics used in hedonic regressions include object criteria (e.g. artist, medium, signature, and dimensions) and other market factors such as the seller, location, and time of the sale. Limitations of this method arise from the frequent use of dummy variables, which flatten historical complexity. Despite the sophistication of recent studies, hedonic regression also runs the risk of focusing on factors that are irrelevant or else ignoring aspects that could be important, such as the social and cultural aspects of value formation.

Social scientists have taken a different approach to the value formation of artworks. Rather than focusing solely on valuation outcomes, such as prices, these authors emphasise the importance of investigating the processes that lead to them.[12] With his in-depth study of galleries in the Amsterdam and New York contemporary art market in the late 20$^{\text{th}}$ century, the sociologist Olav Velthuis has shown, mainly through interviews, that value is a function of social context. He writes:

> The value of an artwork does not reside in the work itself, but is, under conditions of uncertainty, produced and constantly reproduced by artists, intermediaries, and audiences, subject to numerous conventions and codes of art worlds.[13]

11. For recent studies on prices, see e.g. Crotta and Vermeylen (2020); Oosterlinck and Radermecker (2019); Oosterlinck (2016); Renneboog and Spaenjers (2013).

12. Dekker (2015), p. 312.

13. Velthuis (2005), p. 160.

A similar cast of actors in value construction was identified by Samuel Fraiberger et al. in their data-driven analysis of 497,796 exhibitions in 16,002 galleries, 289,777 exhibitions in 7,568 museums, and 127,208 auctions in 1,239 auction houses spanning 143 countries and 36 years (1980 to 2016).[14] One of the key findings of their study showed how artists who were exhibited early on in their careers in specific galleries and museums were more successful. These galleries and museums were usually located in Western art market centres and had the highest cultural and/or economic cachet. Success was measured here by higher auction prices later in life and a lower risk of dropping out of the art market altogether. The study focused on network analysis, investigating the interdependence between actors such as galleries and museums, their role in the exhibition circuit, and the effects on artists' careers and their auction prices. It demonstrated how the decision of specific galleries and museums on which artists to include in their exhibitions directly impacts their future success and the valuation of their work. These particular actors can therefore be considered gatekeepers since they influence the value of artists' work like few others do.

By not studying prices but looking at art market activity across time and space, Schich et al. have recently identified network dimensions in their important study in the field of cultural data analytics.[15] Using auction house data from four European markets (England, France, Belgium, and the Netherlands) between 1801 and 1820, they analysed 267,000 market transactions involving around 22,000 actors.[16] Despite their dataset's narrow temporal and geographical focus, they have revealed insights into social, temporal, spatial, and conceptual network dimensions. They have shown how auction activity was spread over four countries, with markets peaking at different times, measured by the number of transactions. Indeed, the European auction market cannot be fully understood by studying a single region since it is only via a comparison of four countries that it becomes evident that most protagonists stay within one region, where they stick to exclusively buying or selling. Equally, Schich et al. have described the behaviour of market participants, demonstrating the relative importance of specific dealers and the buying and selling activities of collectors. They observe: "Top sellers tend to appear on the market exactly once and top buyers tend to dominate the market for two to three years."[17]

Their study has also shown links between art market clusters and highlighted the existence of brokers between communities who both buy and sell. Looking at the lo-

14. Fraiberger et al. (2018).

15. Schich et al. (2017).

16. This data was taken from the Getty Provenance Index, the world's largest single repository of art market data. Their limited focus was due to the uneven quality of the data available, which forced them to work with a relatively complete, clean, and standardized subset.

17. Schich et al. (2017), p. 6.

cation of sales, Schich et al.'s results have confirmed the known ranking dynamics of cultural centres in Europe; Paris is essentially synonymous with the French market (with only a few side locations, which never see auctions in consecutive years); London and Amsterdam are predominant in their countries; Belgium is a market with multicentric competition. Conceptually, their study has looked at the nationality of the artist attribution of objects recorded in auctions and related them to sales in the artist's respective countries. They found that top artist attributions function similarly to a "product category" in a supermarket. On this basis, they tried to identify sense-making time-frames of shopping cycles by looking at all transactions by a buyer in a given week, year, or in total.

This led them to conclude that a large portion of the European auction market was "a system that functions like a super-slow supermarket on an annual grocery-cycle." Schich et al. nevertheless admit that their conclusions are made "on thin ice" since it is unclear, for example, whether the protagonists involved believed the attribution of a given artist was true or not.[18]

What Schich et al. have in common with the many studies based on auction data analysis is that they rely on vague catalogue descriptions. Auction cataloguing, whether analogue or digital, is a marketing tool that may or may not camouflage uncertain attributions of artworks. This means the significance of these studies for analysing art market behaviour is circumscribed, in particular, by the absence of actual artworks. How can we draw conclusions from the analysis of discrete economic decisions of buying and selling unique goods when the commodities are reduced to a set of very few characteristics? Equally unreliable is the public reporting on the results of auction sales. The common practice of auction houses and consignors, especially in times of crisis, to unofficially guarantee prices, withdraw lots, and buy back art stock, makes publicly available auction results a poor source for studying prices or understanding buying and selling behaviour in the art market.[19] Furthermore, buyers, such as museums and collectors, usually enlist the help of dealers and other agents to bid at auction while they remain anonymous.[20]

Here is where provenance data from museums or catalogues raisonnés differs fundamentally. The information contained in provenances has been studied on a

18. Schich et al. (2017), p. 8.

19. See, e.g. Ashenfelter and Graddy (2020), pp. 19-21; for auctions during the Great Depression, e.g. the sales of the so-called Collection of a Swiss Nobleman, Albert Figdor, or Marczell von Nemes, see Rother (2017), pp. 26-51.

20. See, e.g. Zalewski (2019), p. 101.

micro-level, enhanced, and vetted by art historical expertise and knowledge.[21] What is more, provenance data includes not only auction sales, which represent a fraction of art market activities, but also a wide variety of information on how works of art change hands, especially in non-market circumstances, such as intra-familiar ownership changes, private sales, and donations to museums. In fact, from our experience of working with provenances, we have identified at least five relatively frequent types of activity related to objects owned by an individual. The first is the sale of individual works of art at auction or to a dealer.[22] The second is the exchange of artworks directly with a dealer or with another collector through the intermediary help of a dealer or gallery. The third is the liquidation of an entire collection by way of auction. The fourth is the passing on of artworks through inheritance, which may or may not be followed by the liquidation of the inheritance by the receiving party for various reasons – not least to do with tax. The fifth activity, which frequently appears in provenances, is donations to museums and institutions.

Since provenances are chains of socio-economic events – involving actors such as museums, galleries, and auction houses and describing their role in a given transaction, the location and time of the event, and, increasingly, prices – they are destined for complex network analysis. Given that their data points come from multiple potentially related events, they allow us to exploit the advantages of the repeat sales method and hedonic regression analysis. Bringing such queries together with specific object information of unique artworks (such as artist, title, subject matter, genre, style, medium, size, and date) makes aggregate provenance data an inevitable source for analysing the construction of value. This is all the more so since both the economic and social sciences have acknowledged that object *and* social factors impact the construction of value in the art market.

With the potential of analysing interdependencies of an unprecedented wide spectrum of socio-economic activities in relation to the objects, aggregate provenance data will enable us to specify the role of institutions or people and to see patterns

21. We are cognisant of the fact that provenance records also have an inherent bias. They are only produced for works considered worthwhile spending resources for research and documentation. However, as constituents such as museums, authors of catalogues raisonnés, auction houses, collectors, and artists write provenances, the quality of the data is better and more representative for the works circulating in the art market than any other data containing such interactions and transactions. For computational analysis, object and provenance data would ideally be combined with exhibition data, which has been used more frequently and is easier to structure, and has also an impact on the construction of value and the concentration, circulation, and distribution of artworks. Art criticism data has so far not been systematically recorded digitally, but would equally enhance such analysis.

22. Many transfers recorded in provenances are complex transactions, involving not only a single sender and a single receiver of a work, but possibly two or more senders or receivers, and also agents enlisted by senders or receivers.

and outliers in the concentration, circulation, and distribution of artworks. Indeed, network analysis can account for a whole host of historical factors that have an impact on the frequency and outcome of socio-economic activities such as inheritance, sales, or donations: from wars and economic crises to changes in import and export or tax regulations. The impact of gatekeepers for specific artists, subject matters, genres, styles, or media can also be analysed from social, temporal, spatial, and conceptual perspectives. To show their importance, or indeed insignificance, would not only be possible within their period of activity and geographic reach but also in comparison to their peers and across historical periods. Identifying such complex network dimensions for the understanding of art markets and collecting practices requires machine-readable data of the histories of objects, however. Since provenance data is already digitally available on a large scale thanks to the work of museums and, increasingly, catalogues raisonnés, its structuring ushers in a new era for art market studies.

4.3 From Text to Data: Structuring Provenance

The various paths of analysing aggregate provenance data that we have laid out are based on the different types of information stored in provenance texts. To address a seemingly simple inquiry, however – such as a comparative understanding of how men and women may differ in their engagement with inherited artworks – we first need to extract the historical knowledge contained in provenance texts. In the following sections, we illustrate how deep learning is the digital method most suited to this challenge, given the syntactic and semantic peculiarities of provenance texts and the variety of information contained in them. First, we discuss which natural language processing tasks are most appropriate for extracting information from provenance texts and why. Then we describe how we implemented deep learning models to address the two most promising tasks concerning provenances: Sentence Boundary Detection and Span Categorization.

Although the study of both provenance and the art market has an established historiographical tradition, the use of computational methods for analysing large-scale phenomena in this field is still in its early and experimental stages. This is because provenances have thus far been recorded and published as texts understandable only by humans and not by machines. Many provenance repositories, like museums, adhere to guidelines on recording provenance, such as those published by the American Alliance of Museums (AAM) or guidelines loosely based on them.[23] According to the AAM guidelines, provenance texts are a succession of sentences, where each sentence stands for a provenance event. These provenance events are

23. Yeide, Akinsha, and Walsh (2001).

listed in chronological order, with the first event ideally identifying the creation of the object or the moment of its archaeological discovery. According to the guidelines, an event contains information about the participants involved, the method of transfer, the date, and the location. Events are divided by punctuation marks, which carry significant meanings. The guidelines determine that if two events are divided by a semicolon, there was a direct transfer between the two owners. However, if two events are separated by a period, there was no direct transfer between the two owners, and a gap in the historical reconstruction of events can be inferred. Parentheses can provide additional knowledge, such as biographical information. Further information about events, such as historical sources, can be provided in footnotes. It should be noted that, aside from this guidance on the form and structure of provenance events, the AAM guidelines leave much room for interpretation when museums come to write their provenances. This means that there may be differences between the provenance texts of different museums that conform to the AAM guidelines – an issue we will address in the next section.

To analyse provenance events on a large scale, we need to extract information from provenance texts with the help of digital methods. Automatic information extraction is one of the purposes of Natural Language Processing (NLP). NLP is an area of research at the intersection of computer science, linguistics, and, more recently, artificial intelligence. It investigates tasks that can be performed by digital methods to automatically process and analyse natural human language on a large scale. Specifically, as in the case of provenance, the NLP problem of extracting event information from a written text is defined as event extraction, for which several NLP tasks are available – something we come to discuss.[24] An event can be understood as an entity represented in space and time in which the parties involved take actions that result in a change.[25] Where provenances are concerned, the change caused by an event affects the ownership of a work of art.

We must take four event elements into account to perform event extraction.[26] The first element is the *event mention*, the text portion containing a potential event, which usually forms a sentence. When extracting events from provenance, we already have the advantage that punctuation marks delimit each event according to the AAM guidelines. Therefore, each sentence is considered an event. The second element is the *event trigger*, that is, the element explicitly indicating that an event is taking place. The most common *event trigger* is a verb. This is also true for provenances since they contain verbs that indicate various methods of transfer, such as "sold", "bequeathed", "exchanged", or "purchased". We may also encounter nominal

24. For event extraction from historical texts, see: Cybulska and Vossen (2011); Sprugnoli and Tonelli (2019); Lai et al. (2021).

25. Xiang and Wang (2019).

26. Xiang and Wang (2019); Linguistic Data Consortium (2005).

sentences using expressions such as "by descent" or "in exchange". A peculiarity of provenance texts is that we may encounter no *event trigger* at all for a particular sentence. The text gives nothing but the object's owner in such a case. Nevertheless, we can still consider it a provenance event since we can extract information about the owner who received the object. The third element of event extraction is the *event argument*, which is to say, the different components of an event. For example, a provenance event, in addition to the *event trigger*, may also have a date, location, and involved parties as *arguments*. It is also quite common to find other information in a provenance event referring to biographical subevents of the parties involved, such as their birth, death, or movement to another location.

The *event argument* role is the last element to consider in event extraction. Where participants are concerned, we have to determine who gives the item (sender) and who receives the item (receiver). But then again, a participant can also take the role of the intermediary (agent). It is, moreover, crucial to distinguish whether a place or date in a provenance is an *argument* referring to the main event or a biographical sub-event, such as birth.

In light of these four elements, we have divided the event extraction problem into two sub-problems. The first of these problems concerns splitting the provenance text into discrete provenance events. Such an approach allows us to preliminarily isolate the different *event mentions*. From these mentions, we can, in turn, extract the *event trigger*, *event arguments*, and *event argument roles*. To split a provenance text into discrete events, we can apply a Sentence Boundary Detection (or Disambiguation, SBD) task. This task involves recognising which characters start and end a sentence by disambiguating punctuation. We are reminded that, according to the AAM guidelines, provenance events are divided by semicolons or, if followed by an unknown event, by periods. More established digital methods to address SBD tasks include rule-based techniques, such as decision trees or regular expressions.[27] Although more straightforward to implement, the disadvantage of these techniques lies in their failure to disambiguate punctuation marks in some contexts. Consider, for example, the punctuation marks used for abbreviations, which are easily confused with periods and can coincide, in rare instances, with the end of sentences. More sophisticated methods, such as deep learning, allow us to address these ambiguities.

Deep learning models have recently addressed NLP tasks with impressive results.[28] Such digital methods, inspired by the structure of the human brain, use neural networks to extract features from digital sources, such as images or, in our case, texts. Neural networks are hierarchically organised and interconnected layers of mathematical functions, i.e. artificial neurons. From digital sources, they can

27. Riley (1989); Aberdeen et al. (1995).
28. Alzubaidi et al. (2021).

identify and extract different characteristic features through a process of abstraction, from one neural network layer to the next.[29] This provides neural network models more flexibility than an algorithm-oriented approach, which is centred on a specific goal with certain predetermined parameters.

Extracting features means finding the correct representation of raw data.[30] For example, in the case of text, the representation is generated by extracting semantic and syntactic features from tokens, i.e. smaller textual units such as words. To provide a representation of a word, it is necessary to analyse the different contexts in which a word appears and infer patterns. As Firth summarises: "You shall know a word by the company it keeps."[31] A neural network can infer patterns from a context through a process of multiplied, layered abstraction. However, such a network needs sufficient examples to effectively generalise a token's features across different contexts. Therefore, we need to train the model on a set of examples before we can use it on a larger scale. In addition, the examples provided for training should be annotated according to the desired output so as to calibrate the model to perform specific tasks. For instance, in the case of SBD, we need to train the model with texts in which sentence boundaries are annotated.

The effectiveness of deep learning in performing SBD has stimulated both the creation of cross-domain models (trained on large corpora of generic and consistent texts such as newspaper articles) and domain-specific models (where the text has a peculiar structure, as in the clinical and legal domains).[32] Performing an SBD task on provenance texts compiled according to AAM guidelines requires us to train a domain-specific model from scratch. This is because provenance texts do not always follow a logical and linguistic structure comparable to standard texts. As previously mentioned, sometimes they do not even have a verb present. Finally, according to a provenance-specific SBD model, sentences can be divided not only by periods but also by semicolons.

The second sub-problem of event extraction from provenances concerns identifying the *event trigger*, *event arguments*, and *event argument roles* for each provenance event. We can solve these three issues with three specific tasks. For example, we can identify the *event trigger* by performing a Part-of-Speech tagging (PoS tagging) task. This task deals with recognising each word's grammatical role in the text. For example, PoS tagging involves identifying which words are verbs, the primary *event triggers*. *Event arguments* are named entities, such as people, organisations, places, and temporal expressions. We can therefore use a Named Entity Recognition (NER) task to recognise and categorise such entities. Finally, we can perform a Relation

29. Zhang et al. (2018).
30. LeCun, Bengio, and Hinton (2015).
31. Firth (1957).
32. Griffis et al. (2016); Sanchez (2019).

Extraction (RE) task to identify event argument roles.[33] As the name suggests, relation extraction involves identifying semantic relationships within the text. In this way, we can represent the *event argument roles* in relation to the *arguments* and the *event trigger*.

PoS tagging, NER, and RE are NLP tasks that deep learning models can successfully perform. However, as with SBD, we also need to train domain-specific models for these tasks, annotating training examples for each. In addition to the SBD model, we would need to train three new models from scratch to extract provenance events. The effort involved in such an approach, however, made us consider alternative options and seek out a more efficient process that could potentially address all three tasks at the same time.

We thus experimented with identifying the *event triggers*, *event arguments*, and *event argument roles* using a Span Categorization (or Classification) task.[34] This task is similar to that of NER. Unlike NER, however, Span Categorization does not focus on individual tokens but spans, i.e. portions of text. Span Categorization involves recognising and classifying spans, whether they are named or unnamed entities. In addition, unlike NER, multiple spans can be recognised from the same text portion. This characteristic allows us to create a hierarchy of spans, possibly assigning more than one category to the same text portion. In this way, we can assign an *event argument role* as an additional category to a span recognised as an *event argument*.

Figure 4.1 shows how, conceptually, Span Categorization can be applied to a provenance event.[35] First, we identified the *event trigger*. In this example, the *event trigger* was the term "by descent," classified as *method* (method of transfer). Then we identified the different *event arguments*. While we tagged the span "his wife, Florene May Schoenborn (1903-1995), New York and Chicago" as party, "1964" indicates the *time* of the event. The advantage of Span Categorization is the ability to assign additional classifications to a party's span. Firstly, we identified the role: the party is the one who received the object (*receiver*). In addition, we assigned the type of party, whether it be a *person* or a *group*. Finally, given our research interest in the role of women in the history of collecting, we also classified the participant as a *female party*. Through this last categorisation, we performed another NLP task by addressing the issue of gender classification.[36]

33. Zhou et al. (2005).

34. Xu, Jiang, and Watcharawittayakul (2017); Sohrab and Miwa (2018); Tan et al. (2020).

35. The event is excerpted from the aforementioned provenance of Pablo Picasso, Head of Young Boy, as published on the website of The Art Institute of Chicago (https://www.artic.edu/artworks/158479/head-of-young-boy, 16.07.2022).

36. Hu et al. (2021). For the feasibility of the NLP task and given the inconsistent nature of the underlying historical documentation, we have opted to apply a binary gender classification. We acknowledge that gender identity is more diverse.

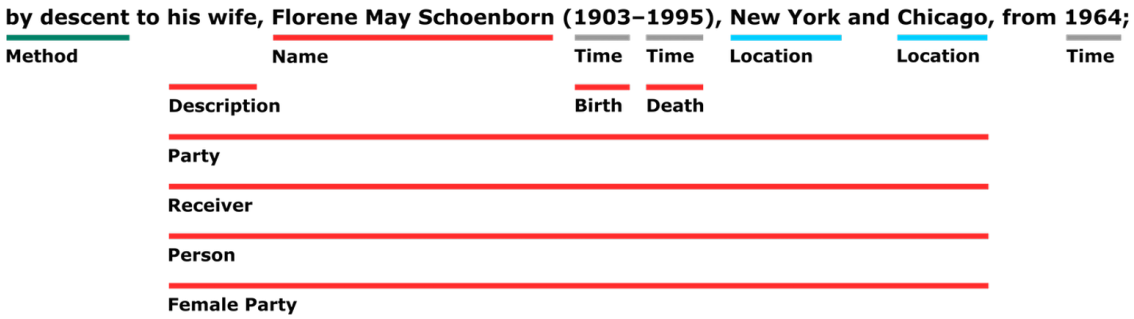

Figure 4.1: Conceptual Example of Span Categorization Applied to a Provenance Event. Source: based on the provenance of Pablo Picasso's Head of Young Boy, as published on the website of the Art Institute of Chicago (https://www.artic.edu/artworks/158479/head-of-young-boy,16.07.2022); created by Fabio Mariani.

The ability to overlap multiple spans also allows us to assign categories to smaller spans within a portion of text that we have already classified. Returning to the example above, we can classify "Florene May Schoenborn" as the party's *name*. We can also distinguish a *description* of the party, i.e. the span "his wife". After extracting the data, this allows us to trace the relationship of the party to the previous owner, from whom they inherited the object. Moreover, the text provides the party's biographical dates of "1903" and "1995", which are identifiable as two spans belonging to the *time* category, the first of which can be tagged as *birth* and the second as *death*. Finally, the text portions "New York" and "Chicago" refer to the party's locations, which are both classifiable as *location*.

For all the advantages that Span Categorization offers, however, there are, of course, disadvantages. Compared to NER, Span Categorization is more computationally expensive since it does not operate on individual words but rather portions of text. When assigning categories, it, therefore, has more candidates to consider, which may or may not be overlapping. In a NER scenario, the potential candidates are equal to the number of words in the text. In a Span Categorization case, the number of candidates is

$$n \cdot \frac{n+1}{2}$$

where n is the number of words in the text.[37] Furthermore, increasing complexity makes the task of establishing an entity's boundaries more difficult. For example, the span "Paul Rosenberg & Co." refers to an organisation and not a person. And yet this situation may prove ambiguous for a Span Categorization model, which might only recognise "Paul Rosenberg" as a person. Finally, even in the case of

37. Tan et al. (2020).

Span Categorization, we must still train a domain-specific deep learning model from scratch.

Despite these issues, Span Categorization, combined with SBD, nevertheless allows us to deal successfully with event extraction by having to train only two models. In the next section of this article, we illustrate how we experimented with event extraction in provenance texts by training SBD and Span Categorization models on a museum dataset.

4.4 Training SBD and Span Categorization

The Art Institute of Chicago was founded in 1879 and is a prominent museum with strong collections across multiple departments ranging from the Ancient Americas to the Arts of Africa and Asia. It is particularly known for its exceptional holdings of European Modern art, extending to Contemporary global art.

It is, however, one of the few museums that have made its collection dataset available to the public, including provenance texts – be it online via the museum website or as a download.[38] That is why we chose the Art Institute for our experiment, hoping that the results will only inspire more museums to make their collection data available to the public. The version of the dataset we used in the experiment was downloaded on April 7th, 2022, and contained data for 122,317 objects, varying in medium, culture, and period. For the experiment, we focused on those objects with provenance texts, which counted 11,504 objects (9.4 percent of the dataset). Generally, the Art Institute's provenances follow the AAM guidelines, with the peculiarity that notes for each provenance event are given in square brackets within the text, as opposed to in the footnotes.

Before training the SBD and Span Categorization models, we pre-processed the data. This step is crucial in standardising texts, helping to clean up any human errors or stylistic peculiarities. Firstly, the spaces between words were standardised. We replaced inconsistent spacing caused by tabulation and removed multiple spaces, including those found before periods, semicolons, or slashes. Parentheses and quotation marks were also made uniform. Furthermore, where notes were found in curly brackets, they were replaced with square brackets. As for quotation marks, we replaced any curly quotation mark with a straight one. Different dashes used for hyphenation were also standardised as the hyphen-minus character. And finally, all HTML tags were removed. After this cleaning process, we discovered that 112

38. The Art Institute's Github page (https://github.com/art-institute-of-chicago/api-data, 08.07.2022) provides examples of the data as well as the external link to download the entire dataset. Data from a single object is also available via API (https://www.artic.edu/open-access/public-api, 08.07.2022), however for using the whole dataset it is recommended to download the dataset.

problematic provenance texts (0.97 percent of the texts) contained typos and errors, such as open/unclosed parentheses or not conforming to AAM guidelines. Since correcting such typos and errors requires human intervention, we decided to discard them altogether. The dataset we experimented on thus consisted of 11,392 objects with provenance texts, as opposed to the 11,504 we started with.

Once the provenance texts had been pre-processed, we were able to proceed with model training. For both models, we used the open-source library *Spacy*, written in Python.[39] *Spacy* offers a set of generic models trained to perform various NLP tasks, including those already discussed: SBD, PoS tagging, and NER.

In addition, the library offers users the opportunity to configure new neural network models and train them for custom tasks.

From the dataset, we randomly selected 6,000 objects whose provenance texts we then annotated to train a domain-specific SBD. For each text, we tagged the boundaries delimiting the provenance events.[40] We then divided the 6,000 annotated texts into three groups: the training set, the validation set, and the test set. We used 3,600 texts (60 percent of the corpus) in the training set. This set contained the examples with which we trained the model to predict the boundaries of a provenance event. Since training is an iterative process, finding the right number of iterations is essential, i.e. the times the model is updated. Too few iterations can cause what is known as underfitting (poor learning), while too many iterations can cause overfitting (poor generalisation ability).[41] For this reason, we used 1,200 texts (20 percent) for the validation set to control the model's performance during training. This so-called early stopping technique adjusts the number of iterations and stops the learning process before overfitting occurs.[42] Which is to say, before the model no longer improves in testing against the validation set after a predetermined number of iterations (in our case, 1,600 iterations). Our SBD training continued for 3,000 iterations before early stopping interrupted the process to prevent overfitting. Finally, we used the remaining 1,200 texts (20 percent) for the test set, which helped us to evaluate the quality of the model's predictions after training.

The model reached peak performance after 1,400 iterations, with an F1 Score of 0.99 recorded on the validation set. The F1 Score is a measure by which a model's training is evaluated. It is given by the following formula:

$$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

39. ExplosionAI, Spacy 3.2.4 (https://spacy.io/, 18.07.2022). We conducted the experiment on a MacBook Pro (2019, 2.6 GHz 6-Core Intel Core i7), using Python 3.8.

40. We annotated the texts using Doccano (https://doccano.github.io/doccano/, 20.09.2022). Doccano is an open source data labelling tool.

41. Ying (2019).

42. Prechelt (2012).

Precision is a score that measures the quality of the model's predictions, that is, the ratio of true positives (tp) to false positives (fp), which is calculated as:

$$\frac{tp}{tp + fp}$$

Recall rewards the number of predictions and is inversely proportional to the number of missed predictions, i.e. false negatives (fn):

$$\frac{tp}{tp + fn}$$

The F1 Score is the harmonic mean of these two fundamental measures. The maximum value for F1 Score, Precision, and Recall is 1.

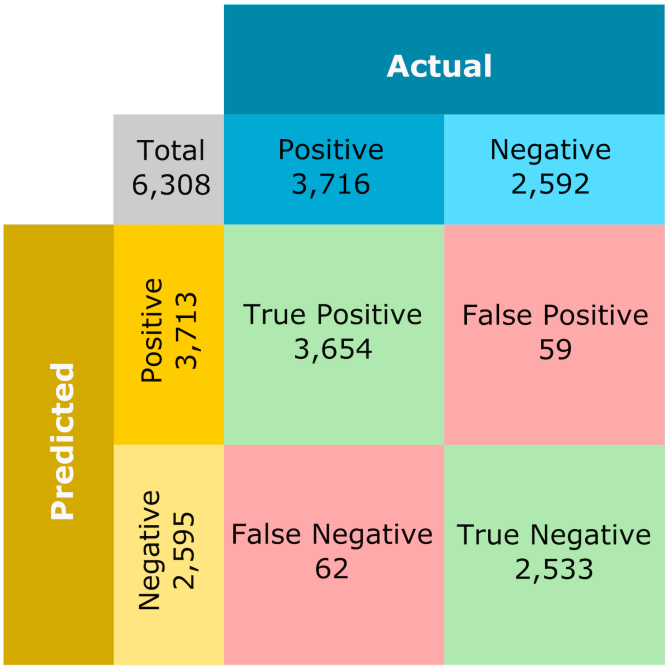

Figure 4.2: Confusion Matrix Showing the Model's Results of the Test Set. Source: 1,200 provenance texts from the Art Institute of Chicago, selected for testing an SBD model; created by Fabio Mariani.

The performance of the test set gave the quality assessment of the model, as summarised in figure 4.2. The model predicted the boundaries of 3,713 provenance events (predicted positives), 3,654 of which were correct (true positives), which means that 59 times the prediction turned out to be incorrect. We can speak of false positives in this case since the model incorrectly divided the text into provenance events. Indeed, the model failed to recognise 62 provenance events, otherwise known as false negatives. Finally, the model disambiguated 2,533 characters, which, although they take the form of a period or semicolon, do not delimit a provenance

event (true negatives). Calculating Precision, Recall, and the F1 Score, the model registered 0.98 in all three measures on the test set.

Despite such an excellent result, we noticed while analysing the errors that provenance texts with notes inside square brackets (as practised by the Art Institute) created difficulties for the model. Sometimes these notes were several sentences long and created ambiguous scenarios for the SBD model. We, therefore, strengthened the model by adding an auxiliary rule which stipulated that any event boundary recognised inside brackets should not be considered. We then implemented an algorithm that corrected these predictions during the workflow of the model. Such an approach creates a hybrid model that benefits from the generalisation of deep learning and the definition of strict rules. The results of the test set for the hybrid model confirmed the efficacy of the auxiliary rule. Total predictions dropped to 3,695, compared to the previous count of 3,713. Of these predictions, those that were correct rose to 3,672, whereas they previously counted 3,654. Furthermore, false positives dropped to 23, less than half of the 59 cases mentioned above. This, in turn, led to the identification of more correct provenance events, reducing false negatives to 44. In light of these refinements, the Precision, Recall, and F1 Score values rose to 0.99. That is all to say that while the model had already achieved a high quality of efficacy through deep learning, we increased the result ever so slightly through hybridisation. This intervention, moreover, ensures that even better results are reached when applying SBD to the whole dataset.

Having completed the training on the SBD model, we turned our attention to the Span Categorization model. Again, the first step was to select and annotate data for training, validation, and testing. Since Span Categorization applies not to the entire provenance text but to each provenance event, we used the new SBD model to divide all provenance texts into events. From the 11,392 provenance texts in the Art Institute's dataset, we extracted 35,554 provenance events.

As discussed in the previous section, one of the problems of Span Categorization is the question of complexity, which is proportional to the number of words in any given text. As we have already seen with SBD training, the peculiarity of the Art Institute's provenance events, with their long notes in square brackets, creates ambiguity. For this reason, we pre-processed the texts of provenance events by extracting the notes in square brackets, transforming them into footnotes, and then replacing the missing text with their respective number in square brackets. We performed this intervention on all notes that were either longer than 20 characters or contained numbers (which usually refer to bibliographic citations).

Since Span Categorization is more complex and text-dependent than SBD, we tried to select training, test, and validation data that were as representative as possible of the entire dataset. For this reason, we clustered the provenance events

according to text similarity by calculating cosine similarity among the different texts.[43] Cosine similarity is a measure to calculate the similarity of two vectors and can range from 0 (totally different) to 1 (identical). In our case, we compared two provenance events as two vectors having each n dimensions, where n is the sum of the total words of the two texts. For each dimension, we assigned a value of 1 to it if the respective word was present in the text represented by the vector or 0 if not. Once both provenance texts were represented as vectors, we calculated their cosine similarity. This calculation allowed us to collect event texts with a cosine similarity of 0.5 or greater into 6,531 clusters. For each cluster, we randomly selected one event, creating an annotation corpus of 6,531 items.

We tagged 17 span categories in the provenance events of the annotation corpus. First, we annotated the *method*, *time*, *location*, and *party*. As discussed in the previous section, in Span Categorization, it is possible to have multiple categories for a span or its sub-portions. For example, each span annotated as *party* was also categorised by type (*group* or *person*), role (*sender*, *receiver*, or *agent*), and gender (*female party*). Within each party, we also tagged the span representing biographical information, such as *name* (even more than one, e.g. maiden name), *birth* and *death*, with the associated *time* and *location* span. In addition, we annotated any life *location* within the *party* span. Finally, through the *description* tag, we annotated portions of text providing additional information about the *party*, such as their profession or relationship with other parties. Where relationships were concerned, it was also possible to annotate any parties mentioned within the description. We maintained a similar approach for groups of people acting together (e.g. couples), tagging them as a group-type party containing multiple person-type parties. For some events, especially auctions, where the inventory or lot number is often given, we annotated with the *inventory* tag. Given the possibility of overlapping annotations, we used a *vagueness* tag to categorise approximate *time* or *location* spans (e.g. "circa 1945" or "near Paris") and an *incompleteness* tag for *party* spans whose information was incomplete (e.g. "unknown collector"). During the annotation stage, 80 provenance events were discarded as SBD errors (1.2 percent of the annotation corpus).

Training of the Span Categorization model was carried out along similar lines to that of the SBD model, dividing the annotation corpus into three datasets: the training set (3,871 texts making up 60 percent), the validation set, and the test set (each containing 1,290 texts amounting to 20 percent). Training continued for 5,800 iterations, reaching peak performance on the validation set after the 4,200th iteration, with an F1 Score of 0.95. By evaluating the model on the test set, we obtained Precision and Recall results of 0.95 and 0.93, respectively, and an F1 Score

43. Gomaa and Fahmy (2013).

of 0.94. While these results amount to overall good performance, they are lower than the SBD model's, which performed a more straightforward task. For instance, we can see that the Span Categorization model has a lower Recall result than it does Precision, which means that it fails to identify categories for some spans. Still, when it does, it is fairly accurate. This shows that the model did not encounter any difficulties when handling entity boundaries, however, which we previously identified as one of the drawbacks in dealing with the Span Categorization task. If it had encountered such difficulties, we would have found ourselves with a model with a Precision result lower than its Recall result, which would have meant that it was more adept at identifying spans but more prone to errors.

Looking back at the excellent results of the experiment with the Art Institute's data, we are confident that the method can be applied on a larger scale, training models using provenance texts from several institutions. By using deep learning to address the two NLP tasks of SBD and Span Categorization, we are able to identify discrete provenance events and the various elements presented in them with reliable accuracy. Moreover, Span Categorization offers the possibility to go beyond event extraction, as it allows us to extract layered and complex information on individual provenance elements, especially those involving parties. With the help of Span Categorization, researchers analysing data can query more complex phenomena, which depend on individuals occupying multiple roles at once. For example, they will be able to identify a person of a certain age, gender, and relation to another person or group, not to mention acting in a certain place, time, and role.

4.5 Preliminary Analysis

On the basis of the publicly available collection data from the Art Institute of Chicago that we structured through our experiment, we were able to undertake preliminary data analysis. This analysis allowed us to identify not only certain dimensions of the historic acquisition patterns of the museum but also gender differences regarding the question of how individuals that inherit artworks engage with them subsequently.

The collections of museums are not only their most important source for attracting and educating visitors, but they are also a signifier of their wealth, which is ultimately the cumulative result of a string of economic decisions regarding collecting strategies that can be traced from their founding years to the present. Their standing and recognition rely on the capacity to acquire works of art, either by spending money (or offering other works) or by attracting donors. Across aggregate provenance data, it is possible to discern whether or not the decision of a museum to acquire a certain artist – through purchase, gift, donation, or some other method of

transfer – has had a measurable effect on their performance with collectors and/or other museums.[44]

Any acquisition occurs in a market setting where museums, dealers, and collectors compete for artworks. When a museum acquires an artwork, this can have multiple effects. We already know from Fraiberger et al. that the inclusion of an artist in an exhibition by a (specific) museum positively affects their reputation and success. When museums acquire specific works, they exert even more influence on the recognition of objects and, ultimately, how the canon is formed. When an object enters a museum collection permanently, the object is automatically elevated to the level of so-called museum quality, which at once marks it as cultural heritage – worth collecting, storing, and preserving.

When a museum department buys an object, there is always a decision to purchase one object over other objects. This is because museum departments are usually restricted by financial constraints or infrastructural factors, such as limited storage. While a lack of storage also concerns donations and bequests, their acceptance does not affect museums' generally scarce acquisition budgets. Moreover, with their expert judgements, we can assume that curators are more invested in and have a bigger influence on the selection process of what museums buy and when than they do on donations and bequests. It is, therefore, logical to assume that of all the acquisition methods, purchases by museums have the highest symbolic meaning. This can, in turn, affect the growing reputation and value of an artist, style, or a related group of objects when observed on a micro-level by market participants. Although we would like to acknowledge the many nuances of acquisition processes, such as the influence of curators on future donors, for our preliminary analysis on a macro-level, we have focussed on the simplified categorisations of *active* and *passive* purchases.[45]

So how has the Art Institute of Chicago collected objects since its founding? From their collection of circa 300,000 objects, 122,317 object records are publicly available through their data dump. We could draw from 10,776 of these objects and their accompanying provenance texts for our preliminary analysis.[46] Our main objective was to differentiate between objects that were purchased actively by the Art Institute (as indicated by the terms *sold* and *purchased*) or acquired passively through donation (indicated by the terms *given*, *gifted* or *gift*) or bequest (indicated

44. An example of an analysis of the impact of museums on the art market can be found in Pommerehne and Feld (1997).

45. For a discussion of acquisition categorisation, see MacDonald (2022), p. 7.

46. A total of 616 provenance texts could not be included in this analysis as they did not contain an explicit record of the Art Institute as a receiver of the object. The analysis therefore considered 10,776 provenance texts from the total dataset of 122,317 object records.

by *bequeathed*).[47]

Such a birdseye view of a collection's history of acquisition patterns provides insights into the institution's history (fig. 4.3). Where the Art Institute is concerned, the institution began with an active purchasing period, which was followed by an increasing reliance on donations in the first decades of the 20th century. From the 1930s to the 1950s, the institution then increased its share of purchases compared to gifts, followed by a period reaching into the 1980s of relative stability in the ratio of purchased to donated objects. Since the 1980s, the institution's activities have been marked by an increase in passive acquisitions rather than purchases.

What emerges from the data beyond this broad view are, of course, outlier events that punctuate the institution's history of acquisitions. Of particular interest are spikes in purchasing that are easily identifiable. The first peak of purchases occurred in the 1890s. Then, after steady growth, a second peak occurred in the 1950s. This latter peak also marks the decade in which the highest number of objects were directly purchased by the museum. A closer look at the data reveals that these two peaks can be attributed to specific departments, namely, the Arts of Africa in the 1890s and the Arts of the Americas in the 1950s. Both purchasing peaks were related to singular and exceptional circumstances in which collections of objects became available on the market or were offered, and the museum decided to act. It should be noted that the numbers here reflect the number of objects rather than their prices. A graph representing the monetary value of purchases and donations may look much different due to single objects, such as modern paintings, being worth multiple times that of objects from other collecting areas.

From the perspective of researchers, such preliminary insights into acquisition patterns already offer a range of questions to be investigated. For example, researchers could explore the wider significance of the 1955 purchase by the Art of the Americas department of 571 objects from an illustrious collection of Pre-Columbian objects. Did the event change the museum's collecting strategy? Was it an event unique to the Art Institute, or did similar events with similar provenance details occur at other museums? Did the purchase impact the wider collecting of Pre-Columbian objects in the US? Did the purchase (or the potential simultaneity of similar purchases) have lasting effects on the valuation of Pre-Colombian artefacts?

Our second preliminary analysis relates to the market behaviour of specific people. From experience, we know that provenances usually record the locations of owners rather than the actual locations of their objects. Aggregate provenance data thus provides a general indication of the capacity for conspicuous consumption at a particular location. Conspicuous consumption in the form of collecting can also

47. Of the 10,776 acquisitions made by the Art Institute of Chicago, 4,329 are a purchase, 5,686 relate to a donation, and 523 are a bequest. The remaining 238 acquisitions refer to a variety of less frequent methods of transfer, excluded from this analysis.

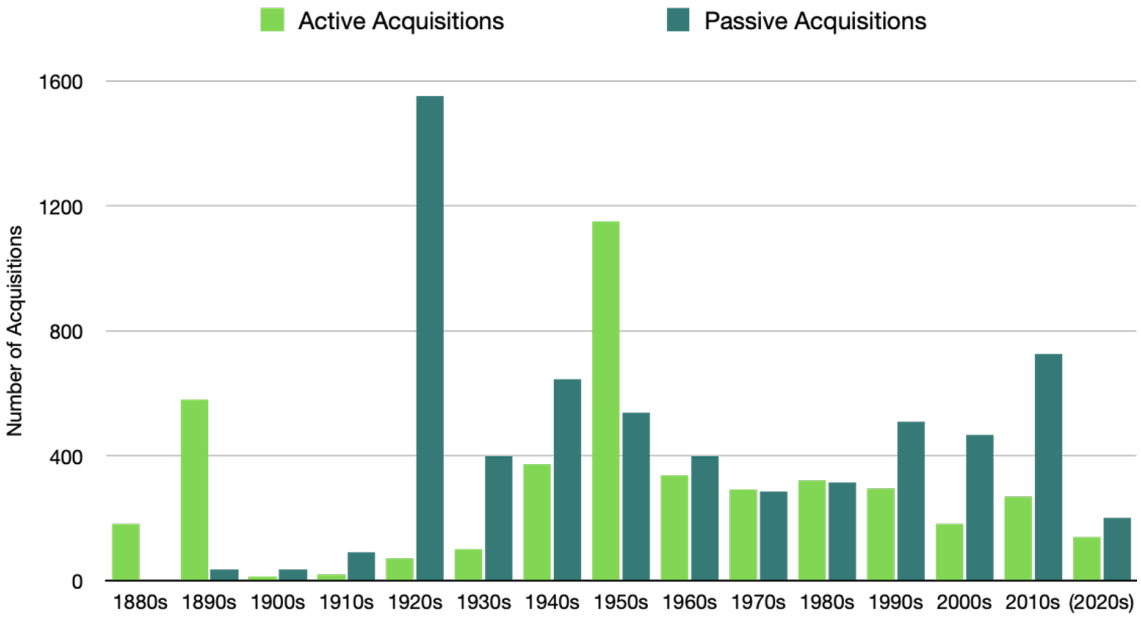

Figure 4.3: Comparison of Active and Passive Acquisitions of the Art Institute of Chicago by Decades from the 1880s to the 2020s. Source: 10,538 acquisition events extracted from provenance texts from the Art Institute of Chicago dataset downloaded on April 7th, 2022; created by Fabio Mariani.

function as an indicator of elite behaviour and can support a social and economic analysis of the differentiated developments of the elite and, indeed, their wealth.

On an individual level, provenances contain varying degrees of information about the fates of objects at the hands of their owners or at the hands of those who receive works of art due to death or divorce, for example. With aggregate provenance data, we can also contextualise the sale of an artwork and better understand whether this activity is part of a collector's larger, long-term collecting strategy or not. The ability to observe such behaviours allows the art historian to describe, for example, the changing shape of any given collection, which, when analysed together with the collecting activity of other collectors, can amount to a history of taste. For the economic and social historian, such a sequence can indicate the consolidation of wealth or the liquidation of one type of asset for another, i.e. art for money. From aggregate provenance data, we can thus gain insights into not only the relationship between collecting and larger macroeconomic trends but also the historically shifting role of art as an asset class within those trends.

Where aggregate provenance data records interdependencies of socioeconomic events, an individual can appear not only in one activity (e.g. as a sender or receiver) but also in multiple activities across provenances, where one activity (e.g. inheritance) relates to another (e.g. selling or donating). Inheritances are especially promising for the analysis of provenances, as they are not usually recorded in art

market data. And yet inheritances remain a key social phenomenon, where gender and wealth, in particular, intersect in meaningful and quantifiable ways. To study gender is to study mechanisms of exclusion within the art market.

A closer look at the behaviour of women who inherit objects can provide knowledge of how women have historically adapted to their positions as inheritors. With aggregate provenance data, it will be possible to study their attitudes toward the wealth passed on to them, for example, and compare their attitudes with those of men. Are they more or less likely to sell their inherited artworks on the market, pass them on to family members, or gift them to institutions? Regarding the latter, it would also be possible to measure gender disparities in philanthropic engagement and how they have developed over time and across geographies.

For our preliminary analysis, we identified a total of 3,151 inheritance events in which an owner passed an object to another owner by one of the several types of activity that fall under the umbrella of inheritance.[48] We then analysed the category of the event following the inheritance event. We found that in 41.3 percent of the cases, those parties who inherited a work of art subsequently sold it. In 11.5 percent of the cases, they made a bequest; in 13.5 percent of the cases, they made a donation (96 percent of which went to the Art Institute); and in 18.4 percent of the cases, it was the museum itself that inherited the artwork, thereby bringing an end to the object's provenance.[49]

Based on this information, we were able to analyse the data further. Of the 3,151 inheritance events, only 1,056 occurrences involved individuals, and for 831 of those, we can identify the following event.[50] These 831 events could then be broken down further by gender (fig. 4.4). If we consider the actions of men (403 events), visualised in figure 4.4, we see that they are most likely to sell an object they inherited (45.2 percent of the time), followed by making a bequest (38.2 percent) and, lastly, making a donation (13.1 percent).[51]

For women (428 events), the ranking looks different. They are most likely to bequeath an inherited object (43 percent), followed by donating it (29.7 percent). Selling an inherited object is ranked third in their list of activities (22.9 percent), thus highlighting significant behavioural differences between men and women.

While the underlying data for such statistics was taken from one specific collec-

48. The terms we gathered under the category of "inheritance" are: bequeathed, descended, bequested, inherited, by descent, by bequest, by descendent, by inheritance, bequest, and variations thereof with typographic errors.

49. The remaining 15.2 percent were divided among other methods (3.7 percent), the complete absence of information on the next event in the original provenance text (6.4 percent), and the absence of information on the activity of transfer to the immediately following owner (5.1 percent).

50. We excluded groups of people because they did not represent individual decisions.

51. Both groups engage in other types of activities as well, but we have grouped them under *other*, given their statistical insignificance.

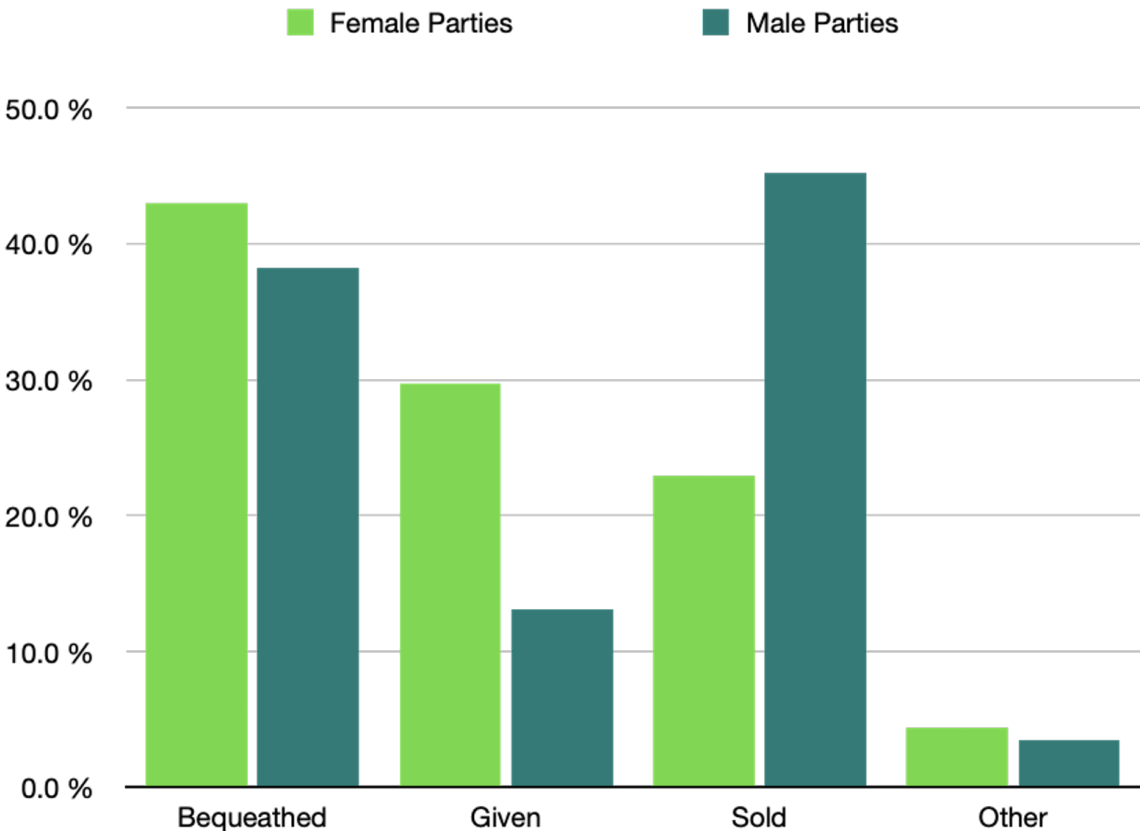

Figure 4.4: Comparison of Different Methods of Transfer Following an Inheritance and Their Relative Importance According to Gender. Source: 831 events following an inheritance involving an individual, extracted from provenance texts from the Art Institute of Chicago dataset downloaded on April 7th, 2022; created by Fabio Mariani.

tion and can therefore not be treated as representative in the way that a much larger, cross-institutional and geographically diverse set of aggregate provenance data could be, the results nevertheless point out the way for specific avenues of further research. Indeed, by comparing men and women alone and their actions when passing on an inherited object, we can observe stark gender differences. This may, in turn, indicate a broader difference in gendered attitudes towards direct engagement in the market. To what extent such a difference is a product of external factors, such as barriers when accessing the market, or other cultural and societal factors, is a question that would have to be addressed with further research. Suffice it to say, with the help of aggregate provenance data as presented in this paper, we can not only explore large-scale gender dynamics of art circulation but also bring them into relief.

4.6 Conclusion

Transforming provenances into sources for computational analysis by historians is not only a promising endeavour for interdisciplinary research, spanning art, social, and economic history but also technologically feasible. As we have demonstrated, making provenances available as a source for quantitative analysis depends on structuring the information contained within them. Only when we have structured provenances into machine-readable data can we analyse aggregate provenance data. By describing how we can do just that with the help of digital methods, our paper hopes to bridge the gap between computer science, the humanities, and the social sciences.

Our paper has spotlighted several related issues for social and economic historians to consider, which can be addressed with the help of aggregate provenance data: from the socio-economic construction of value to questions of wealth, and the role of gender contained therein. At the same time, we have shown that, based on aggregate provenance data, any element in provenances can be analysed, be it parties, locations, time periods, or methods of transfer. With this possibility to pursue narrow, specialised inquiries into the fates of particular artists, collectors, or institutions, aggregate provenance data also brings the potential for complex comparative analyses on much larger scales. Moreover, the discreteness of the information found in provenance descriptions allows us to map a whole host of networks across time and geography.

Lastly, we acknowledge that the act of making the information contained within provenances searchable and analysable through digital methods is but a preliminary step in the development of provenance data. Ultimately, efforts to digitise provenances and structure provenance data will deliver a different kind of infrastructure altogether: provenance linked open data. In its connectivity across institutions – including museums, libraries, archives, and other repositories – provenance linked open data promises to deliver unprecedented insights into the circulation of artworks.[52] It is towards this vision of provenance that our paper means to contribute.

Acknowledgement

The authors would like to thank the Art Institute of Chicago for making publicly available and downloadable detailed object data on its collection, including provenance information. We are grateful, in particular, to Amanda Block and Jennifer Cohen. We extend our gratitude to the anonymous peer reviewer and to Liza Weber for her rigorous and incisive editing of multiple versions of this article.

52. Rother, Koss, and Mariani (2022).

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

Bibliographic Information
Mariani, Fabio, Max Koss, and Lynn Rother. 2024. "People Information in Provenance Data: Biographical Entity Linking with Wikidata and ULAN." *Život umjetnosti*, no. 114, 148–161. https://doi.org/10.31664/zu.2024.114.07.
CRediT Roles: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization

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.

5.1 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 the integration and navigation of 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, including biographical information.[3] ULAN, for

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, and Mariani (2022).

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 were more likely to sell an inherited artwork. For more insights into the method and analysis, see Rother, Mariani, and Koss (2023).

3. Vrandečić and Krötzsch (2014).

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 in regard to information stemming from provenance research.

5.2 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), set forth in 2001. Published in response to the watershed 1998 Washington Conference Principles on Nazi-Confiscated Art that codified in a legally non-binding way 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

4. Harpring (2010).

5. Comparisons between cultural heritage LOD repositories have been examined in several studies. Sugimoto (2023) compares platforms connectivity, including Wikidata and ULAN, across different aspects, including people. A comparison between platforms focused on individuals is found in Freire, Manguinhas, and Isaac (2020) and Goldfarb and Merkl (2018). Context-specific comparisons based on the analysis of a dataset in relation to Wikidata and ULAN have been conducted, for example, in Faraj and Micsik (2021).

6. Yeide, Akinsha, and Walsh (2001).

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 rules of punctuation.

The Art Institute provenance dataset that we were thus able to build contains 11,392 provenance texts divided into 35,554 distinct provenance events.[7] 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"). Complicating 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

7. Rother, Mariani, and Koss (2023).

8. The experiment was carried out using Art Institute of Chicago data downloaded from the museum repository on April 7th, 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 (Rother, Mariani, and Koss (2023)).

9. Mariani, Rother, and Koss (2023b).

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, lacking a clear ownership period, we know that the owner either separated from an object before their date of death or that it passed 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 reconcili-

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.

ation.

In light of this, it is crucial to consider how female individuals are named in provenance texts. 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.

5.3 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, it became necessary to follow a two-step match discovery process involving quantitative and qualitative approaches.

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 of these figures are stark expressions of the underrepresentation of female par-

11. The selection process was conducted using OpenRefine reconciliation API services and SPARQL queries to the respective platform endpoints on February 5th, 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.

ties 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 number of entities, its generalist encyclopedic scope makes it prone to ambiguity. Addressing this limitation would require museums to record individuals consistently 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 (F. 5.1).

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 6th, 2024. On these platforms, disagreements related to birth and death dates can be found. Therefore, multiple dates were taken into account when necessary.

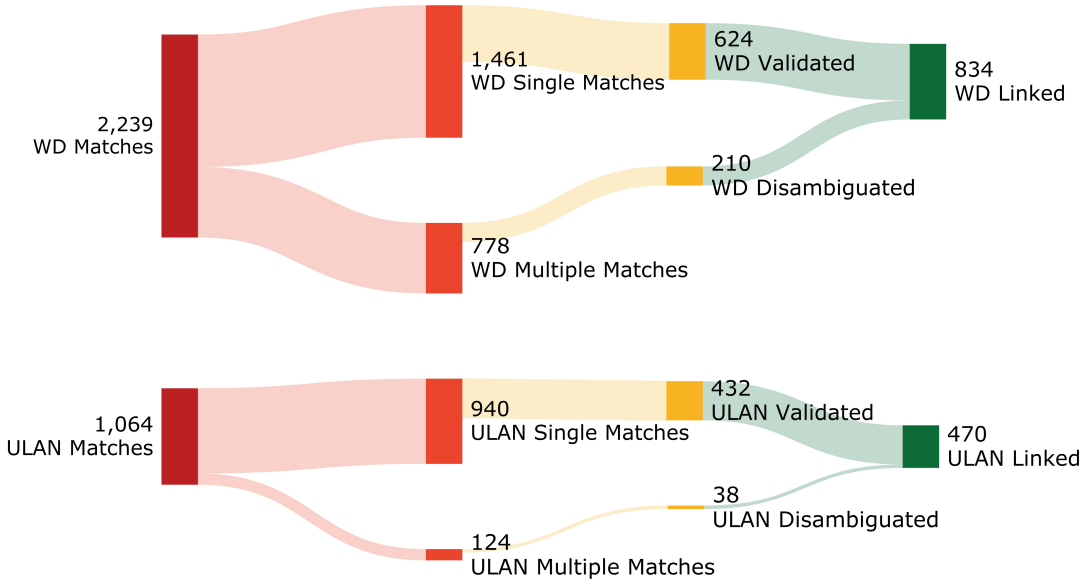

Figure 5.1: Sankey diagram summarizing the process of match finding, validation, disambiguation, and entity linking for both Wikidata (WD) and ULAN entities.

5.4 Entity Linking: Authority Control and Data Enrichment

In the context of an 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]

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

13. An example of network analysis applied to the study of the art market can be found in Schich et al. (2017).

14. A comparable classification approach was also introduced in Goldfarb and Merkl (2018).

15. This trend was also noted in a broader analysis of ULAN entities (Goldfarb and Merkl (2018)).

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.

5.5 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.

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% (F. 5.2).[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

16. SPARQL queries were executed at the respective platform endpoints on February 6th, 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.

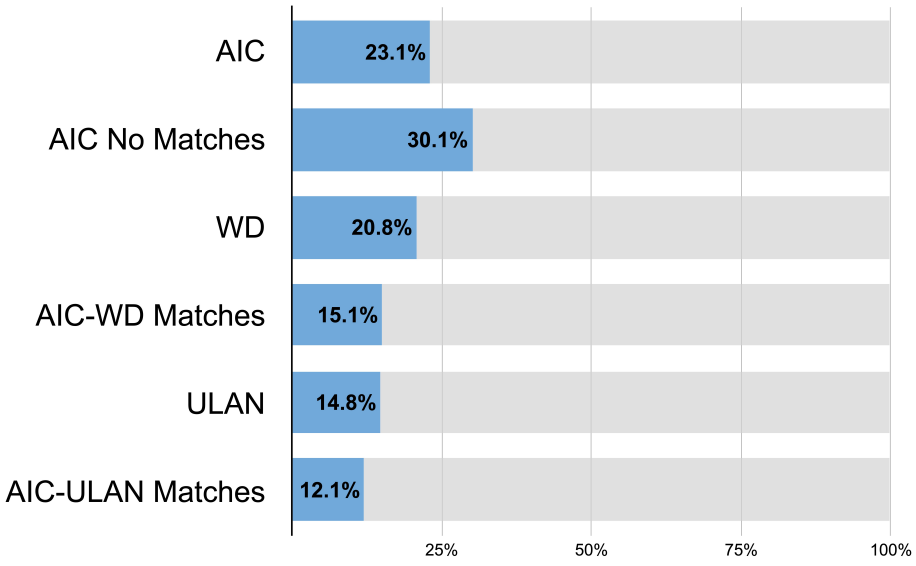

Figure 5.2: Female parties representation across Art Institute of Chicago (AIC), Wikidata (WD), and ULAN.

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 5.1 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 crowd-sourced 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 the fore. Such contribution not only streamlines the museum's data management processes, avoiding information

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)	721		
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		

Table 5.1: Table of the 13 individuals documented in the provenance records of the Art Institute of Chicago who participated in more than 100 provenance events.

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.

5.6 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.

6. Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data

Bibliographic Information
Mariani, Fabio. 2023. "Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data." In *Proceedings of the Workshop on Computational Methods in the Humanities 2022. Lausanne, Switzerland.* https://ceur-ws.org/Vol-3602/paper5.pdf.

Abstract
The acronym VISU refers to Vagueness, Incompleteness, Subjectivity, and Uncertainty found in provenance records, which document the history of ownership and socio-economic custody changes of an object. VISU information represents the intellectual effort of researchers and its limits in reconstructing historical events from archival sources. Although provenance has mainly been used in the past to assess an object's artistic and economic value, it has recently become crucial information from an ethical and legal viewpoint. In light of this, there is a growing interest in structuring provenance information in a machine-readable format and making this data openly accessible to anyone, e.g., by publishing provenance data as linked open data. However, with the impetus to publish provenance linked open data, we risk losing or simplifying VISU information. After describing VISU information and analysing current community standards, this article illustrates how to represent such information in publishing provenance linked open data.

6.1 Introduction

Provenance records document chains of events of ownership and socio-economic custody changes of an object. These records contain historical information that answers the question: from where did it come? This article focuses on the provenances of objects with artistic or cultural value held by a gallery, library, archive, or museum (GLAM).

In the art market, documenting provenance has been a means of establishing

the value of artworks since the eighteenth century [1]. For example, if a well-known and highly respected collector owned an object, then they would contribute to its supposed authenticity and aesthetic value, determining its economic value [2]. By the late twentieth century, however, provenance's moral, ethical, and legal entanglements became a subject of scrutiny and debate. As a consequence of colonialism, totalitarian regimes and two world wars, many objects improperly changed hands due to seizures, confiscations, and looting. For this reason, documenting and establishing the life story of an object has become crucial in establishing its rightful owner. The 1998 Washington Conference on Holocaust-Era Assets foregrounded the importance of provenance research to find and return art and cultural property confiscated by the Nazi regime [3]. At the conference, the 44 participating governments and 13 non-governmental organisations agreed on eleven non-binding principles ("The Washington Conference Principles on Nazi-Confiscated Art") resolving disputes over Nazi-looted art through the study of provenance. As a result of these principles, provenance research has become more professionalised, acquiring interdisciplinary characteristics. In fact, it has become something of an academic field in its own right [4]. The increased importance of provenance from not only an economic perspective, but also an ethical and legal one, has put a spotlight on the responsibility of institutions. Indeed, the accountability and transparency to which GLAM institutions are being held also depend on researching and publishing the provenance records of the objects for which they are responsible.

However, recording provenance is a complex process and requires a considerable investment of resources. On the one hand, careful research of sources is necessary to reconstruct the history of an object. On the other hand, this effort requires consideration in curating and publishing any information obtained. Moreover, the efforts of a single institution must be coordinated with other stakeholders in the GLAM domain. Recently, digital tools and methodologies have opened up new possibilities to assist the curation, publishing, and analysis of provenance data. In particular, the publication of provenance linked open data promises unprecedented levels of standardisation, enabling researchers to analyse the context of object histories in their cross-institutional complexity [5].

Considering the benefits of provenance linked open data, it is crucial to identify and address its related risks and challenges. The exclusion or simplification of historical complexity could reduce the quality of information, which could, in turn, cause harm when considering the ethical and legal implications of provenance. It is

1. Raux (2012).
2. Gramlich (2017).
3. U.S. Department of State, Office of the Special Envoy for Holocaust Issues (1998).
4. Fuhrmeister and Hopp (2019).
5. Rother, Koss, and Mariani (2022), Luther (2020), and Newbury and Lippincott (2019).

no coincidence that among the principles that emerged from the 1998 Washington Conference, it is advised that "consideration should be given to unavoidable gaps or ambiguities in the provenance..." [6].

In this article, we aim to categorise such "unavoidable gaps or ambiguities in the provenance" as they are likely to be compromised in publishing provenance linked open data. Indeed, recording provenance requires considerable intellectual effort in interpreting sources and formulating hypotheses about an object's history. Such a hermeneutic process is prone to produce Vagueness, Incompleteness, Subjectivity, and Uncertainty (VISU). In publishing provenance linked open data, it is, therefore, critical to maintain the integrity of the intellectual process, with its hypothetical statements and its dealing with gaps in knowledge. Given provenance's complexity, this article, in addition to identifying and classifying VISU information, introduces implementation solutions to represent it as linked open data. These solutions comply with current data publishing standards in the cultural heritage domain.

6.2 Vague, Incomplete, Subjective, and Uncertain Information

The growing requirement for institutions to be more transparent and accountable has prompted them to publish information about the provenance of objects in their collections. Currently, the provenance of an object is recorded manually as textual metadata through collection management software. Although there is not yet a shared standard for transcribing this information, the American Alliance of Museums (AAM) has drafted guidelines for compiling provenance texts [7]. To give an example, below is the provenance text of a painting by André Derain from 1910 titled "Cagnes", which is published on the Art Institute of Chicago website and has been compiled according to the AAM guidelines:

> Galerie Kahnweiler, Paris, probably acquired directly from the artist. Louis Lion & Co., New York, by Feb. 1957 [verso inscription; this and the following according to letter from Knoedler and Co., Apr. 8, 1975, copy in curatorial file]; sold to Knoedler & Co., New York, Feb. 1957; sold to the Art Institute of Chicago, 1960.[8]

According to the AAM guidelines, provenance editors should list events in chronological order, from the object's creation to the acquisition by its current owner.[9] An

6. U.S. Department of State, Office of the Special Envoy for Holocaust Issues (1998).

7. Yeide, Akinsha, and Walsh (2001).

8. https://www.artic.edu/artworks/12402/cagnes (accessed 2023-08-11).

9. Usually, the creation event is omitted in the provenance text as it is recorded in other appropriate metadata fields, such as author, date, and place of creation.

event represents a change of ownership, or custody, of the object from one party to another. Each event consists of the acquisition method, location, date, names of the parties, and their related biographical information.

Punctuation separating events has a specific meaning: a semicolon implies that the transaction from one party to another was direct; a period indicates a gap in the reconstruction of the events. For example, the period at the end of the first recorded event listed above, when Galerie Kahnweiler received the object, indicates a gap in the provenance record of the painting "Cagnes". This means, therefore, that it is unknown how the painting passed from Galerie Kahnweiler to Louis Lion & Co., its next recorded owner. Potentially, there could have been other owners of the object that have yet to be identified.

When there is no sufficient certainty about an event, the AAM guidelines suggest using the terms "probably" and "possibly", depending on the level of uncertainty. In analysing the provenance text of the painting "Cagnes", we can see that the authors were not certain about the first recorded event, and therefore used the phrase "probably acquired directly from the artist".

Finally, notes can provide additional information regarding the provenance. In the above example, the Art Institute of Chicago uses notes in square brackets. Notes in compiling a provenance text are necessary since the chronology of events results from careful research of disparate archival sources, such as inventories, letters, and even photographs. Indeed, sometimes a provenance expert can find a source for reconstructing an event on the object itself. For example, we know that Louis Lion & Co. owned Derain's artwork through an inscription on the back of the painting ("verso inscription").

From what has been discussed, it is clear that reconstructing ownership histories is not a straightforward process since it requires intellectual and critical effort in analysing the available historical sources and formulating hypotheses. Moreover, sources are not always available to reconstruct events, and some information may not be immediately evident. We have classified these phenomena into four categories: Vagueness, Incompleteness, Subjectivity, and Uncertainty. We have gathered them under the acronym VISU, from the Latin *de visu*, meaning with your own eyes. Vagueness refers to information that is given with certainty but in an approximate way. An approximation can occur when describing spatial information (e.g., near Paris) or temporal information (e.g., circa 1945). In either case, the vagueness of the information does not affect the certainty of the event. Incompleteness refers to a lack of information in the reconstruction of an object's provenance. In this case, provenance experts may not have formulated any hypotheses yet to address the missing information. Subjectivity concerns the expert's interpretive context when reconstructing an object's provenance—how they formulated hypotheses through

source analysis and deduction. Moreover, different assumptions may contradict each other. Finally, uncertainty refers to the level of confidence with which a provenance expert has expressed a hypothesis, using terms such as "possibly" or "probably". Unlike vagueness, uncertainty questions the very occurrence of a given event.

The categories of what we define as VISU have already been a topic of interdisciplinary debate, from philosophy and mathematics to, more recently, computer science [10]. In Smithson's taxonomy of ignorance, for example, the concept of uncertainty represents a generic term that, in turn, can be divided into more specific concepts such as vagueness and probability [11]. The latter is closer to our definition of uncertainty. In contrast, Smets distinguishes more sharply between uncertainty and imprecision in providing a taxonomy of imperfection [12]. Smets' imprecision can be compared to the vagueness of VISU information. At least lexically speaking, a classification close to that of VISU is provided by Nagypál and Motik [13]. Here, the categories of uncertainty, subjectivity, and vagueness are defined in relation to expressing temporal knowledge. However, the meaning given to each term is different from that intended in VISU. In fact, according to their classifications, uncertainty (e.g., circa 1918), subjectivity (e.g., the dating of the Russian Revolution), and vagueness (e.g., in February 1918) are all ascribable to the concept of vagueness in VISU. In analysing uncertainty in the digital humanities domain, Piotrowski recognises the conflict of interpretations between scholars as an additional aspect of dealing with "uncertain, vague, incomplete, or missing information" [14]. In doing so, Piotrowski partially anticipates the classification we propose with the acronym VISU, since the conflict of interpretations is one aspect of what we define as subjectivity.

6.3 Provenance Linked Open Data

As previously discussed, institutions currently create and share provenance records in text format. Although provenance texts are stored and published online digitally via collection management systems, the text format limits the use of provenance as a research and study tool. Indeed, it is currently impossible to use provenance data to perform large-scale analyses across multiple institutions through the application of, for example, digital methods such as big data queries, network analysis, and spatial analysis [15]. These limitations can be attributable, on the one hand, to the

10. Piotrowski (2019).
11. Smithson (1989).
12. Smets (1997).
13. Nagypál and Motik (2003).
14. Piotrowski (2019).
15. Jaskot (2020).

fact that textual information is not machine-readable and, on the other hand, to the fact that it is not published according to FAIR principles [16]. Indeed, as it stands, provenance information, which is siloed as text in collection databases of institutions, is not findable, accessible, interoperable, and reusable. For these reasons, publishing provenance linked open data (LOD) has recently emerged as a promising possibility to address the standardization of provenance information produced by institutions in a machine-readable format compliant with FAIR principles [17]. Moreover, LOD respects the open data principles: that is, provenance LOD can be used by anyone for any purpose.[18] Provenance data should be published as open data, not only because it involves historical facts but also because of the significance of provenance for institutional accountability and transparency.

A significant early experiment in publishing an institution's provenance records as LOD was carried out within the Art Tracks project, an initiative of the Carnegie Museum of Art (CMOA), which took place from 2014 to 2017 [19]. In particular, Art Tracks implemented the CMOA Digital Provenance Standard for modelling provenance LOD following the CIDOC CRM schema, the international standard for exchanging digital information regarding cultural heritage (ISO 21127).[20] The schema of CIDOC CRM is event-based since its semantic structure has temporal entities (crm:E2_Temporal Entity) as its core [21]. A temporal entity, such as an event (crm:E5_Event), can link to time (crm:E52_Time-Span), space (crm:E53_Place), or event actors (crm:E39_Actor). However, the centrality of the temporal entity means that an actor, such as a person (crm:E21_Person), cannot link directly to a time or place. For example, CIDOC CRM does not express an individual's birth date as a person's attribute, but rather as a specific event, birth (crm:E67_Birth), linked to a time and involving that person. In turn, the birth event can be linked to the location of the event.

In order to make CIDOC CRM modelling more accessible to institutional practitioners, Linked Art, a community of cultural heritage institutions, developed a CIDOC CRM application profile.[22] In addition to CIDOC CRM, the Linked Art Data Model integrates the Getty's controlled vocabularies, such as the Art and Architecture Thesaurus (AAT), to identify domain-specific terms via URI.[23] The

16. Wilkinson et al. (2016).
17. Rother, Koss, and Mariani (2022).
18. https://opendefinition.org/ (accessed 2023-08-11).
19. Newbury (2017).
20. CIDOC CRM (version 7.2) is the Conceptual Reference Model (CRM) implemented by the International Committee for Documentation (CIDOC) of the International Council of Museums (https://www.cidoc-crm.org/, accessed 2023-08-11).
21. Doerr (2003).
22. https://linked.art/model/ (accessed 2023-08-11).
23. https://www.getty.edu/research/tools/vocabularies/aat/ (accessed 2023-08-11).

integration of CIDOC CRM and Getty vocabularies, combined with the support of a large and active community behind the Linked Art Data Model, make this application profile an ideal candidate for the standardisation of publishing provenance LOD. Indeed, modelling provenance LOD is one of the aspects that the Linked Art Data Model covers in detail. According to Linked Art, a provenance record structured as LOD is a succession of provenance events (or activities), structured in CIDOC CRM as crm:E7_Activity. An activity can itself consist of multiple activities (sub-activities), expressing more complex events. Linked Art provides a pattern for defining the characteristics of events: the object(s) involved, the actors participating, the location, and the time. Examples of how to structure the data are given depending on the different types of activities. For example, an activity describing the purchase of an object may contain two sub-activities. The first activity consists of the acquisition of the object given by the seller and received by the buyer, while the second constitutes the payment made by the buyer to the seller. Similarly, exchanging two objects involves two sub-activities, each describing the respective ownership change.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com/activity/4> a crm:E7_Activity ;
    rdfs:label "Purchased by the Art Institute of Chicago from Knoedler & Co. in
        1960" ;
    crm:P2_has_type <http://vocab.getty.edu/aat/300055863> ;
    crm:P2_has_type <http://vocab.getty.edu/aat/300417642> ;
    crm:P4_has_time-span [ a crm:E52_Time-Span ;
        crm:P82a_begin_of_the_begin "1960-01-01T00:00:00Z" ;
        crm:P82b_end_of_the_end "1960-12-31T23:59:59Z" ] ;
    crm:P9_consists_of [ a crm:E8_Acquisition ;
        crm:P22_transferred_title_to [ a crm:E74_Group ;
            rdfs:label "The Art Institute of Chicago" ] ;
        crm:P23_transferred_title_from [ a crm:E74_Group ;
            rdfs:label "Knoedler & Co." ] ;
        crm:P24_transferred_title_of [ a crm:E22_Human-Made_Object ;
            rdfs:label "Cagnes" ] ] .
```

Listing 6.1: RDF description, serialized in Turtle format, of the purchase of the painting "Cagnes" by the Art Institute of Chicago from Knoedler & Co. in 1960.

Listing 6.1 shows the RDF description of the last provenance event of André Derain's painting "Cagnes": the purchase of the artwork by the Art Institute of Chicago from Knoedler & Co. in 1960. RDF, Resource Description Framework, is a World Wide Web Consortium standard for information exchange as LOD. The activity (crm:E7_Activity) is classifiable according to AAT vocabulary as "provenance" (aat:300055863) and "purchase" (aat:300417642). Moreover, the activity took place in 1960, a time span expressed through its time limits: the begin of the begin "1960-01-01T00:00:00Z" (the minimum possible date) and the end of the end "1960-

12-31T23:59:59Z" (the maximum possible date). The activity has a sub-activity (crm:E8_Acquisition), which describes the acquisition of the painting "Cagnes" by the Art Institute of Chicago from Knoedler & Co.

In making CIDOC CRM modelling usable to practitioners, Linked Art deliberately leaves out some aspects that would complicate the accessibility of the data model, such as uncertainty and data provenance. However, this choice compromises the integrity of VISU information when modelling provenance LOD. As discussed in the previous section, VISU information is based on the intellectual work of provenance experts, who research and record provenance. Moreover, with VISU information, historical debate and hypothesis-making become critical to achieving the most scientifically accurate reconstruction of an object's history. Forgoing VISU information thus not only compromises the integrity of the data but also prevents debate, thereby reducing its usefulness for research. This phenomenon, also referred to as the "lure of objectivity", is one of the major challenges in digital humanities [24]. We, therefore, intend to safeguard the complexity of VISU information by making it machine-readable according to LOD standards and compatible with the Linked Art Data Model. The following sections describe the challenges, opportunities, and solutions in dealing with VISU information as LOD.

6.4 Vagueness

By introducing VISU information, we have established a clear distinction between the concepts of vagueness and uncertainty that previous scholarship, as noted, has not made consistently. Vagueness indicates the approximation of a datum. Approximating a datum per se does not compromise the statement's certainty. For example, to say that an event occurred near Paris is to approximate the geographical location of a temporal entity. The fact is not called into question. Similarly, the existence of an event that occurred circa 1945 is not questioned by the temporal approximation of the date.

Since vagueness concerns spatial and temporal information, it depends on the measures and language used by historical sources. Indeed, whereas technology allows us to calculate space and time with utmost precision, human language can hardly replicate its accuracy. Compare, for example, the limitations of language in traditional art market information, such as the inventory of an art dealer, with the measures of modern digital data, such as an online auction house database. In the written inventory of an art dealer, an event date can achieve maximum accuracy by expressing the year, month, and day of the event. Usually, however, vague reference systems such as months or seasons are used. Seldom does an author of a source

24. Rieder and Röhle (2012).

go into details such as the exact hour of an event. In contrast, an online auction house database can capture the moment of purchase to a thousandth of a second. Similarly, whereas human language cannot go beyond the precision of an address to indicate spatial information, technology allows us to pinpoint the geographical coordinates of a place with greater accuracy.

In addition to measuring instruments and human language, an approximation can result from a lack of information. For example, in the provenance text of the painting "Cagnes", the second provenance event states that Louis Lion & Co. owned the work "by Feb. 1957". The author of the provenance record used this expression because they had no sources to establish when Louis Lion & Co. received the object precisely. We do not even know who the previous owner was, expressed using a period that signifies a gap in the painting's provenance text. What the provenance expert can establish from the historical information available, however, is that Louis Lion & Co. had the object in February 1957 since the sources show that they sold the work to Knoedler & Co. in that month. Thus, we can assert that the acquisition of the painting by Louis Lion & Co. took place between 1910, the previous known date and thus the lower limit of the possible time interval, and 28 February 1957, the last day of the month in which Knoedler & Co. acquired the object. Experts can formulate subjective hypotheses with different degrees of uncertainty based on a vague time expression such as "by Feb. 1957". For example, according to the available information, it is possible, although very unlikely, that Louis Lion & Co. acquired the object on 28 February 1957 and sold it to Knoedler & Co. on the same day.

In CIDOC CRM and Linked Art, one can already model some vague information. This ensures that information is not falsified when publishing provenance LOD, which runs the risk of making vague information seemingly precise. It also opens up possibilities for data analysis and visualization that include this layer of complexity. Concerning the approximation of spatial data, CIDOC CRM introduces the property crm:P189_approximates. Using this property makes it possible to establish an approximation relation between two places. For example, in Listing 6.2, we see how the place "Paris", defined as a point in space, approximates the expression "near Paris". In this way, we preserve the vagueness of the information on the one hand. And on the other hand, we model a point in space that, albeit approximate, allows us to query the geospatial datum and visualize it on a map. The Linked Art Data Model already includes this modelling solution.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com/place/1> a crm:E53_Place ;
    rdfs:label "Paris" ;
    crm:P168_place_is_defined_by "POINT(2.2769957 48.8589466)" ;
    crm:P189_approximates [ a crm:E53_Place ;
        rdfs:label "near Paris" ] .
```

Listing 6.2: RDF description, serialized in Turtle format, of the "near Paris" approximation.

As far as temporal information is concerned, as we have already seen when introducing CIDOC CRM and Linked Art in Listing 6.1, it is represented as a time span. Thanks to the properties crm:P82a_begin_of_the_begin and crm:P82b_end_of_the_end, this type of modelling makes it possible to model several vague chronological pieces of information [25]. For example, Listing 6.3 shows the modelling of the time span in which Louis Lion & Co. acquired the painting "Cagnes". As previously discussed, the activity occurred sometime between 1910 (begin of the begin) and February 1957 (end of the end). In addition, this approach allows for modelling other approximate expressions in which an event occurred, such as months, seasons, years, decades, centuries, and millennia.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com/timepsan/1> a crm:E52_Time-Span ;
    rdfs:label "between 1910 and February 1957" ;
    crm:P82a_begin_of_the_begin "1910-01-01T00:00:00Z" ;
    crm:P82b_end_of_the_end "1957-02-28T23:59:59Z " .
```

Listing 6.3: RDF description, serialized in Turtle format, of the time span between 1910 and February 1958.

Although CIDOC CRM allows us to model temporal information as a time span, it does not allow the representation of an approximation that occurs around a date, such as the expression "circa 1945". However, it is possible to integrate the CRM-geo module to overcome this limitation. This extension of CIDOC CRM, dedicated to a more complex representation of spatiotemporal data, introduces the property crmgeo:Q13_approximates [26]. Like the crm:P189_approximates for places, the crm-geo:Q13_approximates property establishes an approximation relation between two time spans [27]. As an example, Listing 6.4 describes how the time span "1945"—with a begin of the begin as 1 January 1945 and an end of the end as 31 December 1945—approximates the vague time span "circa 1945". Therefore, we believe this

25. Holmen and Ore (2010).
26. Hiebel, Doerr, and Eide (2017).
27. Hiebel et al. (2014).

solution, similar to the one adopted for spatial approximation, can be integrated into the Linked Art Data Model.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix crmgeo: <http://www.cidoc-crm.org/rdfs/1.2/crmgeo#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com/timespan/1> a crm:E52_Time-Span ;
    rdfs:label "1945" ;
    crm:P82a_begin_of_the_begin "1945-01-01T00:00:00Z" ;
    crm:P82b_end_of_the_end "1945-12-31T23:59:59Z" ;
    crmgeo:Q13_approximates [ a crm:E52_Time-Span ;
        rdfs:label "circa 1945" ] .
```

Listing 6.4: RDF description, serialized in Turtle format, of the time expression "circa 1945" approximated by the time span 1945.

6.5 Incompleteness

In dealing with incompleteness, we must consider a trivial but essential fact: it is impossible to model as LOD what is unknown. Indeed, incompleteness is the only VISU information we cannot address directly in the modelling phase. However, conscious modelling of known information can help to address incompleteness through subsequent data analysis and in the hypotheses-making phase. Although we cannot model what we do not know, we can establish patterns of incompleteness against which we analyse the available information [28]. This approach first allows us to identify where and what information is missing and, secondly, to formulate new hypotheses with the help of data analysis.

The first pattern of incomplete provenance information that we can identify concerns gaps in the object's chain of activities. The importance of considering this kind of incompleteness for the integrity of a provenance record already emerges from the AAM guidelines. As we have already described, in provenance texts, events are divided by semicolons if transactions are direct and by periods if there are gaps in the ownership history of an object. Since we cannot directly model the presence of a gap as LOD, we must define a pattern to detect this incompleteness in the data. Linked Art describes the chronological linkage of provenance activities through the properties crm:P183_ends_before_the_start_of and crm:P183i_starts_after_the_end_of. These two properties allow us to determine whether an event occurred before or after another [29]. While they may help establish a chronological order of events, these properties are insufficient for identifying gaps between them. To detect such gaps, we must formulate the incompleteness pattern

28. Destandau and Fekete (2021).
29. Papadakis and Doerr (2015).

of the chain of activities: there is a gap between two events, A and B, linked in chronological succession (Activity_A crm:P183_ends_before_the_start_of Activity_B) if the party who receives the object in Activity_A is not the one who parts with it in Activity_B.

In Listing 6.5, we describe the activities involving the acquisition of the painting "Cagnes" by Galerie Kahnweiler from the artist and the subsequent acquisition by Louis Lion & Co. In this case, the scenario respects the incompleteness pattern of the chain of activities insofar as Galerie Kahnweiler was not the owner who gave the object to Louis Lion & Co. Identifying such a gap in analysis can lead to the formulation of new hypotheses since there may have been one or more intermediate owners prior to Louis Lion & Co. The gap in question is of significant interest to scholars as it conceals the events that caused the object to be moved from Paris to New York. Moreover, the gap overlaps with two world wars that affected, among other aspects, the circulation of artworks, legal or otherwise. In this scenario, the publication of provenance LOD is valuable because it allows us to analyse large amounts of provenance data from different institutions. Indeed, through network analysis, we can identify the most frequent pathways of artworks that, at some point in their lives, passed through Galerie Kahnweiler, as well as the most prominent agents from whom Louis Lion & Co. purchased artworks, thus opening up new hypotheses that try to bridge the gap.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com/activity/1> a crm:E7_Activity ;
    rdfs:label "Acquired by Galerie Kahnweiler from André Derain" ;
    crm:P2_has_type <http://vocab.getty.edu/aat/300055863> ;
    crm:P183_ends_before_the_start_of <https://example.com/activity/2> ;
    crm:P9_consists_of [ a crm:E8_Acquisition ;
        crm:P22_transferred_title_to [ a crm:E74_Group ;
            rdfs:label "Galerie Kahnweiler" ] ;
        crm:P23_transferred_title_from [ a crm:E21_Person ;
            rdfs:label "André Derain" ] ;
        crm:P24_transferred_title_of [ a crm:E22_Human-Made_Object ;
            rdfs:label "Cagnes" ] ] .

<https://example.com/activity/2> a crm:E7_Activity ;
    rdfs:label "Acquired by Louis Lion & Co." ;
    crm:P2_has_type <http://vocab.getty.edu/aat/300055863> ;
    crm:P9_consists_of [ a crm:E8_Acquisition ;
        crm:P22_transferred_title_to [ a crm:E74_Group ;
            rdfs:label "Louis Lion & Co." ] ;
        crm:P24_transferred_title_of [ a crm:E22_Human-Made_Object ;
            rdfs:label "Cagnes" ] ] .
```

Listing 6.5: RDF description, serialized in Turtle format, of the acquisition of the painting "Cagnes" by Galerie Kahnweiler from the artist, and the subsequent acquisition by Louis Lion & Co.

Different patterns of incompleteness can result from other missing constituents of an activity. As we discussed in introducing Linked Art, the data model introduces a pattern of event constituents. An activity is determined not only by its participating actors, but also by time and place and the object(s) involved. In addition, an activity can consist of several sub-activities, depending on its type. Thus, Activity_A is incomplete if the time, place, or object(s) involved are not expressed, or if one or more of the sub-activities associated with its type are missing.

When the time of an activity is unknown, incompleteness can be solved by generating vague information, that is, by defining that the event occurred in a time interval between the last previously known date before the activity and the first subsequently known date after the activity. As previously discussed in the section on vagueness, while we do not know when Louis Lion & Co. acquired the painting "Cagnes", we can infer that the activity occurred sometime between 1910 and 28 February 1957. The incompleteness of an activity's location proves to be a more challenging piece of information to reconstruct from the sources, except for when an event is specific, like an auction. In provenance texts, we find mainly geographical information about the actors. This can sometimes be useful in hypothesising the locations where events occurred. For example, we can infer that the purchase of "Cagnes" by Knoedler & Co. from Louis Lion & Co. occurred in New York since both companies were located there.

In contrast, the incompleteness concerning an activity and its sub-activities depends on the type of event, for which the Linked Art Data Model introduces a distinct structure. For example, as discussed, a purchase activity involves two sub-activities: 1) the acquisition of the sold object and 2) payment. In the previous section, we presented the LOD example of modelling the purchase of "Cagnes" in 1960 by the Art Institute of Chicago (Listing 6.1), the last event in the provenance record of that object. We can therefore assert that the activity is incomplete, since there is no sub-activity related to the payment made by the Art Institute of Chicago to the seller. Similarly, a provenance activity that concerns the exchange of one object for another will be incomplete if it consists of only one sub-activity, since one of the objects involved is not registered.

Additional types of incompleteness, which are difficult to ascribe to a fixed pattern, concern the biographical information of the actors involved. Missing biographical information of interest to the reconstruction and study of provenance may be: birth and death (or formation and dissolution, in the case of organisations), period and place of activity, and relationships to other actors. In addition to the direct intervention of historians, it is possible to use external knowledge published as LOD to fill in these gaps, such as the Getty's Union List of Artist Names (ULAN).[30] This

30. https://www.getty.edu/research/tools/vocabularies/ulan/ (accessed 2023-08-11).

controlled vocabulary can enrich our understanding of the actors of provenance activities with additional biographical information. In turn, enriching biographical information can help fill in other types of incompleteness. For example, by using the ULAN entity information of Louis Lion & Co. (ulan:500449799), we learn that the company has been in business since 1949. This new information allows us to, in turn, narrow down its purchase of the painting "Cagnes" to a time interval from 1949 to 28 February 1957.

Finally, it should be noted that provenance texts have a considerable bias in the representation of women. Many women are represented by their husbands' names ("Mrs John Doe") or even by the expression "the artist's wife". Such expressions compromise the historical representation of women and make it difficult for historians to identify female actors. For example, expressions such as "the artist's wife" are of little help if an artist had multiple wives. Modelling provenance LOD thus becomes an opportunity for historians to remedy such bias and finally give proper representation to people.

6.6 Subjectivity

Reconstructing the history of an art object is the result of laborious research by provenance experts, who hypothesise through the interpretation of sources what might have happened. Of course, the hypotheses of different experts may contradict each other, evolve with time, and become obsolete in light of new findings. As provenance texts stand, however, they cannot capture the hermeneutic and dialectical complexity of this intellectual process. In fact, except for notes to provide additional context for specific hypotheses, the texts are not accompanied by any publication information. For example, the author's name and publication date are critical metadata for information authority and versioning. The lack of versioning, in particular, can lead to the harmful practice of deleting a provenance text whenever an institution produces a new version. In this way, a debate concerning the provenance of an object is arbitrarily steered in a single direction, collapsing the idea that different historical interpretations can coexist.

It is possible to include publication information and versioning when publishing provenance LOD by implementing what is known as the data provenance of provenance data [31]. Just as we can trace an artwork's ownership history, we can trace the recording history of a given datum through data provenance. The recording history tracks when a datum was created, by whom, and when it was modified.

CIDOC CRM introduces the class crm:E13_Attribute_Assignment, a subclass of crm:E7_Activity, to describe the context in which an assertion is made regarding

31. Huemer (2020), Newbury and Lippincott (2019), and Al-Eryani, Bucher, and Rühle (2018).

an entity. An attribute assignment is the entity with which CIDOC CRM represents the n-ary relationship between the asserted entity and the assertion information. In this way, in addition to defining the asserted value, we can add additional statements to describe the context of the assertion, such as the author and date. Although this solution is also adopted in Linked Art to define, for example, authorship attribution, it tends to be verbose and redundant [32]. Focusing on the case of data provenance of provenance data, we found issues related to using attribute assignments to represent this type of information. An n-ary relation enables us to describe the context of an assertion pertaining only to a single statement. However, in the case of provenance, hypothesis-making does not concern a single statement but the assertion of an entire event and, thus, multiple statements. In this scenario, should we model an attribute assignment for each statement, we would need to repeat the same information multiple times. This situation would be even more complex in case of contradictory assumptions, as this requires us to produce multiple attribute assignments to describe conflicting hypotheses, resulting in an additional increase in statements. Moreover, such a solution would result in the coexistence in the same RDF graph of different and contradictory information about the same fact, compromising the usability of the data.

Given the nature of provenance information and the issues arising using attribute assignments, we considered other approaches. Among the many methods to represent data provenance as LOD, nanopublication is one of the most suitable [33]. Nanopublication is a way of publishing an atomic unit of information as LOD, providing data provenance and publication information [34]. In this way, it is possible to trace and reference these atomic units of information independently of the entire dataset, making the knowledge expressed more authoritative and compliant with FAIR principles [35].

In presenting provenance LOD modelling according to the Linked Art Data Model, we have seen how provenance activities are the constitutive elements of an event-based model. In light of this, we consider the provenance activity as the atomic unit of a provenance record published as a nanopublication. Thus, publishing provenance LOD as a nanopublication implies publishing each provenance activity as a stand-alone, referenceable, and citable unit. In this way, two conflicting hypotheses about the same activity can coexist while older hypotheses that have become obsolete can remain accessible to scholars [36]. In addition, each nanopublication expresses metadata about the creation of the information and its publication.

32. Daquino et al. (2022).
33. Sikos and Philp (2020).
34. Groth, Gibson, and Velterop (2010).
35. Sustkova et al. (2020).
36. Asif, Tiddi, and Gray (2021).

We can thus publish the data provenance of provenance data.

The structure of a nanopublication consists of three separate named graphs. A named graph is an RDF graph identified by a URI, which allows one to assert information about it [37]. The first graph of the nanopublication, the assertion graph, is devoted to the information on the published atomic unit. In the case of provenance data, it contains statements about a single provenance activity.

The second graph, the provenance graph, is dedicated to the data provenance related to the assertion graph. It contains statements about how the knowledge expressed in the assertion graph was produced. For example, in a nanopublication of a provenance activity, the provenance graph describes the context of the hypothesis formulated by an expert, including the author's identity, the date, the scientific method used, and the sources consulted by the author. Finally, the publication info graph, the third graph of a nanopublication, provides metadata about the entire nanopublication, such as the creator, creation date, and license.

HiCO	**CIDOC CRM (CRMinf)**
hico:InterpretationAct	crminf:I1_Argumentation
hico:InterpretationCriterion	crm:E55_Type
hico:hasInterpretationType	crm:P2_has_type
crm:P14_carried_out_by	crm:P14_carried_out_by
prov:startedAtTime	crm:P4_has_time-span
cito:citesAsEvidence	crm:P16_used_specific_object
prov:wasGeneratedBy	crminf:J2_concluded_that → crminf:I2_Belief → crminf:J4_that
hico:hasInterpretationCriterion	crm:P32_used_general_technique
hico:isExtractedFrom	crm:P70i_is_documented_in
prov:wasInfluencedBy	crm:P15_was_influenced_by

Table 6.1: HiCO classes and properties alignment with CIDOC CRM, importing the CRMinf module.

In expressing the subjectivity of information, such as the contexts of different hypotheses and the possible conflicts between them, it is necessary to focus on modelling the interpretation context in the provenance graph. The Historical Context Ontology (HiCO) is dedicated to expressing as LOD the context of a hermeneutic activity performed by a scholar in formulating a hypothe-

37. Carroll et al. (2005).

sis through the interpretation of sources [38]. HiCO is an extension of the PROV ontology, the standard model dedicated to modelling data provenance on the web [39]. Given its purpose, such an ontology is ideal for representing provenance graph information. HiCO revolves around one activity: the interpretation (hico:InterpretationAct). This activity represents the action of the scholar in formulating a hypothesis of which, among other types of information, we can express the type of interpretation (hico:hasInterpretationType), the criterion of interpretation (hico:hasInterpretationCriterion), the time frame in which the interpretation was carried out (prov:startedAtTime), the resources used (cito:citesAsEvidence), and the influence of other hypotheses (prov:wasInfluencedBy). To integrate HiCO into the Linked Art Data Model, we propose aligning HiCO and CIDOC CRM, as shown in Table 6.1. In aligning the two ontologies, it is necessary to use CRMinf, a CIDOC CRM module dedicated to modelling inference-making activities.[40] Specifically, CRMinf introduces the argumentation activity (crminf:I1_Argumentation) semantically comparable with HiCO's interpretation act (hico:InterpretationAct). In addition, the module allows for more granular modelling of the argumentation result, expressed in the assertion graph. While HiCO uses the PROV ontology property prov:wasGeneratedBy to indicate that the assertion graph resulted from an interpretation act, CRMinf uses an n-ary relation. As a result, the argumentation generates a belief (crminf:I2_Belief), which is, in turn, expressed by the assertion graph. As discussed in the next section on uncertainty, an n-ary relation allows one to assert information about the relation, which is impossible in a binary relation.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix crminf: <http://www.cidoc-crm.org/cidoc-crm/CRMinf/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix dct: <http://purl.org/dc/terms/> .

<https://example.com/nanopub/3/head> {
    <https://example.com/nanopub/3> a np:Nanopublication ;
        np:hasAssertion <https://example.com/nanopub/3/assertion_graph> ;
        np:hasProvenance <https://example.com/nanopub/3/provenance_graph> ;
        np:hasPublicationInfo <https://example.com/nanopub/3/pubinfo_graph> .
}

<https://example.com/nanopub/3/assertion_graph> {
    <https://example.com/activity/3> a crm:E7_Activity ;
        rdfs:label "Purchased by Knoedler & Co. from Louis Lion & Co. in February
            1957" ;
        crm:P2_has_type <http://vocab.getty.edu/aat/300055863> ;
        crm:P2_has_type <http://vocab.getty.edu/aat/300417642> ;
        crm:P4_has_time-span [ a crm:E52_Time-Span ;
            crm:P82a_begin_of_the_begin "1957-02-01T00:00:00Z" ;
            crm:P82b_end_of_the_end "1957-02-28T23:59:59Z" ] ;
```

38. Daquino and Tomasi (2015).
39. Moreau and Groth (2013).
40. https://www.cidoc-crm.org/crminf/ (accessed 2023-08-11).

```
        crm:P9_consists_of [ a crm:E8_Acquisition ;
            crm:P22_transferred_title_to [ a crm:E74_Group ;
                rdfs:label "Knoedler & Co." ] ;
            crm:P23_transferred_title_from [ a crm:E74_Group ;
                rdfs:label "Louis Lion & Co." ] ;
            crm:P24_transferred_title_of [ a crm:E22_Human-Made_Object ;
                rdfs:label "Cagnes" ] ] .
}

<https://example.com/nanopub/3/provenance_graph> {
    <https://example.com/argumentation/3> a crminf:I1_Argumentation ;
        crm:P2_has_type <http://vocab.getty.edu/aat/300444173> ;
        crm:P14_carried_out_by [ a crm:E74_Group ;
            rdfs:label "The Art Institute of Chicago" ] ;
        crm:P16_used_specific_object [ a crm:E33_Linguistic_Object ;
            rdfs:label "letter from Knoedler and Co., Apr. 8, 1975." ;
            crm:P2_has_type <http://vocab.getty.edu/aat/300026879> ;
            crm:P94i_was_created_by [ a crm:E65_Creation ;
                crm:P4_has_time-span [ a crm:E52_Time-Span ;
                    crm:P82a_begin_of_the_begin "1975-04-08T00:00:00Z" ;
                    crm:P82b_end_of_the_end "1975-04-08T23:59:59Z" ] ;
                crm:P14_carried_out_by [ a crm:E74_Group ;
                    rdfs:label "Knoedler & Co." ] ] ] ;
        crminf:J2_concluded_that [ a crminf:I2_Belief ;
            crminf:J4_that <https://example.com/nanopub/3/assertion_graph> ] .
}

<https://example.com/nanopub/3/pubinfo_graph> {
    <https://example.com/nanopub/3> dct:created "2023-08-11T16:31:08Z" ;
        dct:creator <https://orcid.org/0000-0002-7382-0187> ;
        dct:source <https://www.artic.edu/artworks/12402/cagnes> ;
        dct:license <https://creativecommons.org/publicdomain/zero/1.0/> .
}
```

Listing 6.6: Nanopublication, serialized in TriG format, of the purchase of the painting "Cagnes" by Knoedler & Co. from Louis Lion & Co. in February 1957.

Listing 6.6 shows the nanopublication of the provenance activity in which Knoedler & Co. purchased the painting "Cagnes" from Louis Lion & Co. in February 1957. The structure of the nanopublication is defined using the Nanopublication Ontology.[41] According to the note in the original provenance text, the assumption made by the Art Institute of Chicago is based on a "letter from Knoedler and Co., Apr. 8, 1975." The information is structured using HiCO's alignment to CIDOC CRM. The Getty AAT vocabulary is used to assign the entity types, as standard practice in Linked Art. In particular, the argumentation has the entity type "provenance remark" (aat:300444173), while the linguistic object used to formulate hypotheses has the entity type "letter" (aat:300026879). The metadata of the publication info graph, such as creation date, creator, source and license, are structured using properties from the Dublin Core Metadata Initiative (DCMI) Metadata Terms, a set of standardized metadata elements to describe digital resources.[42]

41. https://nanopub.net/guidelines/working_draft/ (accessed 2023-08-11).
42. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (accessed 2023-08-11).

6.7 Uncertainty

According to the literature, the terms "uncertainty" and "vagueness" are related, if not conflated. For example, in documenting evidence interpretation in archaeology using CIDOC CRM, Niccolucci and Hermon merge the concepts of vagueness and uncertainty in the same concept of reliability [43]. However, as discussed in previous sections, we distinguish between vagueness and uncertainty. The reliability of vague information lies in the accuracy of the data approximation. In contrast, the reliability of uncertain information lies in the probability of the data's factuality. In light of what was discussed in the previous section, we can therefore correlate the concept of uncertainty to subjectivity, as it expresses the degree of confidence in making a hypothesis. As we have already seen, AAM guidelines introduce the possibility of expressing uncertainty about a piece of information. Terms such as "possibly" or "probably" express levels of uncertainty depending on the provenance expert's degree of confidence. Regarding provenance LOD modelling, Art Tracks uses a boolean value to express certainty about some information [44], and Linked Art deliberately avoids adding this degree of complexity. Examining other attempts to model uncertainty in CIDOC CRM, in the previously mentioned work by Niccolucci and Hermon, the reliability of information is expressed through fuzzy logic, with a subjective coefficient ranging from 0 (not credible) to 1 (absolutely true) [45].

When analysing provenance texts, we noticed that uncertainty coincides with the patterns we identified when dealing with incompleteness. In the presence of a gap, hypotheses become less confident. Since uncertainty is related to making hypotheses, we could have multiple contradictory hypotheses of varying degrees of certainty to fill a given gap. For this reason, the nanopublication solution is effective since it can separate various hypotheses with their degrees of certainty in different assertion graphs, allowing for the coexistence of multiple hypotheses with varying degrees of certainty.

While modelling uncertainty as information associated with the act of interpreting has already been proven possible using HiCO [46], we take a different approach. We align HiCO with CIDOC CRM, particularly with CRMinf. The use of this module to model uncertainty in provenance data has already been hypothesised by Smith in analysing the potential of provenance LOD [47]. As we have seen when dealing with subjectivity modelling, our alignment involves describing the product of the crminf:I1_Argumentation expressed in the assertion graph with an n-ary rela-

43. Niccolucci and Hermon (2017).
44. Berg-Fulton, Newbury, and Snyder (2015).
45. Niccolucci and Hermon (2017).
46. Daquino, Pasqual, and Tomasi (2020).
47. Smith (2018).

tion through the crminf:I2_Belief entity. The argumentation does not generate an assertion graph but instead concludes with a belief that is, in turn, expressed in the assertion graph. Thus, we can link additional information to the crminf:I2_Belief entity, such as the crminf:I6_Belief_Value. The belief value represents the truth value of a belief produced by an argumentation. The CRMinf module requires determining a belief value scale with at least three values.

Staying true to the approach of the Linked Art Data Model, we delineate a belief value scale within Getty's AAT vocabulary. A crminf:I2_Belief can have as crminf:I2_Belief_Value: "true" (aat:300068765), "probably" (aat:300435721), "possibly" (aat:300435722), and "obsolete" (aat:300404908). The uncertainty terminology already used according to the AAM guidelines reoccurs through this new scale of values. In addition, we include the option of assuming the obsolescence of a given assumption. This option is fundamental to the data provenance of provenance data as it allows hypotheses to be discarded without eliminating them permanently, thus leaving them as evidence of the hermeneutic process concerning a given fact. What is obsolete for one provenance expert may not be obsolete according to another.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix crminf: <http://www.cidoc-crm.org/cidoc-crm/CRMinf/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix dct: <http://purl.org/dc/terms/> .

<https://example.com/nanopub/1/head> {
    <https://example.com/nanopub/1> a np:Nanopublication ;
        np:hasAssertion <https://example.com/nanopub/1/assertion_graph> ;
        np:hasProvenance <https://example.com/nanopub/1/provenance_graph> ;
        np:hasPublicationInfo <https://example.com/nanopub/1/pubinfo_graph> .
}

<https://example.com/nanopub/1/assertion_graph> {
   <https://example.com/activity/1> a crm:E7_Activity ;
        rdfs:label "Acquired by Galerie Kahnweiler from André Derain" ;
        crm:P2_has_type <http://vocab.getty.edu/aat/300055863> ;
        crm:P183_ends_before_the_start_of <https://example.com/activity/2> ;
        crm:P9_consists_of [ a crm:E8_Acquisition ;
            crm:P22_transferred_title_to [ a crm:E74_Group ;
                rdfs:label "Galerie Kahnweiler" ] ;
            crm:P23_transferred_title_from [ a crm:E21_Person ;
                rdfs:label "André Derain" ] ;
            crm:P24_transferred_title_of [ a crm:E22_Human-Made_Object ;
                rdfs:label "Cagnes" ] ] .
}

<https://example.com/nanopub/1/provenance_graph> {
    <https://example.com/argumentation/1> a crminf:I1_Argumentation ;
        crm:P2_has_type <http://vocab.getty.edu/aat/300444173> ;
        crm:P14_carried_out_by [ a crm:E74_Group ;
            rdfs:label "The Art Institute of Chicago" ] ;
        crminf:J2_concluded_that [ a crminf:I2_Belief ;
            crminf:J5_holds_to_be <http://vocab.getty.edu/aat/300435721> ;
            crminf:J4_that <https://example.com/nanopub/1/assertion_graph> ] .
```

```
}

<https://example.com/nanopub/1/pubinfo_graph> {
    <https://example.com/nanopub/1> dct:created "2023-08-11T16:35:12Z" ;
        dct:creator <https://orcid.org/0000-0002-7382-0187> ;
        dct:source <https://www.artic.edu/artworks/12402/cagnes> ;
        dct:license <https://creativecommons.org/publicdomain/zero/1.0/> .
}
```

Listing 6.7: Nanopublication, serialized in TriG format, of the probable acquisition of the painting "Cagnes" by Galerie Kahnweiler from the artist.

Listing 6.7 shows the nanopublication of the provenance event in which Galerie Kahnweiler probably acquired the painting "Cagnes" directly from the artist. In this case, since the level of certainty was expressed with the term "probably" in the original text, we can describe the value held by the belief generated by the argumentation, with the entity aat:300435721.

6.8 Discussion and Conclusion

The classification of VISU information differentiates among four distinct yet correlated types of information, each pertaining to a specific intervention by the provenance expert. Vagueness, subjectivity, and uncertainty represent information categories we depend on when provenance records are incomplete. In the absence of information, the provenance expert can fill the gap by approximating data, formulating hypotheses, and expressing varying degrees of confidence in reconstructing facts.

Although these terms are often used synonymously, the VISU classification distinguishes between vagueness and uncertainty. In the classification's context, vagueness pertains to the approximation of spatial and geographical information, thereby addressing the precision of the data. CIDOC CRM offers valuable elements for representing vague temporal information by modelling dates as time spans. Additionally, it enables the representation of vague spatial information by utilising the property crm:P189_approximates. Linked Art already includes such solutions. As we have discussed, to extend the modelling of temporal information approximation, we integrate the CRMgeo module. In this way, it is possible to describe a relation between a vague time span and its approximation using the property crmgeo:Q13_approximates.

By its nature, incompleteness is the only VISU information we cannot model in LOD. However, we can address incompleteness in analysis and hypothesis-making by carefully modelling the available information. Thanks to the event-based schema of CIDOC CRM and the application profile of Linked Art, we can formulate patterns for analysing incompleteness between and within different events. Thanks to

these patterns, on the one hand, conscious modelling of the available information is possible, for example, by always including the sender, the one who parts with the object in an event. On the other hand, identifying and analysing gaps in provenance records makes it possible to gain new insights into the state of provenance research on a large scale, helping to determine which artworks, collectors, and historical periods to prioritise in research efforts.

In the classification of VISU information, subjective and uncertain information is correlated. It requires a change of approach from what CIDOC CRM proposes since modelling the assertion context for each triple related to a single provenance event proves inefficient and repetitive. For this reason, we introduced a different approach by publishing provenance LOD as a nanopublication. The nanopublication of provenance LOD involves publishing each provenance event as an atomic unit, of which we describe the data provenance information, thus implementing the data provenance of provenance data. In this way, we model the information asserted and the context of the hypothesis, such as author, date, and sources used. In addition, we can include conflicting hypotheses by modelling them in distinct RDF graphs. In the literature, there are already ontologies suitable for modelling the context in which a hypothesis is formulated, such as HiCO. We, therefore, aligned HiCO with CIDOC CRM, using the CRMinf module to describe inference-making activities. Since uncertain information is related to the degree of confidence with which an expert makes a hypothesis, it is possible to model this uncertainty as LOD by qualifying the hypothesis-making context and implementing a belief value scale using terms from the Getty's AAT vocabulary.

Although compatible with the Linked Art Data Model, the solutions discussed for including VISU information in publishing provenance LOD should be considered an external module rather than a proposed extension. The representation and analysis of VISU information involve areas that, as we have seen, are deliberately outside the scope of Linked Art. One of the purposes of the Linked Art application profile is to make LOD information accessible and usable to institutional insiders. In this way, solutions such as nanopublications, although an established good practice in sharing scientific data compliant with FAIR principles, can be barriers to institutional practitioners. Indeed, the large volume of provenance texts from which we need to extract data would make it even more challenging to publish provenance LOD as nanopublications.

As discussed in this paper, VISU information is critical to the integrity of provenance LOD, and we cannot do without it in the name of simplicity. On the contrary, VISU information represents the complexity inherent in the effort to reconstruct historical events, as well as the contradictory assumptions that arise from the plurality of historical debates.

Balancing the effort of structuring provenance information as LOD with the qualitative care of VISU information requires a human-in-the-loop approach [48]. This means that, on the one hand, quantitative data structuring from provenance texts can be performed automatically by addressing natural language processing tasks through AI [49]. On the other hand, the qualitative curation of the data remains the responsibility of domain experts who can evaluate *de visu*, with their own eyes, the most ambiguous information.

Acknowledgments

The author would like to thank the three anonymous reviewers for their constructive feedback. I extend my gratitude to Marilena Daquino for valuable input and to Max Koss, Lynn Rother, and Liza Weber for their efforts in editing the article.

48. Rother, Mariani, and Koss (2024).
49. Rother, Mariani, and Koss (2023).

7. Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI

Bibliographic Information
Rother, Lynn, Fabio Mariani, and Max Koss. 22024. "Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI." In *Sammlungsforschung im digitalen Zeitalter: Chancen, Herausforderungen und Grenzen*, edited by Katharina Günther and Stefan Alschner, 93–103. Göttingen: Wallstein. https://doi.org/10.1515/jbwg-2023-0005.
CRediT Roles: Conceptualization, Methodology, Writing – Original Draft, Visualization

7.1 Introduction

The provenance of an artwork is the record of its ownership and socioeconomic custody changes. Traditionally, provenance was recorded as texts in ledgers. More recently, it has shifted to free text fields in collection management systems in museums. The researching and writing of provenances have long been characterized by a high degree of complexity as well as fuzziness, both of which are now emerging as points of concern for the digitization of provenance. To begin with, there is the very heterogeneous and incomplete historical archives that source provenances. While new archival materials continue to be found and made available to researchers, more often than not, some information remains missing, leading to gaps in provenances. At the same time, the information researchers have at their disposal, previously used to compile and record provenances, is open to interpretation. For each generation of researchers and scholars producing provenance, the historical sources underpinning provenances can be reinterpreted with new perspectives, leading to the updating, rewriting, and reinterpretation of provenances in light of contemporary concerns and conventions. Provenances are generally texts without single authorship, amalgamations of the work of multiple authors active at different moments in time, working to more or less scientific standards.

Today, museums are called upon to structure their provenance records and, eventually, transform them into provenance-linked open data (PLOD), which is based on standards for publishing information as structured data on the web. This enables the interlinking and reusability of any such information and, consequently, the enhancement of shared knowledge. For example, linked open data allows using

already vetted data of other institutions and querying datasets across institutions and repositories for complex research questions. While this process has yet to be widely adopted in the cultural heritage field, the advantages of such an approach are clear.

Today, museums are called upon to structure their provenance records and, eventually, transform them into provenance-linked open data (PLOD), which is based on standards for publishing information as structured data on the web. This enables the interlinking and reusability of any such information and, consequently, the enhancement of shared knowledge. For example, linked open data allows using already vetted data of other institutions and querying datasets across institutions and repositories for complex research questions. While this process has yet to be widely adopted in the cultural heritage field, the advantages of such an approach are clear.

A PLOD approach would allow the identification of objects unlawfully appropriated during contexts of injustice, such as Nazi-era expropriation or colonial looting, serving restitution and decolonization efforts. Large-scale analysis of ownership and socio-economic custody changes may also be relevant for other research questions in such fields as art history, anthropology, sociology, and social and economic history.[1] PLOD will also make provenance as a knowledge practice more accessible, if not more democratic. Where the writing of provenance is still predominantly tied to institutions that often function as gatekeepers of knowledge, a digital provenance approach will transform provenance into a distributed and collaborative knowledge practice. Such an undertaking can potentially counteract the various historical biases, for example, sources or subjective interpretation. Last but not least, the digital future of provenance allows tackling the issue of the authority of provenance, as it provides the possibility to publish the provenance of provenance, clearly identifying authors and sources of every bit of data contained in a given digital provenance.

Today, the digital transformation of provenance faces two interrelated questions addressed in this paper: First, how can vast quantities of provenance texts be transformed into high-quality data, and second, what would such a process need to look like for the benefits of digital transformation to outweigh the efforts and costs required? This paper focuses on the use of artificial intelligence in the transformation of provenance. It lays out which AI techniques are particularly suited for the process and expands on the limitations of a technology-only approach. Indeed, given the complexity and fuzziness mentioned at the outset, it will become clear that the production of digital provenance will continue to require expert knowledge to make judgment calls where the machine cannot.

1. Rother, Koss, and Mariani (2022); Rother, Mariani, and Koss (2023).

7.2 From Provenance Texts to Provenance Data

A look at the current state of provenances, especially in the United States of America, reveals that, ever since the publication of the Washington Conference Principles on Nazi-Confiscated Art in 1998, there have been rigorous research efforts to record the provenance of hundreds of thousands of works across numerous institutions.[2] Although no shared standard was established, the publication of the American Alliance

Offered by the artist as a New Year's gift to "Mme X." [1] M and Mme Jules Féral, Paris, by 1932 until at least 1938.[2] Possibly (Galerie Charpentier, Paris) in 1951. [3] Capt. Edward H. Molyneux [1891-1974], Paris, by 1952;[4] sold 15 August 1955 to Ailsa Mellon Bruce [1901-1969], New York; bequest 1970 to NGA.

[1]According to Paul Jamot and Georges Wildenstein, *Manet*, Paris, 1932, no. 508.

[2]Lent by Féral to exhibitions in London in 1932 and Amsterdam in 1938. Eugène and Jules Féral [died c. 1949] acted as experts at sales at Hôtel Drout and elsewhere, the former betweeen c. 1876-1901 and the latter in the 1920s.

[3] A 1949 sale of objects from Jules Féral's collection held at the Galerie Charpentier did not include the NGA picture. However the picture was included in an 1951 exhibition held at Charpentier, with no owner listed, and was probably sold to Charpentier by Mme Féral by that time.

[4]Lent by Molyneux to the National Gallery of Art in 1952.

Figure 7.1: The provenance of Édouard Manet's *Flowers in a Crystal Vase*, as published on the website of the National Gallery of Art, Washington, DC (https://www.nga.gov/collection/art-object-page.52181.html, 3 February 2023).

of Museums (AAM) guidelines on how to write provenance has resulted in many U. S. institutions adopting a similar approach to documenting provenance.[3] In addition, institutions outside the U. S. have adopted comparable guidelines, proposed by the International Foundation for Art Research (IFAR) recommended.[4] Figure 7.1 shows the provenance of Édouard Manet's *Flowers in a Crystal Vase*, as published on the National Gallery of Art website in Washington, DC.

Recording provenance according to the AAM guidelines involves compiling the chain of provenance events in chronological order up to the acquisition by the current owner. In our example, the first event in the history of any object – the creation of the painting – is omitted. The first recorded event is the gift of the object by its creator Édouard Manet to an anonymous »Mme X«; the last recorded event is the bequest by Alisa Mellon Bruce to the National Gallery of Art, the work's current owner. Each event corresponds to a sentence in the text divided from the previous one by a semicolon when the transfer between the parties was direct. »If a direct

2. U.S. Department of State, Office of the Special Envoy for Holocaust Issues (1998).
3. Yeide, Akinsha, and Walsh (2001).
4. International Foundation for Art Research (IFAR) (2023).

transfer did not occur or is not known to have occurred,« then the AAM guidelines suggest dividing the events by a period.[5] For example, this type of gap appears between the ownership of »Mme X« and that of »M and Mme Jules Féral,« the two parties named in the first and second event of the provenance text in Figure 7.1. Footnotes should be used to document historical sources and clarify uncertain events. Terms such as »probably« and »possibly« can be used to indicate hypotheses about events not entirely accounted for. For example, the provenance text in Figure 7.1 indicates that the ownership of the Galerie Charpentier is considered possible; a footnote explains the reason for this uncertainty, namely, that this work was on view at the gallery in 1951, which does not necessarily imply ownership.

While many institutions have now adopted the AAM guidelines to record the provenance of thousands of objects, a stricter standardization of writing provenance has yet to be achieved. In light of this continued heterogeneity, we cannot consider provenance compiled according to the AAM guidelines to be structured, machine-readable knowledge. At the same time, AAM-compliant provenances can be considered a foundation for employing advanced knowledge extraction techniques to streamline the process of creating and publishing PLOD.

7.3 The Role of AI in Structuring Provenance Texts

Although provenance texts written according to the AAM guidelines are unstructured, one can use artificial intelligence (AI) to automatically extract information from the text and structure it in a machine-readable format. Because this is a challenge involving texts, the research area for this process is Natural Language Processing (NLP), which develops computational methods that automatically process human language to solve specific problems. One such problem is the extraction of events from a text. In our case, the chronological nature, lining up event after event, helps to extract information from provenance texts.

We have successfully experimented with event extraction from provenancetexts by approaching the problem with two NLP tasks.[6] The first task is sentence boundary detection (or disambiguation, SBD). SBD aims to identify and disambiguate punctuation marks that separate sentences in a text. As discussed earlier, events in a provenance text may be separated by a semicolon or a period, depending on whether the change of ownership is direct or not. However, characters such as a period can be ambiguous. For example, a period indicating an abbreviation may or may not mark the end of an event.

Once we have divided provenance texts into events based on punctuation, we im-

5. Yeide, Akinsha, and Walsh (2001), p. 33.

6. The experiment is discussed in Rother, Mariani, and Koss (2023).

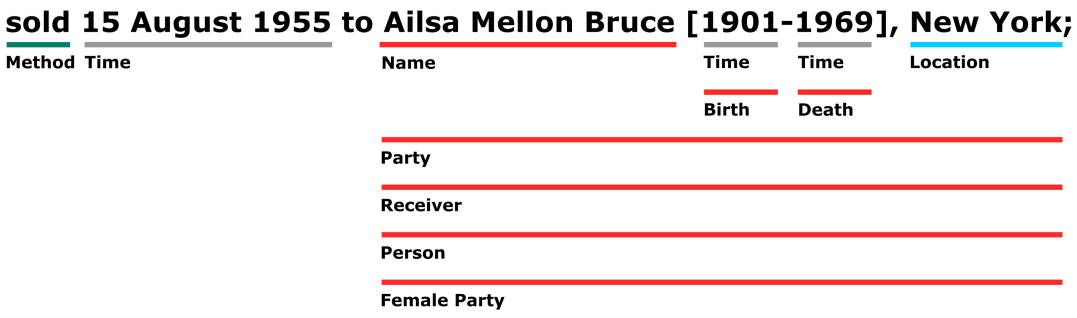

Figure 7.2: Example of span categorization. The event is taken from the provenance of Édouard Manet's *Flowers in a Crystal Vase*, as published on the website of the National Gallery of Art, Washington, DC (https://www.nga.gov/collection/art-object-page.52181.html,3February2023).

plement the second task identified for event extraction: span categorization (or classification). This task involves identifying and categorizing portions of texts (spans) by assigning them to a category.[7] We can apply span categorization to any event previously extracted by SBD. Figure 7.2 shows an example of span categorization applied to a provenance event extracted from the provenance of Édouard Manet's *Flowers in a Crystal Vase* discussed previously. The different spans of the text assigned to a category are highlighted. For example, the span »sold« corresponds to the method of transfer used in the event, so »method« is the category assigned to the span. Similarly, the text portion »15 August 1955« represents the »time« when the event occurred.

A distinctive feature of the span categorization task is that spans may overlap. For example, the span »Ailsa Mellon Bruce [1901-1969], New York« can be categorized as »party,« in the form of a »person,« and with its role as »receiver,« in other words, the party who receives the object in this event. In addition, we can indicate the presumed gender of the party by assigning the category »female party.« Finally, within the span, we can find additional spans. In the event annotated in Figure 7.2, the span »Ailsa Mellon Bruce« is the »name« of the party, »1901« is the date of birth, to which we assign both the »birth« and »time« categories, while »1969« is the date of death, to which we instead assign both the »death« and »time« categories. Lastly, the span »New York« is the party's location.

We trained two deep learning models to address the SBD task and the span categorization task through AI. Both models performed remarkably well in our experiment on the provenance texts published by the Art Institute of Chicago. The

7. For the documentation of the provenance specific annotation scheme, see Mariani, Rother, and Koss (2023b).

SBD model achieved an F1 score of 0.99, while the span categorization model scored an F1 score of 0.94.[8] Given these results, we can automatically extract information from provenance texts written according to the AAM guidelines with high degrees of accuracy.[9]

The use of AI to extract knowledge from provenance texts reveals promising scenarios for the fast publication of large amounts of data. However, introducing a heuristic process, such as deep learning models, in dealing with historical information requires a critical awareness of the technology used and how it shapes its results. Indeed, behind the output of the AI's black box lie substantial human interventions that influence the heuristic process. Therefore, we must not be tempted by AI's lure of objectivity to accept its results but rather maintain a critical approach in supervising them.[10]

7.4 Interpreting Strings, Weaving Threads

Despite the satisfactory results that AI models achieve in extracting information from provenance texts, the production of PLOD cannot be considered complete with these computational methods. In fact, two main issues require human intervention when using AI to structure provenance texts on a large scale: First, despite good test performance, techniques such as span categorization are not error-free. Although a low percentage of errors is not statistically significant when analyzing large amounts of data (distant reading), each error becomes noteworthy when analyzing individual provenances published in LOD (close reading).[11] Given the accuracy of the tests, ignoring the low error rate of AI can be a concern should any of these errors involve legally and ethically problematic provenance events. Consider, for example, potential errors in data extraction for events involving looting or confiscation. Such neglect would go against the principles of transparency and accountability that museums aim to uphold, not least by publishing PLOD. Therefore, it is essential to always monitor the output of AI models to prevent the publication of erroneous historical information.

The second reason for human intervention in AI-extracted data concerns certain types of historical information that require expert interpretation to be recorded and published in LOD in a manner commensurate with their complexity. One can divide

8. The F1 score is a measure to assess the accuracy of an AI model. Its value is between 0 and 1.

9. We refer to Rother, Mariani, and Koss (2023) for a comprehensive description of the models' implementation and training.

10. The »lure of objectivity« of computational methods is one of the five challenges of the digital humanities presented in Rieder and Röhle (2012).

11. For a focus on the concepts of close and distant reading in art history, see: Klinke (2020).

this information into four categories: vague, incomplete, subjective, and uncertain. To emphasize that this information requires human supervision, we grouped the four categories under the acronym VISU, from the Latin *de visu*, which translates as ›with your own eyes‹.[12] Vague information concerns approximations of spatial and temporal data, examples of which are expressions such as »near Florence« or »by 1932.« The expert's task in such cases is to evaluate the vague information in the provenance data and reconstruct the information as accurately as possible.

Incompleteness refers to the lack of provenance information, which may occur as a gap when the transfer between two owners is not known to have been direct. As indicated earlier, the AAM guidelines recommend recording such gaps by separating the events with a period.[13] In dealing with such gaps, experts may formulate new hypotheses for what may have occurred by interpreting available historical sources or analyzing already structured provenance data. Indeed, data analysis can support this process, revealing patterns and insights that can help suggest new hypotheses.[14] However, this machine intervention in the historian's hermeneutic approach should not be understood as an automatic process in which the machine generates new hypotheses. Rather, data analysis becomes a new research tool for the historian, albeit not exempt from the source criticism required by historiographical methods.[15] Finally, further incomplete information may exist in the components of a provenance event, such as biases in the representation of female parties or minorities. Indeed, it is not uncommon to find female parties in the text recorded with the husband's surname or even with the husband's first and last name, as in the example of »Mme Jules Féral.« In this case, the bias is explicit in the text, and the information is propagated in the data without appropriate human intervention.

Writing a provenance text and supervising the data extracted by AI are both hermeneutic processes that call on domain experts to formulate hypotheses. Individual scholars create information that – while following scientific criteria – is subjective insofar as any use of historical sources is an act of individual interpretation by the domain expert. Any provenance event is recorded following such standard historical practice of interpreting sources; however, as discussed in the Introduction, current provenance writing usually neglects to identify authorship and, to a lesser extent, sources. The intervention of a domain expert on the data extracted by AI enables the reconstruction of the history of an object parallel to the history of documenting its provenance, by whom it was conducted, when, and with what sources.

Documenting this information means recording, in addition to the provenance,

12. Mariani (2023).

13. Yeide, Akinsha, and Walsh (2001).

14. An experiment on the use digital methods and analysis for reconstructing missing art market information is presented in: Lincoln and Ginhoven (2018).

15. Zundert (2015).

the provenance of provenance.[16] This approach enables a further step in the process of professionalizing and raising the scientific profile of provenance research. In particular, the provenance of provenance meets the need for transparency and accountability in museum documentation. It is imperative to record the author, date, and sources used to formulate each piece of information. As discussed earlier, AI and data analysis can assist historians in producing and enhancing provenance data without replacing their role as experts and critics. The provenance of provenance also records the use of computational methods to ensure a transparent account of how provenance data were generated and by whom (or by what AI). Finally, documenting the provenance of provenance enables recording contradictory hypotheses. For example, two historians may disagree on the interpretation of a particular source, drawing different conclusions regarding the life trajectory of an object. In this way, institutions can publish provenance data without discarding one hypothesis in favor of another, recording both hypotheses and documenting their relative provenance of provenance.

Lastly, uncertainty relates to the interpretability of provenance information discussed above. A scholar most certainly has varying degrees of confidence in formulating different hypotheses, which the AAM guidelines also reflect. They suggest using terms such as »possibly« (more confident) or »probably« (less confident), depending on the degree of certainty with which the statement can be made.[17] Historians must therefore evaluate the uncertain information after structuring the provenance data using AI. The certainty of a hypothesis, which is related to the interpretability of provenance information, is thus additional information to be included in the provenance of provenance.

The example of the provenance of Flowers in a Crystal Vase is enlightening in demonstrating the importance of human supervision of AI-extracted data, mainly VISU information. Regarding vague information, we note several approximations of dates. For example, the text records that Capt. Edward H. Molyneux had acquired the work »by 1952.« In this case, we need to turn to the last known date before the event in question, which delimits a time interval for locating the vague acquisition date. Since the earlier date is »1951,« we can infer that Molyneux acquired the work between 1951 and 1952. Nevertheless, to validate this inference, we must first consider the incomplete, subjective, and uncertain information in the text: Some of the names of the previous owners are unknown, and there is no record of the name of »Mme X« nor of Jules Féral's spouse, represented as »Mme Jules Féral.« This incomplete information coincides with gaps in the provenance. Indeed, we do not know what happened to the painting once it was given to »Mme X.« The

16. Al-Eryani, Bucher, and Rühle (2018); Huemer (2020); Newbury and Lippincott (2019).
17. Yeide, Akinsha, and Walsh (2001), p. 33.

object, created circa 1882, reappears 50 years later as the property of Jules Féral and his anonymized spouse, after which a further gap occurs. In trying to fill this gap, the editor of the provenance text formulated the hypothesis that the Galerie Charpentier might have owned the object. In fact, through a note in the provenance text, we learn that the gallery presented the painting in a 1951 exhibition. The editor of the provenance speculates that Jules Féral's spouse sold the object to Galerie Charpentier by that year. Since the AI models we introduced previously do not involve extracting knowledge from notes, this valuable information would have been lost without human intervention. However, in this instance, in reconstructing the provenance of provenance and assessing the reliability of such a hypothesis, we must note the omission of an appropriate reference to the historical sources used to formulate a hypothesis.

Through the term »possibly,« the provenance text indicates the uncertainty of the hypothesis that Galerie Charpentier owned the object. The historian supervising the extracted data may accept this uncertainty or engage in further research. For example, an archival search might turn up new documents related to purchases made by Galerie Charpentier, filling the gap. Further help for the historian might come from analyzing other provenance data. For example, one could analyze the data and identify the main parties who sold objects to Galerie Charpentier, particularly whether there were other instances of Jules Féral's widow selling objects. However, we need an appropriate record of the parties' names to facilitate this analysis. In the case of »Mme Jules Féral,« it is necessary to record this person's name for her proper historical representation and consistency across provenance data. Indeed, we might have cases where »Mme Jules Féral« is recorded as »Mme Féral« or »Mrs. Jules Féral.« The analysis could be even more arduous if, for instance, Jules Féral had more than one wife.

Based on the considerations and decisions we introduced, a historian could finally assess whether Molyneux acquired the work between 1951 and 1952, thus accepting that Galerie Charpentier acquired the object in 1951. Otherwise, discarding this hypothesis, one might infer that Molyneux could have acquired the object between 1938, the last date when Jules Féral and his spouse owned the object with any certainty, and 1952 when there is the certainty that the object was already in his possession.

7.5 Conclusion

This paper discusses how the provenance of museum objects can be (semiautomatically) structured and published. By leveraging the power of artificial intelligence, in particular deep learning models, we can process large quantities of data relatively

quickly. Nonetheless, when dealing with the qualitative nature of much historical information, we found it necessary to consider human intervention to monitor the results of AI. This approach is essential for error correction and appropriately handling VISU information. Thus, we developed a two-step, two-speed digitization process: fast digitization, enabled by AI and its quantitative benefits, followed by and combined with slow digitization, performed by domain experts, evaluating and ensuring the scientific quality of the data.

The domain expert is not replaced by technology but becomes an essential factor in the digitization process. The historian need not participate in the time-consuming data structuring process, which AI can successfully perform. Instead, the expert is involved in critiquing sources and formulating historical hypotheses. This demarcates a precise boundary between the tasks delegated to AI and the tasks appropriate for domain experts. After all, history is written by humans, not machines.

8. PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data

Bibliographic Information
Mariani, Fabio. 2025. "PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data." *Zeitschrift für digitale Geisteswissenschaften*, no. 10. https://doi.org/10.17175/2025_012

Abstract
Provenance documents the history of cultural objects, providing evidence of authenticity and ownership, and ensuring ethical accountability. Publishing provenance as Linked Open Data (LOD) enhances accessibility, interoperability, and large-scale analysis, addressing the limitations of textual records. However, the transformation of provenance information into structured LOD remains constrained by labour-intensive extraction processes and technical barriers to adoption. This paper introduces PROV-A (the Provenance App), a web-based tool designed to streamline the creation and publication of provenance information as LOD. PROV-A facilitates the integration of external automated data extraction workflows, such as natural language processing (NLP), with human validation, balancing efficiency with scholarly rigour. A case study using provenance records from the Art Institute of Chicago illustrates how PROV-A enables users to refine automatically extracted data, preserve historical ambiguities, and support provenance analysis. By lowering technical barriers and fostering a human-in-the-loop approach, PROV-A improves the scalability and accuracy of provenance research, making LOD more accessible to cultural institutions.

8.1 Introduction

Provenance, defined as the history of an object through its creation and subsequent changes in ownership and custody, provides crucial documentation that offers insight into an object's authenticity, artistic value, and historical significance. In addition to these applications, provenance is necessary to address ethical and le-

gal issues related to cultural heritage, particularly in contexts of systemic injustice such as confiscations by totalitarian regimes and colonial-era looting. The establishment of ethical standards for the restitution of cultural property was first initiated by the Washington Principles on Nazi-Confiscated Art (1998), which emphasised the role of provenance research in the proactive return of cultural property to its rightful owners.[1] After this, and fostered by the publication of the American Alliance of Museums (AAM) guidelines, provenance research has achieved a scientific and methodological rigour that has encouraged the compilation and publication of provenance records by numerous institutions in recent years.[2]

Despite its potential to enhance transparency and institutional accountability, provenance research faces significant challenges that hinder its full implementation. The process of recording the ownership and custody history of an artefact requires specialised archival and intellectual work, resulting often in provenance records with fragmented and incomplete information. Furthermore, institutions typically compile provenance records as texts, which impacts their ability to systematically maintain information and inhibits the exchange of data, thus impeding large-scale provenance analysis. The publication of provenance information as linked open data (LOD) has been identified as a solution to the issues of data siloing.[3] Indeed, this approach adheres to the FAIR principles, ensuring that information is findable, accessible on the web, while remaining interoperable and reusable across institutions.[4] However, extracting information from texts and transforming it into LOD requires substantial resources and technical expertise, posing a barrier to entry for institutions.

Recent approaches use natural language processing (NLP) techniques with artificial intelligence (AI) to streamline knowledge extraction from provenance texts.[5] While these methods have proven effective, the complexity of historical records and their need for interpretation also expose their limitations. Without human oversight, an automatic approach can perpetuate errors, inconsistencies, and biases present in the original records. This highlights the need for a human-in-the-loop approach, where experts actively validate, interpret, and correct AI-driven outputs. By integrating the efficiency of AI with human judgement, this approach ensures the accuracy and reliability of provenance data, which is crucial given its scientific and ethical significance.

This paper introduces PROV-A (the Provenance App), an application designed

1. U.S. Department of State, Office of the Special Envoy for Holocaust Issues (1998).

2. Yeide, Akinsha, and Walsh (2001).

3. Smith (2018); Newbury and Lippincott (2019); Luther (2020); Rother, Koss, and Mariani (2022).

4. Wilkinson et al. (2016).

5. Rother, Mariani, and Koss (2023).

to lower barriers to creating and publishing provenance records in LOD.[6] It begins with an overview of the scientific background of the interface, followed by a detailed discussion of its core functionalities and design rationale. A case study then demonstrates how PROV-A helps users enhance AI-extracted information by incorporating human intellectual input while preserving the ambiguities and gaps inherent in historical records. Ultimately, the case study highlights how PROV-A can help transform digitisation challenges—such as missing information and approximate dates—into valuable resources for historical narratives.

8.2 Background

The potential advantages in publishing provenance records as LOD have motivated researchers to develop projects, tools, and strategies. A central aspect of these efforts is organising data around a common standard. In the cultural heritage domain, the reference standard is the CIDOC Conceptual Reference Model (CIDOC-CRM), an ISO standard (ISO 21127) developed by the International Documentation Committee (CIDOC) within the International Council of Museums (ICOM).[7] Recognised as an ISO standard in 2006, CIDOC-CRM provides an event-based ontology that structures relationships between objects, people, places, and events in cultural heritage. Its event-based design is particularly well-suited to provenance modelling, as it allows the history of an object to be conceptualised as a sequence of events that define its creation and subsequent changes in ownership or custody.

The Art Tracks project, a pioneering initiative by the Carnegie Museum of Art (CMOA), was one of the first projects to apply CIDOC-CRM for modelling provenance as LOD.[8] The project, conducted between 2014 and 2017, aimed to establish a standard for compiling provenance records as text—the CMOA Digital Provenance Standard—which could then be automatically converted into LOD in accordance with the CIDOC-CRM data structure. The Elysa Tool software, developed during the project, facilitated this conversion process by enabling users to structure and potentially enrich provenance texts compiled according to the CMOA standard before generating LOD.[9]

More recently, Linked Art emerged as a prominent application profile for mod-

6. PROV-A is accessible at https://prov-a.github.io. The project's source code, development updates, and documentation can be found in the PROV-A GitHub repository (https://github.com/prov-a/prov-a.github.io).

7. Doerr (2003). CIDOC-CRM (https://cidoc-crm.org/).

8. Berg-Fulton, Newbury, and Snyder (2015); Newbury (2017).

9. Berg-Fulton, Newbury, and Snyder (2015); Newbury (2017). Elysa Tool (https://github.com/arttracks/elysa).

elling provenance data as LOD.[10] Unlike the broad scope of CIDOC-CRM, Linked Art specifically addresses the needs of museums by implementing data modelling patterns that facilitate and standardise the representation of common entities within the museum domain. To ensure terminological consistency, Linked Art uses controlled vocabularies, such as the Getty Vocabularies, which include the Art & Architecture Thesaurus (AAT), the Thesaurus of Geographic Names (TGN), and the Union List of Artist Names (ULAN).[11] These vocabularies provide consistent terminology for artistic concepts, geographic locations, and personal identities, thereby enhancing the reliability of data linking and cross-referencing across datasets.

Despite existing standards for publishing provenance as LOD, institutions face significant challenges in implementing this transition on a large scale due to the required resources and expertise. However, the adoption of AAM guidelines, which advocate for a non-standardised yet systematic approach to compiling provenance texts, has made it possible to experiment with NLP techniques for event extraction.[12] The experiment demonstrated how deep learning models can parse provenance texts into distinct events and extract relevant data for each, including acquisition methods, dates, involved parties, their roles, and biographical information.[13]

The encouraging results of NLP experiments must be understood in the context of the complexity and need for interpretation of historical documents, such as provenance records. Indeed, tracing the history of an object requires intellectual effort, producing what has been defined as VISU information—vague (e.g., date approximations), incomplete (e.g., gaps and missing details in records), subjective (e.g., the interpretive context of historical research), and uncertain (e.g., the degree of confidence in formulating hypotheses).[14] While automatic knowledge extraction can help identify some of this information (e.g., extracting vague dates), only human intervention ensures the preservation of VISU information throughout the LOD creation process.

To balance the quantity and quality of information extracted and structured in LOD, a hybrid approach combining AI and human expertise has been proposed.[15] AI enables fast digitisation—the rapid extraction of data at scale—while human intervention becomes essential during slow digitisation. In this phase, domain experts ensure the scientific accuracy of the data, contextualise sources, and formulate historical interpretations. This dual-speed approach allows AI to handle data processing and extract core historical information, while experts focus on tasks requiring

10. Newbury (2018). Linked Art (https://linked.art/).
11. Harpring (2010).
12. Rother, Mariani, and Koss (2023); Mariani, Rother, and Koss (2023b).
13. Mariani, Rother, and Koss (2023b).
14. Mariani (2023).
15. Rother, Mariani, and Koss (2024).

contextual knowledge and historical analysis.

Preserving VISU information is critical not only in extracting knowledge from texts but also in modelling LOD. While CIDOC-CRM provides a standard for structuring information in the context of cultural heritage, VISU information requires more complex data structures. To capture the interpretive context of a provenance record, it has been proposed to model each event as a nanopublication.[16] A nanopublication is a compact, self-contained unit of information designed for representation as LOD, consisting of an assertion (i.e., a historical event) enriched with metadata specifying its data provenance (i.e., the authors and sources involved) and contextual details about the nanopublication itself (e.g., its editor, publication date, and license).[17]

Structuring artefact provenance as a nanopublication facilitates access to the historical details of an artefact's biography by recording changes in ownership and custody. In addition, nanopublications document the data provenance of the record, that is, the archival sources consulted, the responsible agents, and interpretive decisions on which the record is based. This dual focus makes explicit both the artefact's biography and the evidential foundations of the research process, supporting more transparent and critically grounded interpretations. Work on polyvocal knowledge modelling in ethnographic heritage reflects this concern, recording the data provenance of each provenance record to preserve and contextualise multiple, potentially conflicting, interpretations of artefacts' biographies.[18]

In recent years, a variety of platforms have emerged to support the creation and use of LOD in cultural heritage while also documenting data provenance. ResearchSpace is an open-source scholarly workspace that enables researchers to structure, annotate, and publish cultural heritage data. It structures information as LOD following CIDOC-CRM while recording data provenance. Although widely applicable, tailoring the platform to specific projects often requires advanced technical expertise, which limits its accessibility for non-specialists.[19] To lower this barrier, Crowdsourcing Linked Entities via web Form (CLEF) provides a form-based workflow for collaborative LOD creation, supporting contributors with predefined templates and systematically capturing data provenance. Its design makes LOD accessible to non-specialists and effective for building new datasets in a collaborative setting.[20] Similarly, HERITRACE provides a semantic data editor developed for cultural heritage professionals. It preserves detailed data provenance, including change tracking and versioning. Nonetheless, configuring or adapting data models

16. Mariani (2023).
17. Kuhn et al. (2018).
18. Shoilee, Boer, and Ossenbruggen (2023).
19. Oldman and Tanase (2018).
20. Daquino et al. (2023).

still requires technical expertise.[21]

Existing platforms highlight the need to make LOD usable while ensuring rigorous data provenance, yet immediate accessibility for non-specialists remains a central challenge. In this context, PROV-A is deliberately scoped to the provenance of cultural heritage artefacts, adopting a narrower domain focus that enables a lightweight, web-based implementation, in contrast to platforms that must accommodate broader and more heterogeneous cultural heritage data. By adopting a web-based design, PROV-A lowers technical barriers and enables both specialists and non-specialists to structure and publish provenance records as LOD. In addition, it structures provenance records as nanopublications, supporting the documentation of data provenance alongside artefact provenance. The following chapter examines the design principles and technological choices that shape PROV-A.

8.3 PROV-A Design Principles

This section provides a detailed overview of the operational workflow, technological architecture, and the core principles that underpin the development of PROV-A. The primary objective of the interface design is to support the organisation of provenance information as LOD while prioritising accessibility and usability. This approach deliberately reduces technical barriers to ensure the interface is accessible to a wide range of users. By designing for both technical and non-technical users, the goal is to facilitate seamless interaction with LOD and encourage its adoption in various research and institutional contexts.

As a client-side application, PROV-A runs entirely within the user's web browser, avoiding the complexities of server-side architectures, such as hosting costs, database management, and system maintenance. This decentralised design allows users to maintain full control over their data and is beneficial in resource-limited contexts where both technical and financial resources may be constrained. PROV-A was developed using web technologies, including HTML5 and JavaScript, and incorporates Bootstrap 5, a front-end framework, to ensure a responsive design and consistent functionality across a wide range of devices and modern web browsers.[22]

The workflow of PROV-A consists of three sequential steps: initialise project, structure data, and generate LOD. The first step requires users to set up a new project by filling in a form to enter project settings. To begin, users must have an ORCID (Open Researcher and Contributor ID), a persistent digital identifier that ensures proper attribution of authorship.[23] Next, users select a data license for

21. Massari and Peroni (2025).
22. Bootstrap (https://getbootstrap.com/).
23. ORCID (https://orcid.org/).

data generated within the project. Users can choose from three available Creative Commons licenses: CC0 (Public Domain), CC BY (Attribution), and CC BY-SA (Attribution-ShareAlike).[24] CC0 allows for unrestricted use without attribution, while CC BY requires users to credit the original author. CC BY-SA extends this by mandating that derivative works use the same open license. These licenses align with open data principles and promote broad data sharing.[25] Finally, users must enter a project URI (Uniform Resource Identifier), which uniquely identifies LOD produced within PROV-A.

In initialising the project, users need to input metadata related to the artefacts to document. Such metadata is organised in a table with nine columns, including title, author, institution, URL, creation date, medium, accession number, provenance, and credit line. Users can either enter metadata manually into the table or upload it as a CSV file. CSV (Comma-Separated Values) is an open, non-proprietary format for tabular data, supported by various software tools, including free text editors and spreadsheets.

After populating the metadata table, users can download it as a CSV file for backup or future use. Upon completion, users can initiate the project, triggering the generation of a JSON (JavaScript Object Notation) file. This file encapsulates all entered data, including ORCID, license, URI, and artefact metadata, and formats it into a structure suitable for web applications, relieving users of manual formatting tasks. At this stage, experienced users can preprocess the JSON data. For example, they can use external AI models to automatically extract knowledge from provenance texts, structure the information according to the PROV-A JSON schema, and upload it into PROV-A for further supervision and refinement.[26]

The second step in the PROV-A workflow involves structuring data. Users upload the JSON file they generated in the previous phase. The interface organises the workspace into three columns to facilitate navigation and interaction (fig. 8.1). The left column displays artefact metadata alongside tools for filtering by attributes such as institution, author, title, or keywords in the provenance text. The right column presents a modular list of provenance activities, allowing users to document events in the artefact's history. Each activity corresponds to a specific provenance event—such as creation, acquisition, or transfer of ownership—and users can add, remove, or rearrange them chronologically using drag-and-drop functionality. The central column contains a form designed to capture historical data about each event, abstracting the CIDOC-CRM data structure. For example, the form includes a dropdown menu for

24. Creative Commons (https://creativecommons.org/).

25. Open Definition (https://opendefinition.org/).

26. The PROV-A JSON schema (https://github.com/prov-a/prov-a.github.io/blob/main/test/test_JSON/schema.json) defines the structure and validation rules for representing provenance data in JSON format, ensuring compliance with the PROV-A interface.

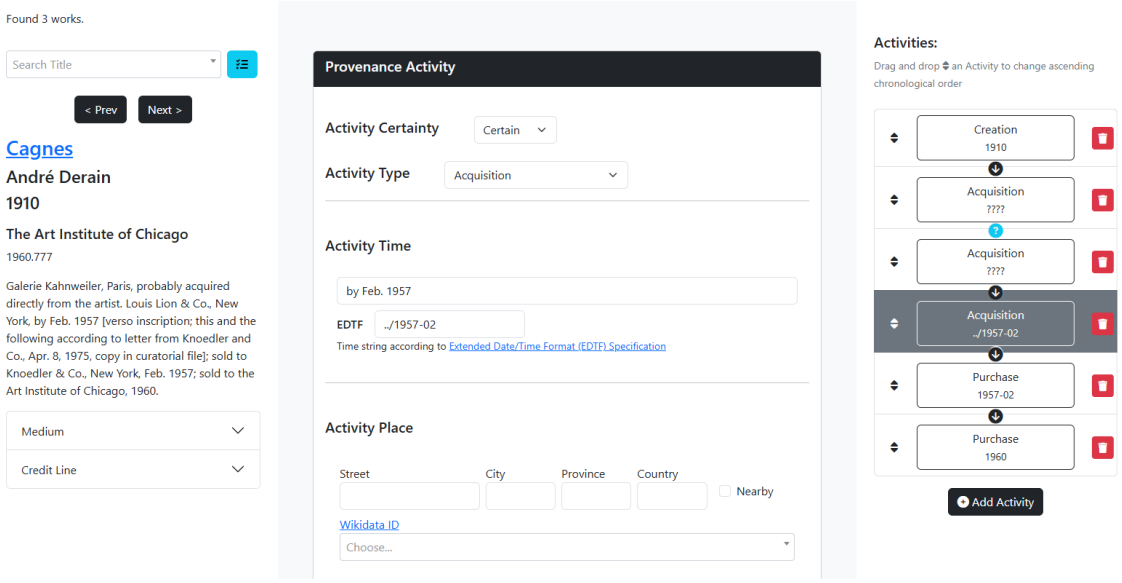

Figure 8.1: PROV-A data structuring interface.

selecting the type of activity, with options like artefact creation, auction, purchase, or looting event. These options align with Getty AAT terminology.

Each provenance event is a spatiotemporal entity that requires detailed temporal and spatial data. To provide flexibility in representing time, PROV-A allows users to enter temporal information as free text, accommodating a range of expressions, including vague information such as "circa 1945" or "between 1856 and 1870." The interface automatically converts these textual inputs into the Extended Date/Time Format (EDTF), a machine-readable system based on ISO 8601 that handles vague and imprecise time references.[27] For spatial data, users can specify address elements—such as street, city, province, and country—and optionally mark locations as approximate with a checkbox. In addition, the interface integrates with Wikidata and suggests potential matches for entered locations to help disambiguate entities.

PROV-A documents all parties involved in each activity, allowing users to specify roles such as sender, receiver, and agent. The sender is the party from whom the artefact departs, the receiver is the party obtaining it, and the agent is the individual or entity responsible for carrying out or mediating the event. Users can enter detailed data for each party, including distinctions between individuals and groups, onomastic details (e.g., names, titles), biographical information (e.g., birth and death dates), relationships, and location data.

To enrich party records, PROV-A cross-references entity names with two external repositories: Wikidata and the Getty ULAN. This entity linking feature disambiguates parties and adds supplementary information to their profiles. Once entered,

27. Extended Date/Time Format (EDTF) Specification (https://www.loc.gov/standards/datetime/).

party information is stored for reuse across multiple provenance activities, eliminating the need for redundant data entry. For example, if an individual or organisation is involved in several activities (e.g., a collector acquiring multiple artefacts), users can retrieve and link them to each relevant event, ensuring data consistency and saving time.

In addition to documenting provenance activities, PROV-A enables users to record the context in which provenance information was created. This includes entering details about the author of a historical assertion, the date, and the sources consulted. The interface features a confidence scale—certain, probable, possible, and obsolete—that allows users to indicate the certainty of each assertion. This scale aligns with Getty AAT terms, ensuring a standardised approach. Additionally, PROV-A integrates with Zotero, a free, open-source reference management software, to streamline the citation of historical sources.[28] Through the Zotero API, the interface automatically populates the citation fields when users enter a Zotero entry URL. Finally, recognising that scholarly interpretations may contradict, PROV-A allows overlapping activities. This feature allows users to designate activities as contradictions or alternative viewpoints rather than as continuations of prior ones, thereby capturing a richer, multi-perspectival account of the artefact's history and provenance.

Edits made during the data structuring step are stored in the project's JSON and saved in web storage, a browser-based API that ensures persistent data storage.[29] This guarantees that users retain their progress across sessions, ensuring a seamless workflow. Additionally, users can download their data in JSON format for backup.

The third step in the PROV-A workflow focuses on generating and querying provenance LOD. During the data structuring process, users can generate provenance LOD at any time through the "generate LOD" section. This client-side operation uses the N3.js library to structure and manipulate data based on RDF (Resource Description Framework).[30] RDF represents data as triples, each consisting of a subject, predicate, and object. The subject is the entity being described, the predicate defines a property or relationship, and the object is either another entity or a value. These triples form a graph structure, enabling the representation and querying of relationships between data points.

In PROV-A, each provenance activity is represented as a nanopublication, which structures the activity's data into three interconnected RDF graphs. The assertion graph captures historical data about the provenance activity, structured in

28. Zotero (https://www.zotero.org/).
29. Web Storage (https://html.spec.whatwg.org/multipage/webstorage.html).
30. N3.js (https://zenodo.org/records/10866356).

compliance with CIDOC-CRM.[31] The provenance graph documents the origins and context of the assertion, including the author, date, sources consulted, and confidence expressed by the author. This graph follows the CRMinf data structure, a CIDOC-CRM extension designed to describe inference-making activities and their metadata.[32] The publication info graph contains metadata about the nanopublication itself, structured according to the Dublin Core Metadata Initiative (DCMI) Metadata Terms.[33] It includes key attributes such as authorship (via ORCID identifiers), Creative Commons licensing, and the nanopublication's creation date.

Once generated, users can download LOD as an n-quads file, a plain text serialisation for encoding RDF graphs. Each line in an n-quads file represents a single RDF statement, consisting of a subject, predicate, and object, followed by the graph URI that identifies the graph to which the triple belongs. To store LOD in the browser, PROV-A uses the Quadstore library for managing RDF graph storage through IndexedDB, a client-side database API.[34] Quadstore also integrates Comunica, a framework for querying knowledge graphs through SPARQL (SPARQL Protocol and RDF Query Language).[35]

Integrating a SPARQL endpoint within the application reflects a deliberate compromise between usability and advanced data analysis, serving both research and educational purposes. The interface incorporates a set of predefined SPARQL queries, specifically designed to assist non-expert users in exploring and analysing LOD with minimal prior knowledge of RDF or SPARQL. A central function of these predefined queries is to help users detect incompleteness in provenance data. Since unknown information cannot be directly modelled as LOD, incompleteness is addressed through query patterns that reveal where data is missing.[36] Two main patterns of incompleteness can be distinguished. The first concerns gaps in the chain of activities, where two events are chronologically linked but the party receiving the object in the first event is not the one transferring it in the second. Such gaps suggest the existence of unrecorded intermediate transfers of ownership or custody. The second pattern concerns missing constituents within activities, such as absent temporal or spatial information. These omissions indicate that an event is structurally incom-

31. The PROV-A RDF shape definition (https://github.com/prov-a/prov-a.github.io/blob/main/test/test_RDF/shape.ttl), compiled using Shapes Constraint Language (SHACL), describes the structure and constraints of PROV-A data in RDF format. A description of the data model is also available in the PROV-A data model documentation (https://github.com/prov-a/prov-a.github.io/blob/main/data-model.md).

32. Doerr, Ore, and Fafalios (2023).

33. DCMI Metadata Terms (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/).

34. Quadstore (https://github.com/quadstorejs/quadstore).

35. Taelman et al. (2018). Comunica (https://comunica.dev/).

36. Mariani (2023).

plete. By formalising these patterns in predefined SPARQL queries, the system allows users to identify incomplete provenance records and, where appropriate, revisit the structuring phase to develop new hypotheses. In this way, incompleteness becomes a productive element, prompting further archival research and interpretive reflection.

8.4 Case Study: PROV-A as Human-in-the-Loop Tool

The following section illustrates how PROV-A integrates into a human-in-the-loop process, allowing users to refine and enhance information automatically extracted through NLP. As previously stated, human supervision primarily applies to VISU information, and thus this analysis specifically focuses on these elements.

The case study draws on provenance data from the Art Institute of Chicago (AIC), which was used in a prior experiment evaluating NLP techniques.[37] The experiment, conducted on museum data downloaded on 7 April 2022, involved a dataset of 11,504 objects with available provenance texts. After filtering out samples affected by typos and errors during the preprocessing stage, the dataset was reduced to 11,392 objects. Two deep learning models were trained and deployed for this experiment: one for sentence boundary disambiguation (SBD), which segmented the provenance texts into discrete events by identifying sentence boundaries, and another for span categorization, which extracted and classified specific portions of text within each identified event according to an annotation scheme designed for provenance records.[38] The SBD model achieved an F1 score of 0.99, while the span categorization model reached an F1 score of 0.94.[39]

For the PROV-A case study, a subset of the AIC dataset was selected. It comprises all artefacts from the "Modern Art" department classified as "paintings," totalling 235 objects. This subset provided both a manageable sample size and a coherent scope for testing PROV-A's human-in-the-loop refinement of automatically extracted provenance data.[40]

To carry out the experiment, the project was initiated within the designated "initialise project" section of PROV-A. The resulting project JSON file was subsequently preprocessed by incorporating data extracted from the deep learning models

37. Rother, Mariani, and Koss (2023).

38. Mariani, Rother, and Koss (2023b).

39. The SBD model was trained on 6,000 annotated texts, while the span categorization model was trained on 6,531 annotated provenance events. Both models were trained with a 60/20/20 training/validation/test split, and are available on Zenodo (https://zenodo.org/records/13987656).

40. The case study material is available on the PROV-A GitHub repository (https://github.com/prov-a/prov-a.github.io/releases/AIC_CaseStudy).

according to the PROV-A JSON Schema. Following this preprocessing stage, the JSON file was imported into the "structure data" section, where it underwent supervision and enrichment to refine the preprocessed data before querying results in the "generate LOD" section. The analysis was conducted through the SPARQL endpoint integrated into PROV-A.

Using the NLP models trained in the previous experiment, 975 distinct events were extracted from the 235 artefacts' provenance records, averaging about four events per object. After manual enrichment through PROV-A, the total number of events increased to 1,166, roughly five events per object. This increase stems from the frequent omission of the artefact's creator as the first owner in the provenance texts. In such cases, manual intervention is required to add an initial event representing the creation of the artefact by the artist. This information is typically sourced from metadata associated with the object, such as the author and creation date, rather than from the provenance text itself.

A critical aspect in the analysis of provenance information lies in handling vagueness, particularly when dealing with approximations of dates. The annotation scheme used for span categorization includes a specific category for identifying textual elements that convey vague information. However, appropriate representation of this data is only achievable through supervised processing in PROV-A, where dates are contextualised by a user and converted into EDTF before being modelled into LOD.

Among the 1,166 provenance events recorded in the case study, 1,036 include a date reference, accounting for 89% of all activities (Table 8.1). All dates are modelled as time spans using CIDOC CRM properties (crm:P82a_begin_of_the_begin and crm:P82b_end_of_the_end). For instance, a date recorded only as a year, such as "1901," spans from "1901-01-01T00:00:00Z" to "1901-12-31T23:59:59Z." The majority of events record only the year (526 instances). A higher degree of temporal accuracy is achieved in 42 instances specifying the month, and in 11 instances where the season is indicated, although the latter presents additional challenges in terms of precision. In contrast, 128 instances provide exact dates specifying the precise day, ensuring the highest level of accuracy.

Furthermore, some events reference broader temporal categories, such as decades (9 instances) and time intervals (320 instances), representing a period rather than a specific date. Among the time intervals, 48 instances are closed intervals with both start and end dates specified (e.g., "1911/23"). In 8 instances, the interval is open at the end, signifying that an event must have taken place after a given date (e.g., "after 1907"). More notably, 264 cases involve intervals open at the start, indicating that an event must have occurred by a specific date (e.g., "by 1920").

Approximation markers, such as "circa" or "around," introduce another layer of

ambiguity. Such terms appear in 54 instances, denoting varying degrees of temporal vagueness. These approximations range from single-year references (e.g., "around 1962") to broader temporal spans (e.g., "c. 1917/19"). Their presence underscores a methodological challenge in historical documentation, where exact dates are often unavailable or inferred from contextual evidence.

Category	Instances	Textual Example (EDTF)	Approximation Markers
Day	128	Mar. 29, 1963 (1963-03-29)	0
Month	42	Dec. 1940 (1940-12)	1
Season	11	spring 1909 (1909-21)	0
Year	526	1908	37
Decade	9	1920s (192X)	0
Interval (closed)	48	1911/23 (1911/1923)	10
Interval (open end)	8	after 1907 (1907/..)	1
Interval (open start)	264	by 1920 (../1920)	5

Table 8.1: Temporal approximations in provenance events for artefacts classified as "paintings" in the Modern Art department of the Art Institute of Chicago.

The analysis reveals a high frequency of events with temporal approximations based on intervals open at the start. Specifically, in 219 of these events (83%), the sender—the party who separated from the object—is not documented. This suggests the presence of a gap before the event in question. In this scenario, the approximation gains context: without a recorded prior event, provenance authors can only estimate the latest possible date when the new owner (or custodian) acquired the object. However, it remains unclear from whom the object was received, how it was obtained, or when exactly this transfer occurred.

In the case study under analysis, gaps between two events were identified in 400 instances. Of these, 164 involve a gap between the creation of the object and the subsequent event, leaving the details of how the author parted with the object

unknown. For the 235 objects examined, this indicates that in 70% of the cases, there is a gap following the creation of the artefact.

As analysed above, in the presence of a gap, the authors of the provenance record can intervene through intellectual effort, such as recording approximate dates that at least allow for a temporal delimitation of the gap. Another approach involves formulating hypotheses to bridge the gap. In doing so, the authors may express a level of confidence by employing expressions of uncertainty such as "possibly" or "probably." While vague information approximates a hypothesis without challenging it, uncertain information questions the actual veracity of the claim.[41]

PROV-A provides a structured approach to managing these interpretative challenges by enabling the documentation of contradictory activities. This allows for the representation of conflicting assertions as LOD, where two nanopublications can exist simultaneously, each expressing a different perspective on the same activity.

Examining the 164 gaps following the creation of the object, it is evident that in 17 cases, a contradictory hypothesis is recorded. These hypotheses, which include the artist as the sender of the activity (the individual who parts with the object), effectively fill the gap from the creation event. However, an analysis of the provenance graph of these assertions reveals that these hypotheses have been formulated with a level of uncertainty, described as "probable" (aat:300435721). For instance, in provenance texts, one might encounter statements such as, "Galerie Kahnweiler, Paris, probably acquired directly from the artist."[42] In such cases, while the acquisition of the object by a specific party is certain, the precise nature of the transaction—whether it was directly from the artist—remains uncertain. These speculative hypotheses often rely on information from the artist's network, particularly when a dealer is known to have associations with the artist.

8.5 Conclusion

This article has examined how PROV-A enhances the accessibility, accuracy, and analytical potential of provenance data by supporting the transformation of textual records into LOD structured as nanopublications. Provenance—which documents an artefact's creation, ownership, and custodial history—is fundamental for verifying authenticity and historical significance. Yet, textual records often lack the depth

41. Mariani (2023).

42. Provenance text of the painting Cagnes (André Derain, 1910) published on the Art Institute of Chicago website (https://www.artic.edu/artworks/12402): "Galerie Kahnweiler, Paris, probably acquired directly from the artist. Louis Lion & Co., New York, by Feb. 1957 [verso inscription; this and the following according to letter from Knoedler and Co., Apr. 8, 1975, copy in curatorial file]; sold to Knoedler & Co., New York, Feb. 1957; sold to the Art Institute of Chicago, 1960." (accessed 22 April 2025).

and accessibility required for nuanced analyses, particularly in areas such as restitution and transparency. Transforming provenance information into LOD enables institutions to effectively connect, analyse, and disseminate data, fostering greater collaboration within the cultural heritage sector.

Despite the advantages of publishing provenance as LOD, technical barriers remain, particularly when dealing with vague, incomplete, subjective, or uncertain (VISU) information inherent in provenance texts, which necessitates expert oversight to ensure data reliability and accuracy. PROV-A addresses these challenges by integrating external automated extraction workflows with human supervision, providing a user-friendly platform that allows non-technical users to refine and enrich provenance data. The nanopublication model implemented by PROV-A preserves the complexity of provenance records, enabling the inclusion of alternative hypotheses and the retention of interpretative nuances.

Through its integration of a SPARQL endpoint, PROV-A facilitates advanced querying and analysis of LOD, serving as both a research tool and an educational resource. The case study using provenance records from the Art Institute of Chicago highlights the practical benefits of PROV-A in refining and enriching provenance data. This includes managing approximate dates, identifying gaps in historical trajectories, and recording conflicting assertions while preserving the intellectual work behind the data. By facilitating the preservation of VISU information—a persistent challenge in digitisation—into interoperable, machine-readable resources, PROV-A opens new avenues for constructing historical narratives of cultural artefacts and for analysing the evolution of their documentation.

Acknowledgements

The author would like to thank the team of the Provenance Lab, in particular Lynn Rother and Max Koss for the thoughtful discussions and Svenja Weikinnis for support with data supervision.

9. Conclusions

9.1 Summary

This cumulative dissertation explored how digital infrastructures might accommodate the epistemic complexity of historical knowledge, conceptualised here through the acronym VISU: *vagueness*, *incompleteness*, *subjectivity*, and *uncertainty*. Rather than framing these characteristics as defects to be overcome through data cleaning or rigid standardisation, the research proposed that they are core to how historical meaning is constructed, transmitted, and contested. Using art provenance as a focused site of inquiry, the thesis investigated how VISU information can be identified, modelled, and preserved across the full computational workflow, from automated extraction to structured representation and expert validation.

The dissertation began by grounding its investigation in the curatorial, historical, and political contexts of provenance digitisation. In **"Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums"** (Rother, Koss, and Mariani 2022), four interdependent challenges were identified that shape the digital transformation of provenance: inconsistent legacy practices, gaps in historical knowledge, limited resources, and increasing demands for institutional accountability. In response, the article introduced the PLOD conceptual framework, a modular and layered approach to provenance modelling. Rather than promoting standardisation at the expense of nuance, it advocates for strategies that remain interoperable while respecting the ambiguity, partiality and uncertainty intrinsic to historical records. This commitment to preserving epistemic complexity shaped the dissertation's ethical stance and provided a conceptual foundation for addressing VISU information in the structuring of provenance data.

The first research question examined how automated extraction processes could detect and classify VISU characteristics within provenance information. In **"Teaching Provenance to AI: An Annotation Scheme for Museum Data"** (Mariani, Rother, and Koss 2023b), a provenance-specific annotation scheme was introduced to structure museum records and support the training of NLP models. The scheme formalises the components of provenance events, such as parties, methods, locations and dates, and also allows for the annotation of more interpretative elements. While not limited to VISU information, it enables its identification by accommodating overlapping spans that capture vague dates, incomplete entries, and subjective or uncertain statements. In doing so, the annotation scheme made it possible to capture the interpretative nature of provenance writing, enabling VISU character-

istics to be systematically annotated alongside the factual elements of provenance records.

A practical application of the provenance-specific annotation scheme is demonstrated in **"Hidden Value: Provenance as a Source for Economic and Social History"** (Rother, Mariani, and Koss 2023), which explored the feasibility of large-scale information extraction from provenance texts using NLP. The study successfully applied sentence boundary detection and span categorisation to structure records from the Art Institute of Chicago. These experiments demonstrated that provenance events and their components can be extracted with a high degree of reliability, allowing for scalable modelling of historical data. At the same time, the exclusion of provenance notes, due to their structural inconsistency, highlighted the limitations of current approaches and confirmed the continuing need for interpretive oversight where computational methods remain insufficient.

This experimental setup also served as the basis for a subsequent study, **"People Information in Provenance Data: Biographical Entity Linking with Wikidata and ULAN"** (Mariani, Koss, and Rother 2024), which examined the extraction and reconciliation of biographical entities within museum provenance records. Building on the dataset from the Art Institute of Chicago used in previous studies, it analysed how named individuals could be linked to external authority files such as Wikidata and ULAN. Using both quantitative and qualitative methods, the study evaluated the reliability of name-based matching and the disambiguation potential of life dates. A key finding was that women were disproportionately recorded through ambiguous identifiers such as marital titles, limiting their representation in structured data. At the same time, the article framed entity linking as a critical opportunity to enrich provenance records, correct inherited biases and integrate overlooked individuals into broader knowledge networks. It emphasised the need to examine naming conventions and documentation standards, while recognising the potential of linked data to support more inclusive forms of historical representation.

The second research question turned to how VISU information can be formally represented without sacrificing interpretive complexity. **"Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data"** Mariani 2023 provided the conceptual core of the dissertation. It articulated the VISU framework, proposed strategies for its integration into modelling practices, and addressed the challenge of preserving interpretive nuance within structured formats. For institutions using CIDOC-CRM, the article outlined a modelling strategy that aligns CRMinf with the Historical Context Ontology (HiCO) to represent interpretative acts as structured, semantically explicit entities. This approach treats each provenance event as a distinct interpretative claim, which is modelled and published as a nanopublication comprising three named graphs: the assertion

graph representing a single provenance activity, the provenance graph describing the interpretation context including the scholar, the date, the sources used and the reasoning process, and the publication info graph recording creation and licensing metadata. This approach makes it possible to represent multiple, even conflicting, hypotheses while preserving their provenance and rendering VISU complexity explicit, citable and open to scrutiny in linked data environments.

The third research question asked how expert validation can be operationalised to safeguard VISU information in computational settings. This was explored in **"Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI"** (Rother, Mariani, and Koss 2024), which examined the limits of AI-based extraction and introduced the concept of two-speed digitisation. While deep learning models enable fast, large-scale processing of provenance texts, the article argued that this approach is insufficient for handling the interpretative complexity of VISU information. Minor extraction errors, tolerable in distant reading, can lead to misrepresentation when data is examined in close reading contexts. Moreover, forms of vagueness, incompleteness, subjectivity and uncertainty often depend on curatorial and historiographic judgement to be recognised and interpreted. The study emphasised that expert intervention is not an auxiliary correction but a core component of digital provenance work. Its proposed human-in-the-loop methodology integrates expert decision-making at key stages of the workflow, ensuring that VISU characteristics are not only preserved but made accessible to critical interpretation within computational environments.

These insights culminated in the development of **"PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data"** (Mariani 2025). The tool operationalises the methodological commitments of the dissertation by enabling scholars and curators to refine, annotate, and publish provenance statements as nanopublications. By integrating machine-generated output with expert supervision, PROV-A ensures that epistemic features such as approximation, contradiction, and evidentiary attribution are preserved within computational systems. It completes the digital workflow envisioned in this research, which spans the identification of VISU information in text, its formal modelling, and its expert validation and publication.

Across these studies, the dissertation advanced a framework for digital historiography that foregrounds the epistemic contingency of provenance data and the interpretative agency involved in handling VISU information. It demonstrated that historical records can be rendered computationally processable while preserving the ambiguity, partiality, and uncertainty that constitute their epistemic and historical significance.

9.2 Impact

The research presented in this dissertation has attracted sustained interest across scholarly, curatorial, and institutional contexts. Initially developed within the *Modern Migrants: Paintings from Europe in US Museums* project at the *Provenance Lab*, the work informed project workflow that emphasise interpretive nuance, epistemic transparency, and accountability in the handling of provenance data.

Over the course of the doctoral project, I presented this research at a range of academic conferences and events within the museum and heritage sectors. Three venues in particular provided opportunities to engage with distinct expert communities. At Digital Humanities 2022 in Tokyo, I presented an early prototype of the PROV-A tool (Mariani 2022b). At CIDOC 2023 in Mexico City, I discussed the VISU framework and collaborative workflows with the museum documentation community (Mariani, Rother, and Koss 2023a). At Provenance Loves Wiki in Berlin (January 2024), I presented findings on the representation of biographical data in provenance records, contributing to methodological discussions within the Wikidata community on how such information can be ethically and accurately recorded (Mariani and Rother 2024b).

In addition to these conferences, I was invited to share aspects of the dissertation in expert working groups and thematic events. In April 2024, I presented a talk on *Provenance Language Processing* at the event Artificial Intelligence Meets Art History: Vorabend des Tags der Provenienzforschung (Mariani 2024a). In October 2024, I was invited to speak at the ICOM Italy Working Group on Provenance, where I presented a contribution titled *Provenance and Artificial Intelligence* (Mariani 2024b).

Finally, the research was the basis for the workshop *Re-thinking Object Histories – or: How to Create Structured Provenance Data?*, co-organised with Prof. Dr. Lynn Rother at the Getty Research Institute in May 2024 (Mariani and Rother 2024a). Participants engaged hands-on with the PROV-A tool, examining the integration of expert judgement in structuring complex provenance records.

Together, these engagements supported interdisciplinary dialogue and highlighted the broader relevance of the dissertation across the domains of digital history, curatorial practice, and cultural data infrastructure.

9.3 Limitations

As a contribution to a still-emerging field, this dissertation inevitably encountered several limitations, both methodological and infrastructural, that shaped the scope and execution of the research. Provenance digitisation remains at an early stage of

development in many institutions. In this sense, the project was situated at the frontier of experimentation rather than within a mature ecosystem of standardised data practices. This pioneering status was both a source of innovation and a constraint, particularly in terms of data availability, tooling, and institutional interoperability.

One significant limitation stemmed from the restricted access to provenance records. Most institutions do not yet offer machine-readable corpora for bulk analysis, which made it necessary to rely on manual collection or data scraping, along with all the attendant ethical and technical considerations. As a result, the deep learning pipeline developed for automated extraction and classification of VISU elements was tested exclusively on the corpus from the Art Institute of Chicago, one of the few institutions providing systematically accessible data.

Moreover, while the complete digital workflow spanning automated extraction, ontological modelling, and expert validation was developed in full, it was only tested at scale on a limited subset of records. This was due in part to the inherent tension between the speed of automated structuring and the slowness of human oversight. While AI methods can process large volumes of records rapidly, expert validation remains time-intensive and reliant on expert human labour. This introduced a practical trade-off between large-scale high-level analysis, as pursued in studies like Mariani, Koss, and Rother (2024) and Rother, Mariani, and Koss (2023), and close, statement-level curation as presented in Mariani (2025).

These constraints were not merely technical in nature but reflected deeper tensions within the landscape of digital historiography, particularly the friction between the desire for scalable computational solutions and the realities of historically contingent, institution-specific data. While the research navigated these tensions productively, they nonetheless placed practical limits on the extent to which the proposed workflows could be tested across varied institutional contexts and at scale.

9.4 Future Work

The research presented in this dissertation provides a foundation for further investigations into the modelling and computational treatment of VISU information. While the framework, methods, and tools developed here have demonstrated the viability of this approach within a specific institutional and linguistic context, several avenues for extension remain open.

A first direction for future research involves the application of the methodology to a wider range of provenance records, particularly those drawn from institutions with different cataloguing conventions, linguistic registers, and historical legacies. Provenance documentation practices vary widely, and records are often shaped by institutional habits, localised terminologies, and distinct archival constraints. Ap-

plying the VISU framework across such heterogeneous sources would allow for a more comprehensive assessment of its generalisability and help identify areas where further refinement or adaptation is required.

Another promising trajectory emerges from the growing availability of large language models (LLMs) and their integration into knowledge extraction workflows. While this dissertation employed more conventional deep learning approaches to support automated extraction, future research might explore how generative models can assist with tasks such as VISU-aware annotation and classification. These models potentially offer new ways of scaling the identification of epistemic features without requiring extensive manual data annotation. At the same time, applying these models to historically contingent and ambiguous data raises critical questions about control, transparency, and epistemic accountability. Experimental work will be needed to assess their suitability for knowledge extraction, particularly in addressing VISU-specific challenges such as people recording biases, data approximations, and statement uncertainties, and to evaluate how LLMs output can be critically situated within a human-in-the-loop framework.

A further area worth exploring is the visual representation of VISU information. Given the spatiotemporal nature of provenance, with its grounding in events, actors, locations, and timeframes, visualisation offers a valuable means of interpretation and exploration. However, the characteristics of VISU information complicate traditional approaches to mapping and timelines. Ambiguities such as vague dates, partial trajectories, and competing hypotheses challenge established visual conventions. Future research may explore how visualisation techniques can be adapted to make such ambiguity legible without collapsing interpretive nuance into deterministic representations.

Whichever direction future developments may take, the digital handling of VISU information will continue to depend on a *de visu* approach—anchored in interpretation, guided by critical judgement, and attentive to the evidentiary contours of historical knowledge at every stage of the process.

Bibliography

Aberdeen, John, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1995. "MITRE: Description of the Alembic System Used for MUC-6." In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.* https://aclanthology.org/M95-1012/.

Alzubaidi, Laith, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions." *Journal of Big Data* 8 (1): 53. https://doi.org/10.1186/s40537-021-00444-8.

Amineddoleh, Leila. 2020. "The Role of Provenance in Resolving Art-World Disputes." In *Provenance Research Today: Principles, Practice, Problems,* edited by Arthur Tompkins, 25–38. London: Lund Humphries.

Appadurai, Arjun. 1986. "Introduction: Commodities and the Politics of Value." In *The Social Life of Things: Commodities in Cultural Perspective,* edited by Arjun Appadurai, 3–63. Cambridge: Cambridge University Press.

Ashenfelter, Orley, and Kathryn Graddy. 2020. "Art Auctions." In *Handbook of Cultural Economics,* edited by Ruth Towse and Trilce Navarrete Hernández. Cheltenham: Edward Elgar.

Asif, Imran, Ilaria Tiddi, and Alasdair J. G. Gray. 2021. "Using Nanopublications to Detect and Explain Contradictory Research Claims." In *2021 IEEE 17th International Conference on eScience,* 1–10. New York, NY: IEEE. https://doi.org/10.1109/eScience51609.2021.00010.

Barron, Stephanie, ed. 1991. *"Degenerate Art": The Fate of the Avant-Garde in Nazi Germany.* Los Angeles: Los Angeles County Museum of Art.

Bekiari, Chryssoula, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios, eds. 2021. *Definition of the CIDOC Conceptual Reference Model, version 7.1.1.* International Committee for Documentation (CIDOC). http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf.

Bell, Joshua A., Kimberly Christen, and Mark Turin. 2013. "After the Return: Digital Repatriation and the Circulation of Indigenous Knowledge Workshop Report." *Museum Worlds: Advances in Research* 1 (1): 195–203. https://doi.org/10.3167/armw.2013.010112.

Berg-Fulton, Tracey, David Newbury, and Travis Snyder. 2015. "Art Tracks: Visualizing the Stories and Lifespan of an Artwork." In *Museums and the Web 2015.* Chicago, IL. https://mw2015.museumsandtheweb.com/paper/art-tracks-visualizing-the-stories-and-lifespan-of-an-artwork/.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste.* Translated by Richard Nice. Cambridge, MA: Harvard University Press.

Carroll, Jeremy J., Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. "Named Graphs, Provenance and Trust." In *Proceedings of the 14th International Conference on World Wide Web - WWW '05,* 613–622. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1060745.1060835.

Challis, David M. 2021. *Foreign Currency Volatility and the Market for French Modern Art.* Vol. 12. Studies in the History of Collecting Art Markets. Leiden: Brill.

Claassen, Babette, Jeroen Borst, Ingrid Vermeulen, Victor de Boer, and Chris Dijkshoorn. 2020. "Linked Art Provenance." In *Proceedings of the Network Institute Academy Assistants Programme 2018–2019.* https://doi.org/10.5281/zenodo.4003499.

Cranston, Jodi. 2020. "Mapping Paintings, or How to Breathe Life Into Provenance." In *The Routledge Companion to Digital Humanities and Art History,* edited by Kathryn Brown, 109–119. London: Routledge.

Crotta, Alessia, and Filip Vermeylen. 2020. *Does Nudity Sell? An Econometric Analysis of the Value of Female Nudity in Modigliani Portraits.* Technical report, ACEI Working Paper Series 2. Association for Cultural Economics International (ACEI).

Cybulska, Agata Katarzyna, and Piek Vossen. 2011. "Historical Event Extraction from Text." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* edited by Kalliopi Zervanou and Piroska Lendvai, 39–43. Portland, OR, USA: Association for Computational Linguistics. https://aclanthology.org/W11-1506/.

Daquino, Marilena, Valentina Pasqual, and Francesca Tomasi. 2020. "Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources." *JLIS.it* 11 (3): 59–76. ISSN: 2038-1026. https://doi.org/10.4403/jlis.it-12642.

Daquino, Marilena, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali. 2022. "Expressing Without Asserting in the Arts." In *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022,* edited by Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello. CEUR Workshop Proceedings. CEUR-WS.org.

Daquino, Marilena, and Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects." In *Metadata and Semantics Research,* edited by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, 424–436. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37.

Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, and Francesca Tomasi. 2023. "CLEF. A Linked Open Data Native System for Crowdsourcing." *Journal on Computing and Cultural Heritage* (New York, NY, USA) 16 (3). https://doi.org/10.1145/3594721.

Davis, Kelly. 2019. "Old Metadata in a New World: Standardizing the Getty Provenance Index for Linked Data." *Art Libraries Journal* 44 (4): 162–166. https://doi.org/10.1017/alj.2019.24.

Dekker, Erwin. 2015. "Two Approaches to Study the Value of Art and Culture, and the Emergence of a Third." *Journal of Cultural Economics* 39 (4): 309–326.

Destandau, Marie, and Jean-Daniel Fekete. 2021. "The Missing Path: Analysing Incompleteness in Knowledge Graphs." *Information Visualization* 20:66–82. ISSN: 1473-8716. https://doi.org/10.1177/1473871621991539.

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine* 24 (3): 75–92. https://doi.org/10.1609/aimag.v24i3.1720.

Doerr, Martin, Christian-Emil Ore, and Pavlos Fafalios. 2023. *Definition of the CRMinf: An Extension of CIDOC-CRM to Support Argumentation.*

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. "Named Entity Recognition and Classification in Historical Documents: A Survey." *ACM Computing Surveys* (New York, NY, USA) 56 (2). https://doi.org/10.1145/3604931.

Al-Eryani, Susanne, Gudrun Bucher, and Stefanie Rühle. 2018. "Ein Metadatenmodell für gemischte Sammlungen." *Bibliotheksdienst* 52:548–564. https://doi.org/10.1515/bd-2018-0066.

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

Feigenbaum, Gail, and Inge Reist. 2013. "Introduction to Provenance: An Alternate History of Art." In *Provenance: An Alternate History of Art.* Los Angeles: Getty Research Institute.

Finkel, Jenny Rose, and Christopher D. Manning. 2009. "Nested Named Entity Recognition." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* 141–150. Singapore: Association for Computational Linguistics. https://doi.org/10.3115/1699510.1699529.

Firth, John Rupert. 1957. "A Synopsis of Linguistic Theory, 1930–1955." In *Studies in Linguistic Analysis,* edited by John Rupert Firth, 1–32. Philological Society, Oxford.

FitzGerald, Michael C. 1995. *Making Modernism: Picasso and the Creation of the Market for Twentieth-Century Art.* New York: Farrar, Straus / Giroux.

Fleckner, Uwe. 2015. "Dubious Business: Trade in Modern Art under the 'Third Reich'." In *Looters, Smugglers, and Collectors: Provenance Research and the Market,* edited by Ana María Bresciani and Tone Hansen, 21–34. Köln: Verlag der Buchhandlung Walther König.

Foster, Ian. 2011. "How Computation Changes Research." In *Switching Codes: Thinking through Digital Technology in the Humanities and the Arts,* edited by Thomas Bartscherer and Roderick Coover, 32. Chicago: University of Chicago Press.

Fraiberger, Samuel P., Roberta Sinatra, Magnus Resch, Christoph Riedl, and Albert-László Barabási. 2018. "Quantifying Reputation and Success in Art." *Science* 362 (6416): 825–829. https://doi.org/10.1126/science.aau7224.

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

Fuhrmeister, Christian, and Meike Hopp. 2019. "Rethinking Provenance Research." *Getty Research Journal* 11:213–231. https://doi.org/10.1086/702755.

Gagliardi, Susan Elizabeth. 2020. "Mapping Senufo: Mapping as a Method to Transcend Colonial Assumptions." In *The Routledge Companion to Digital Humanities and Art History,* edited by Kathryn Brown, 135–154. New York: Routledge.

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

Gomaa, Wael H., and Aly A. Fahmy. 2013. "A Survey of Text Similarity Approaches." *International Journal of Computer Applications* (New York, USA) 68 (13): 13–18. ISSN: 0975-8887. https://doi.org/10.5120/11638-7118.

Graham, Mark, and Stefano De Sabbata. 2015. "Mapping Information Wealth and Poverty: The Geography of Gazetteers." *Environment and Planning A* 47 (6): 1254–1264. https://doi.org/10.1177/0308518X15594899.

Gramlich, Johannes. 2017. "Reflections on Provenance Research: Values – Politics – Art Markets." *Journal for Art Market Studies* 1 (2). https://doi.org/10.23690/JAMS.V1I2.15.

Grana-Behrens, Daniel. 2021. "Digitalbasierte Ethnologische Provenienzforschung: Chancen und Herausforderungen am Beispiel WissKI der Bonner Amerikas-Sammlung (BASA-Museum)." In *Digitalisierung Ethnologischer Sammlungen,* edited by Hans Peter Hahn, Oliver Lueb, Katja Müller, and Karoline Noack, 215–238. Bielefeld, Germany: Transcript Verlag. https://doi.org/10.1515/9783839457900-013.

Greenwald, Diana Seave. 2021. *Painting by Numbers: Data-Driven Histories of Nineteenth-Century Art.* Princeton, NJ: Princeton University Press.

Griffis, Denis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M. Lai. 2016. "A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain." *AMIA Joint Summits on Translational Science Proceedings* 2016:88–97.

Grimme, Gesa. 2020. "Systemizing Provenance Research on Objects from Colonial Contexts." *Museum and Society* 18 (1): 52–65.

Groth, Paul, Andrew Gibson, and Jan Velterop. 2010. "The Anatomy of a Nanopublication." *Information Services & Use* 30 (1-2): 51–56. ISSN: 0167-5265. https://doi.org/10.3233/ISU-2010-0613.

Harpring, Patricia. 2010. "Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA." *Art Documentation: Journal of the Art Libraries Society of North America* 29 (1): 67–72. https://doi.org/10.1086/adx.29.1.27949541.

Hiebel, Gerald, Martin Doerr, and Øyvind Eide. 2017. "CRMgeo: A Spatiotemporal Extension of CIDOC-CRM." *International Journal on Digital Libraries* 18 (4): 271–279. ISSN: 1432-5012, 1432-1300. https://doi.org/10.1007/s00799-016-0192-4.

Hiebel, Gerald, Martin Doerr, Klaus Hanke, and Anja Masur. 2014. "How to Put Archaeological Geometric Data into Context? Representing Mining History Research with CIDOC CRM and Extensions." *International Journal of Heritage in the Digital Era* 3 (3): 557–577. ISSN: 2047-4970. https://doi.org/10.1260/2047-4970.3.3.557.

Higonnet, Anne. 2013. "Afterword: The Social Life of Provenance." In *Provenance: An Alternate History of Art,* edited by Gail Feigenbaum and Inge Reist, 197. Los Angeles: Getty Research Institute.

Holmen, Jon, and Christian-Emil Ore. 2010. "Deducing Event Chronology in a Cultural Heritage Documentation System." In *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). 37th International Conference, Williamsburg, Virginia, United States of America, March 22-26 (BAR International Series S2079),* edited by B. Frischer, J. Webb Crawford, and D. Koller, 122–129. Oxford: Archaeopress.

Hu, Yifan, Changwei Hu, Thanh Tran, Tejaswi Kasturi, Elizabeth Joseph, and Matt Gillingham. 2021. "What's in a Name? – Gender Classification of Names with Character Based Machine Learning Models." *Data Mining and Knowledge Discovery* 35:1537–1563. https://doi.org/10.1007/s10618-021-00748-6.

Huemer, Christian. 2020. "The Provenance of Provenances." In *Collecting and Provenance: A Multidisciplinary Approach,* edited by Jane Milosch and Nick Pearce, 2–15. Lanham, MD: Rowman / Littlefield.

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

Jaskot, Paul B. 2020. "Digital Methods and the Historiography of Art." In *The Routledge Companion to Digital Humanities and Art History,* 1st ed., edited by Kathryn Brown, 9–17. New York, NY: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780429505188-3.

Jordal, Ellen, Espen Uleberg, and Brit Hauge. 2012. "Was It Worth It? Experiences with a CIDOC CRM-based Database." In *Revive the Past: Computer Applications and Quantitative Methods in Archaeology (CAA), Proceedings of the 39th International Conference, Beijing, April 12–16, 2011,* edited by Mingquan Zhou, Iza Romanowska, Zhongke Wu, Pengfei Xu, and Philip Verhagen, 255–260. Amsterdam, Netherlands.

Joyeux-Prunel, Béatrice, Catherine Dossin, and Sorin Adam Matei. 2013. "Spatial (Digital) History: A Total Art History?—The Artl@s Project." *Visual Resources* 29 (1–2): 47–58. https://doi.org/10.1080/01973762.2013.761119.

Joyeux-Prunel, Béatrice, and Olivier Marcel. 2015. "Exhibition Catalogues in the Globalization of Art: A Source for Social and Spatial Art History." *Artl@s Bulletin* 4 (2): 26.

Klinke, Harald. 2020. "The Digital Transformation of Art History." In *The Routledge Companion to Digital Humanities and Art History,* edited by Kathryn Brown, 32–42. Routledge Art History and Visual Studies Companions. London: Routledge.

Kuhn, Tobias, Albert Meroño-Peñuela, Alexander Malic, Jorrit H. Poelen, Allen H. Hurlbert, Emilio Centeno Ortiz, Laura I. Furlong, et al. 2018. "Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data." In *2018 IEEE 14th International Conference on e-Science (e-Science),* 83–92. https://doi.org/10.1109/eScience.2018.00024.

Kuhnen, Steven, Rebecca Sepers, Chris Dijkshoorn, Viktor de Boer, and Lora Aroyo. 2018. *Structuring Cultural Heritage PROVenance: The Rijksmuseum Use Case.* Presented at DHBenelux, June 6–8, 2018. Amsterdam, Netherlands.

Lai, Viet Dac, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. "Event Extraction from Historical Texts: A New Dataset for Black Rebellions." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 2390–2400. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.211.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lincoln, Matthew, and Sandra van Ginhoven. 2018. "Modeling a Fragmented Archive: A Missing Data Case Study from Provenance Research." In *Digital Humanities 2018: Book of Abstracts,* edited by Jonathan Girón Palau and Isabel Galina Russell. Presented at Digital Humanities 2018: "Puentes/Bridges", June 21, 2018. Mexico City: Red de Humanidades Digitales A. C. https://dh2018.adho.org/en/modeling-the-fragmented-archive-a-missing-data-case-study-from-provenanceresearch/.

Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events.* Technical report. Philadelphia: Linguistic Data Consortium.

Luther, Anne. 2020. "Digital Provenance, Open Access, and Data-Driven Art History." In *The Routledge Companion to Digital Humanities and Art History,* 1st ed., edited by Kathryn Brown, 448–458. Routledge Art History and Visual Studies Companions. New York, NY: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780429505188-38.

MacDonald, Isobel. 2022. "Counting when, who and how: Visualizing the British Museum's history of acquisition through collection data, 1753–2019." *Journal of the History of Collections* 35 (2): 305–320. ISSN: 0954-6650. https://doi.org/10.1093/jhc/fhac034.

Mandel, Benjamin R. 2009. "Art as an Investment and Conspicuous Consumption Good." *American Economic Review* 99 (4): 1653–1663. https://doi.org/10.1257/aer.99.4.1653.

Mariani, Fabio. 2022a. *'Probably Sold to Paalen, Possibly by Exchange': Vagueness, Incompleteness, Subjectivity and Uncertainty in Digital Art Provenance.* Paper delivered at the Workshop on Computational Methods in the Humanities. https://wp.unil.ch/llist/files/2022/06/COMHUM_2022_paper_5.pdf.

———. 2022b. *The Expert in the Loop: Developing a Provenance Linked Open Data Management Platform.* Presented at Digital Humanities 2022 – Responding to Asian Diversity. 28 July 2022, Tokyo, Japan.

———. 2023. "Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data." In *Proceedings of the Workshop on Computational Methods in the Humanities 2022.* Lausanne, Switzerland. https://ceur-ws.org/Vol-3602/paper5.pdf.

———. 2024a. *Provenance Language Processing: A Human-in-the-Loop Perspective.* Invited talk at "Artificial Intelligence meets Art History: Vorabend des Tags der Provenienzforschung". 9 April 2024.

Mariani, Fabio. 2024b. *Provenienza e Intelligenza Artificiale (Provenance and Artificial Intelligence).* Invited talk at "Workshop Provenienza: Patrimoni dal Mondo in Italia". 11 October 2024, Genova, Italy.

———. 2025. "PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data." *Zeitschrift für digitale Geisteswissenschaften,* no. 10, https://doi.org/10.17175/2025_012.

Mariani, Fabio, Max Koss, and Lynn Rother. 2024. "People Information in Provenance Data: Biographical Entity Linking with Wikidata and ULAN." *Život umjetnosti,* no. 114, 148–161. https://doi.org/10.31664/zu.2024.114.07.

Mariani, Fabio, and Lynn Rother. 2024a. *Re-thinking Object Histories – or: How to create Structured Provenance Data?* Workshop organized at the Getty Research Institute. 15 May 2024, Los Angeles, California, United States.

———. 2024b. *Towards Wikidata: How to Transform Provenance with AI.* Presented at Provenance loves Wiki 2024. 13 January 2024, Berlin, Germany. https://doi.org/10.5281/zenodo.10555215.

Mariani, Fabio, Lynn Rother, and Max Koss. 2023a. *Artificial Intelligence, Human Expertise: the Case of Provenance Linked Open Data.* Presented at the CIDOC 2023 Conference – Frontiers of Knowledge: Museums, Documentation and Linked Data. 24–28 September 2023, Mexico City, Mexico.

———. 2023b. "Teaching Provenance to AI: An Annotation Scheme for Museum Data." In *AI in Museums: Reflections, Perspectives and Applications,* edited by Sonja Thiel and Johannes Bernhardt, 163–172. Edition Museum. Bielefeld: transcript Verlag. https://doi.org/10.14361/9783839467107-014.

Massari, Arcangelo, and Silvio Peroni. 2025. "HERITRACE: A User-Friendly Semantic Data Editor with Change Tracking and Provenance Management for Cultural Heritage Institutions." *Umanistica Digitale* 9 (20): 317–340. https://doi.org/10.6092/issn.2532-8816/21218.

Masurovsky, Marc. 2020. "The Current State of Nazi-Era Provenance Research, and Access to Nazi-Era Research Resources and Archives." In *Provenance Research Today: Principles, Practice, Problems,* edited by Arthur Tompkins, 136–149. London: Lund Humphries.

Milosch, Jane C., Lynn H. Nicholas, and Megan M. Fontanella, eds. 2014. *Provenance Research in American Institutions.* Special issue of "Collections: A Journal for Museum and Archives Professionals", 10, no. 3. Lanham, MD: Rowman & Littlefield.

Mons, Barend, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz O. B. da Silva Santos, and Mark D. Wilkinson. 2017. "Cloudy, Increasingly FAIR: Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud." *Information Services & Use* 37 (1): 49–56. https://doi.org/10.3233/ISU-170824.

Moreau, Luc, and Paul Groth. 2013. *Provenance: An Introduction to PROV* [in en]. Synthesis Lectures on Data, Semantics, and Knowledge. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-79450-6.

Nagypál, Gábor, and Boris Motik. 2003. "A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies." In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE,* edited by Robert Meersman, Zahir Tari, and Douglas C. Schmidt, 906–923. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-39964-3_57.

Newbury, David. 2017. *Art Tracks: Using Linked Open Data for Object Provenance in Museums.* Presentation, Museums and the Web 2017, Cleveland, Ohio, April 19–22, 2017. https://mw17.mwconf.org/paper/art-tracks-using-linked-open-data-for-object-provenancein-museums/.

———. 2018. "Loud: Linked Open Usable Data and linked.art." In *CIDOC 2018 Conference.* Heraklion, Greece.

Newbury, David, and Louise Lippincott. 2019. "Provenance in 2050." In *Collecting and Provenance: A Multidisciplinary Approach,* edited by Jane C. Milosch and Nick Pearce, 101–109. Lanham, MD: Rowman & Littlefield Publishers.

Niccolucci, Franco, and Sorin Hermon. 2017. "Expressing Reliability with CIDOC CRM." *International Journal on Digital Libraries* 18 (4): 281–287. ISSN: 1432-1300. https://doi.org/10.1007/s00799-016-0195-1.

Nicholas, Lynn H. 1995. *The Rape of Europe: The Fate of Europe's Treasures in the Third Reich and the Second World War.* New York: Vintage.

Niederacher, Sonja. 2012. *Eigentum und Geschlecht: jüdische Unternehmerfamilien in Wien (1900–1960).* Vienna: Böhlau.

Oldman, Dominic, and Diana Tanase. 2018. "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace." In *The Semantic Web – ISWC 2018,* edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 325–340. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_20.

Oomen, Johan, and Lora Aroyo. 2011. "Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges." In *C&T 2011: Proceedings of the 5th International Conference on Communities and Technologies,* 138–149. Brisbane, Australia: Association for Computing Machinery. https://doi.org/10.1145/2103354.2103373.

Oosterlinck, Kim. 2016. "Art as Wartime Investment: Conspicuous Consumption and Discretion." *The Economic Journal* 127 (607): 2665–2701. https://doi.org/10.1111/ecoj.12391.

Oosterlinck, Kim, and Anne-Sophie Radermecker. 2019. ""The Master of..." Creating Names for Art History and the Art Market." *Journal of Cultural Economics* 43:57–95.

Papadakis, Manos, and Martin Doerr. 2015. "Temporal Primitives, an Alternative to Allen Operators." In *Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Poznań, Poland, September 17, 2015,* edited by Paola Ronzino, 69–78. CEUR Workshop Proceedings. CEUR-WS.org.

Pergam, Elizabeth A. 2013. "Provenance as Pedigree: The Marketing of British Portraits in Gilded Age America." In *Provenance: An Alternate History of Art,* edited by Gail Feigenbaum and Inge Reist, 104–122. Los Angeles: Getty Research Institute.

Piotrowski, Michael. 2019. "Accepting and Modeling Uncertainty." *Zeitschrift für digitale Geisteswissenschaften* 4. https://doi.org/10.17175/SB004_006A.

Pommerehne, Werner, and Lars Feld. 1997. "The Impact of Museum Purchase on the Auction Prices of Paintings." *Journal of Cultural Economics* 21 (3): 249–271. https://doi.org/10.1023/A:1007388024711.

Prechelt, Lutz. 2012. "Early Stopping — But When?" In *Neural Networks: Tricks of the Trade: Second Edition,* edited by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, 53–67. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_5.

Raux, Sophie. 2012. "From Mariette to Joullain: Provenance and Value in Eighteenth-Century French Auction Catalogs." In *Provenance: An Alternate History of Art,* edited by Gail Feigenbaum and Inge Jackson Reist, 85–103. Los Angeles, CA: Getty Research Institute.

Reitlinger, Gerald. 1961. *The Economics of Taste: The Rise and Fall of Picture Prices 1760–1960.* Vol. I. London: Barrie / Rockliff.

Reitlinger, Gerald. 1963. *The Economics of Taste: The Rise and Fall of Objets d'Art Prices since 1750.* Vol. II. London: Barrie / Rockliff.

———. 1970. *The Economics of Taste: The Art Market in the 1960s.* Vol. III. London: Barrie / Jenkins.

Renneboog, Luc, and Christophe Spaenjers. 2013. "Buying Beauty: On Prices and Returns in the Art Market." *Management Science* 59 (1): 36–53. https://doi.org/10.1287/mnsc.1120.1580.

Rieder, Bernhard, and Theo Röhle. 2012. "Digital Methods: Five Challenges." In *Understanding Digital Humanities,* edited by David M. Berry, 67–84. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_4.

Riley, Michael D. 1989. "Some applications of tree-based modelling to speech and language." In *Proceedings of the Workshop on Speech and Natural Language,* 339–352. HLT '89. Cape Cod, Massachusetts: Association for Computational Linguistics. https://doi.org/10.3115/1075434.1075492.

Rother, Lynn. 2017. *Kunst durch Kredit: Die Berliner Museen und ihre Erwerbungen von der Dresdner Bank 1935.* Berlin: De Gruyter. https://doi.org/10.1515/9783110496093.

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

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

———. 2024. "Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI." In *Sammlungsforschung im digitalen Zeitalter: Chancen, Herausforderungen und Grenzen,* edited by Katharina Günther and Stefan Alschner, 93–103. Göttingen: Wallstein. https://doi.org/10.1515/jbwg-2023-0005.

Rother, Lynn, and Iris Schmeisser. 2020. "Provenance Research in Museums: The Long Run." In *Provenance Research Today: Principles, Practice, Problems,* edited by Arthur Tompkins, 106–116. London: Lund Humphries.

Sanchez, George. 2019. "Sentence Boundary Detection in Legal Text." In *Proceedings of the Natural Legal Language Processing Workshop 2019,* edited by Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, 31–38. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2204.

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

Shoilee, Sarah Binta Alam, Victor de Boer, and Jacco van Ossenbruggen. 2023. "Polyvocal Knowledge Modelling for Ethnographic Heritage Object Provenance." In *Proceedings of the 19th International Conference on Semantic Systems, 20–22 September 2023, Leipzig, Germany,* 127–143. IOS Press. https://doi.org/10.3233/SSW230010.

Sikos, Leslie F., and Dean Philp. 2020. "Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs." *Data Science and Engineering* 5 (3): 293–316. ISSN: 2364-1541. https://doi.org/10.1007/s41019-020-00118-0.

Smets, Philippe. 1997. "Imperfect Information: Imprecision and Uncertainty." In *Uncertainty Management in Information Systems,* edited by Amihai Motro and Philippe Smets, 225–254. Boston, MA: Springer. https://doi.org/10.1007/978-1-4615-6245-0_8.

Smith, Jeffrey. 2018. "Toward "Big Data" in Museum Provenance." In *Big Data in the Arts and Humanities: Theory and Practice,* edited by Giovanni Schiuma and Daniela Carlucci, 41–50. Data Analytics Applications. New York, NY: Auerbach Publishers.

Smithson, Michael. 1989. *Ignorance and Uncertainty: Emerging Paradigms.* Cognitive Science. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-3628-3.

Sohrab, Mohammad Golam, and Makoto Miwa. 2018. "Deep Exhaustive Model for Nested Named Entity Recognition." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* edited by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 2843–2849. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1309.

Sprugnoli, Rachele, and Sara Tonelli. 2019. "Novel Event Detection and Classification for Historical Texts." *Computational Linguistics* (Cambridge, MA) 45 (2): 229–265. https://doi.org/10.1162/coli_a_00347.

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

Sustkova, Hana Pergl, Kristina Maria Hettne, Peter Wittenburg, Annika Jacobsen, Tobias Kuhn, Robert Pergl, Jan Slifka, et al. 2020. "FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources." *Data Intelligence* 2 (1-2): 158–170. https://doi.org/10.1162/dint_a_00038.

Taelman, Ruben, Joachim Van Herwegen, Miel Vander Sande, and Ruben Verborgh. 2018. "Comunica: A Modular SPARQL Query Engine for the Web." In *The Semantic Web – ISWC 2018,* 11137:239–255. Lecture Notes in Computer Science. Cham: Springer. https://doi.org/10.1007/978-3-030-00668-6_15.

Tan, Chuanqi, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. "Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition." *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05): 9016–9023. https://doi.org/10.1609/aaai.v34i05.6434.

Ter Braake, Serge, Antske Fokkens, Niels Ockeloen, and Chantal Van Son. 2016. "Digital History: Towards New Methodologies." In *Computational History and Data-Driven Humanities,* edited by Bojan Bozic, Gavin Mendel-Gleason, Christophe Debruyne, and Declan O'Sullivan, 482:23–32. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46224-0_3.

Towse, Ruth, and Trilce Navarrete Hernández, eds. 2020. *Handbook of Cultural Economics.* 3rd. Cheltenham: Edward Elgar Publishing.

U.S. Department of State, Office of the Special Envoy for Holocaust Issues. 1998. *Washington Conference Principles on Nazi-Confiscated Art.* https://www.state.gov/washington-conference-principles-on-nazi-confiscated-art/.

Velthuis, Olav. 2005. *Talking Prices: Symbolic Meanings of Prices on the Market for Contemporary Art.* Princeton, NJ: Princeton University Press.

Vertan, Cristina. 2019. "Modelling linguistic vagueness and uncertainty in historical texts." In *Proceedings of the Workshop on Language Technology for Digital Historical Archives,* edited by Cristina Vertan, Petya Osenova, and Dimitar Iliev, 34–38. Varna, Bulgaria: INCOMA Ltd., September. https://doi.org/10.26615/978-954-452-059-5_007. https://aclanthology.org/W19-9007/.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57 (10): 78–85. https://doi.org/10.1145/2629489.

Weber-Sinn, Kristin, and Paola Ivanov. 2020. ""Collaborative" Provenance Research: About the (Im)Possibility of Smashing Colonial Frameworks." *Museum and Society* 18 (1): 66–81. https://doi.org/10.29311/mas.v18i1.3295.

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wintle, Claire. 2013. "Decolonising the Museum: The Case of the Imperial and Commonwealth Institutes." *Museum and Society* 11 (2): 185–201.

Xiang, Wei, and Bang Wang. 2019. "A Survey of Event Extraction From Text." *IEEE Access* 7:173111–173137. https://doi.org/10.1109/ACCESS.2019.2956831.

Xu, Mingbin, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. "A Local Detection Approach for Named Entity Recognition and Mention Detection." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* edited by Regina Barzilay and Min-Yen Kan, 1237–1247. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1114.

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

Ying, Xue. 2019. "An Overview of Overfitting and its Solutions." In *Journal of Physics: Conference Series,* vol. 1168. 2. IOP Publishing. https://doi.org/10.1088/1742-6596/1168/2/022022.

Zalewski, Leanne. 2019. "Creating Cultural and Commercial Value in Late Nineteenth-Century New York Art Catalogues." In *Art Crossing Borders: The Internationalisation of the Art Market in the Age of Nation States, 1750-1914,* edited by Jan Dirk Baetens and Dries Lyna, 99–126. Leiden: Brill.

Zhang, W.J., Guosheng Yang, Yingzi Lin, Chunli Ji, and Madan M. Gupta. 2018. "On Definition of Deep Learning." In *2018 World Automation Congress (WAC),* 1–5. https://doi.org/10.23919/WAC.2018.8430387.

Zhou, GuoDong, Jian Su, Jie Zhang, and Min Zhang. 2005. "Exploring Various Knowledge in Relation Extraction." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),* edited by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, 427–434. Ann Arbor, Michigan: Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219893.

Zundert, Joris J. van. 2015. "Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics." In *A New Companion to Digital Humanities,* edited by Susan Schreibman, Ray Siemens, and John Unsworth, 331–347. Chichester, UK: John Wiley & Sons, Ltd.

Zuschlag, Christoph. 2019. "Vom Iconic Turn zum Provenancial Turn? Ein Beitrag zur Methodendiskussion in der Kunstwissenschaft." In *Von analogen und digitalen Zugängen zur Kunst: Festschrift für Hubertus Kohle zum 60. Geburtstag,* edited by Maria Effinger, Stephan Hoppe, Harald Klinke, and Bernd Krysmanski, 409–417. Heidelberg, Germany: Arthistoricum.net. https://doi.org/10.11588/arthistoricum.493.c6573.