# PROV-A, a web-based tool for structuring provenance as Linked Open Data

**Fabio Mariani**

**Angaben zur Veröffentlichung / Publication details:**

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.

ZfdG

Zeitschrift für digitale Geisteswissenschaften

Projektvorstellung aus:
Zeitschrift für digitale Geisteswissenschaften, Heft 10 (2025)

Titel:
PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data

Autor*in:
Fabio Mariani

Kontakt: fabio.mariani@leuphana.de
Institution: Institute of Philosophy and Art History, Leuphana University Lüneburg, Germany / University of Augsburg, Germany
GND: 1293533335 ORCID: 0000-0002-7382-0187
Contribution (CRediT): Conceptualization | Methodology | Software | Writing – original draft | Formal analysis

DOI des Beitrags:
10.17175/2025_012

Nachweis im OPAC der Herzog August Bibliothek:
1941754538

Erstveröffentlichung:
22.12.2025



Letzte Überprüfung aller Verweise:
21.11.2025

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Linked Data | Provenienzforschung | Human-in-the-loop | Kunstgeschichte

Empfohlene Zitierweise:
Fabio Mariani: PROV-A, a Web-Based Tool for Structuring Provenance as Linked Open Data. In: Zeitschrift für digitale Geisteswissenschaften 10 (2025). 22.12.2025. HTML / XML / PDF. DOI: 10.17175/2025_012

Fabio Mariani

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


## 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.

Provenienzen dokumentieren die Eigentums- und Besitzgeschichten von Kulturgütern. Sie belegen die Authentizität von Objekten und fördern ethische Verantwortung gegenüber dem kulturellen Erbe. Die Veröffentlichung von Provenienzen als *Linked Open Data (LOD)* verbessert den Datenzugang, die Interoperabilität sowie großangelegte Datenanalysen. Die Umwandlung von Provenienzinformationen in LOD wird jedoch nach wie vor durch arbeitsintensive Extraktionsprozesse und technische Hürden erschwert. Dieser Artikel stellt *PROV-A* (*the Provenance App*) vor, ein web-basiertes Programm zur effizienten Erstellung und Veröffentlichung von Provenienzinformationen als LOD. PROV-A erleichtert die Integration externer automatisierter Datenextraktionsprozesse, etwa mittels *Natural Language Processing (NLP)*, mit menschlicher Validierung, und schafft so einen Ausgleich zwischen Effizienz und wissenschaftlicher Genauigkeit. Anhand einer Fallstudie mit Daten des Art Institute of Chicago wird veranschaulicht, wie PROV-A es Nutzenden ermöglicht, automatisch extrahierte Daten gezielt aufzubereiten, um historische Ambivalenzen zu erhalten und Provenienzanalysen zu unterstützen. Dank des niedrigschwelligen technischen Zugangs und der Einbindung des Menschen in den Prozess (*Human in the Loop*) verbessert PROV-A die Skalierbarkeit und Genauigkeit der Provenienzforschung und erhöht zugleich die Zugänglichkeit von LOD für Kultureinrichtungen.


## 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 legal 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,[2] 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. [1]

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 [2]

[1] United States Department of State 1998.
[2] Yeide et al. 2001.

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.

[3] 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.

[4] This paper introduces *PROV-A* (*the Provenance App*), an application designed 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.

## 2. Background

[5] 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.

[6] 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]

[3] Cf. Smith 2018; Newbury / Lippincott 2019; Luther 2020; Rother et al. 2022.
[4] Wilkinson et al. 2016.
[5] Cf. Rother et al. 2023.
[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.
[7] Doerr 2003.
[8] Berg-Fulton et al. 2015; Newbury 2017.
[9] Cf. Berg-Fulton et al. 2015; Newbury 2017.

More recently, Linked Art emerged as a prominent application profile for modelling 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. [7]

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] [8]

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. [9]

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 contextual knowledge and historical analysis. [10]

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] [11]

[10] Cf. Newbury 2018.
[11] Harpring 2010.
[12] Cf. Rother et al. 2023; Mariani et al. 2023.
[13] Cf. Mariani et al. 2023.
[14] Cf. Mariani 2023.
[15] Cf. Rother et al. 2024.
[16] Cf. Mariani 2023.
[17] Cf. Kuhn et al. 2018.

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] [12]

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 still requires technical expertise.[21] [13]

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. [14]

# 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. [15]

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 [16]

[18] Cf. Shoilee et al. 2023.
[19] Cf. Oldman / Tanase 2018.
[20] Cf. Daquino et al. 2023.
[21] Cf. Massari / Peroni 2025.

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.

[17] 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. Next, users select a data license for 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). 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.[22] Finally, users must enter a project *URI (Uniform Resource Identifier)*, which uniquely identifies LOD produced within PROV-A.

[18] 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.

[19] 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.[23]

[22] Cf. Open Definition.
[23] The PROV-A JSON schema defines the structure and validation rules for representing provenance data in JSON format, ensuring compliance with the PROV-A interface.

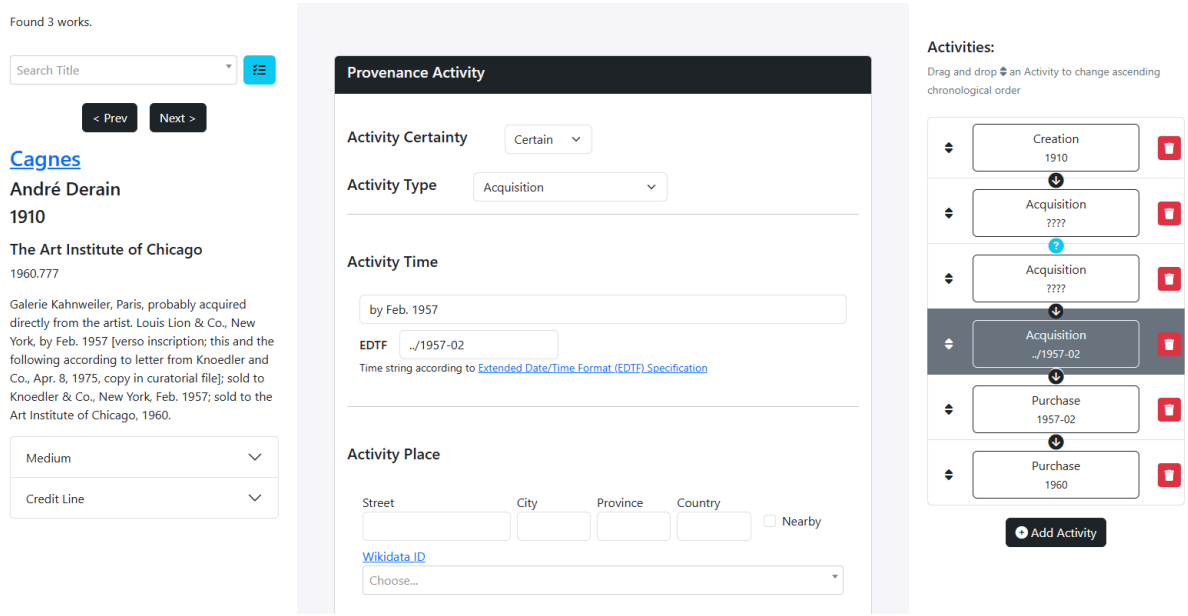


Figure 1: PROV-A data structuring interface. [Screenshot: Fabio Mariani 2025]

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

[21] 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.[24] 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.

[22] 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.

[24] Extended Date / Time Format (EDTF) Specification.

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, 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. [23]

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. 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. [24]

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. This guarantees that users retain their progress across sessions, ensuring a seamless workflow. Additionally, users can download their data in JSON format for backup. [25]

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)*.[25] 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. [26]

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 compliance with CIDOC-CRM.[26] 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.[27] The publication info graph contains metadata about the nanopublication itself, structured according to the Dublin Core Metadata Initiative (DCMI) Metadata Terms. It includes key attributes such as authorship (via ORCID identifiers), Creative Commons licensing, and the nanopublication's creation date. [27]

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 [28]

[25] Verborgh et al. 2024.
[26] The PROV-A RDF shape definition, 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.
[27] Doerr et al. 2024.

browser, PROV-A uses the Quadstore library for managing RDF graph storage through *IndexedDB*, a client-side database API. Quadstore also integrates Comunica, a framework for querying knowledge graphs through *SPARQL (SPARQL Protocol and RDF Query Language)*.[28]

[29] 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.[29] 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 incomplete. 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.

# 4. Case Study: PROV-A as Human-in-the-Loop Tool

[30] 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.

[31] The case study draws on provenance data from the Art Institute of Chicago (AIC), which was used in a prior experiment evaluating NLP techniques.[30] 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.[31] The SBD model achieved an F1 score of 0.99, while the span categorization model reached an F1 score of 0.94.[32]

[32] 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.[33] 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.

[28] Taelman et al. 2018.
[29] Cf. Mariani 2023.
[30] Cf. Rother et al. 2023.
[31] Cf. Mariani et al. 2023.
[32] 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.
[33] The case study material is available on the PROV-A GitHub repository.

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 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. [33]

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. [34]

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. [35]

Among the 1,166 provenance events recorded in the case study, 1,036 include a date reference, accounting for 89 % of all activities (Table 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. [36]

| **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 1: Temporal approximations in provenance events for artefacts classified as paintings in the Modern Art department of the Art Institute of Chicago.

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«). [37]

[38] 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.

[39] 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.

[40] 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.

[41] 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.[34]

[42] 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.

[43] 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.« [35]
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.

[34] Cf. Mariani 2023.
[35] Provenance text of the painting *Cagnes* (André Derain, 1910) published on the Art Insitute of Chicago website: »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).

# 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 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. [44]

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. [45]

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. [46]

# 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. [47]

# Bibliography


Tracey Berg-Fulton / David Newbury / Travis Snyder: Art Tracks: Visualizing the Stories and Lifespan of an Artwork. In: Museums and the Web 2015 (MW2015, Chicago, 08.–11.4.2015). Silver Spring, US-MD 2015. [online]

Marilena Daquino / Mari Wigham / Enrico Daga / Lucia Giagnolini / Francesca Tomasi: CLEF. A Linked Open Data Native System for Crowdsourcing. In: Journal on Computing and Cultural Heritage 16 (2023), no. 3. DOI: 10.1145/3594721

Martin Doerr: The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. In: AI Magazine 24 (2003), no. 3, pp. 75–92. DOI: 10.1609/aimag.v24i3.1720

Martin Doerr / Christian-Emil Ore / Pavlos Fafalios / Athina Krisotaki / Stephen Stead: Definition of the CRMinf: An Extension of CIDOC-CRM to Support Argumentation. Version 1.1: September 2024. PDF. [online]

Patricia Harpring: Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA. In: Art Documentation: Journal of the Art Libraries Society of North America 29 (2010), no. 1, pp. 67–72. DOI: 10.1086/adx.29.1.27949541

Tobias Kuhn / Albert Meroño-Peñuela / Alexander Malic / Jorrit H. Poelen / Allen H. Hurlbert / Emilio Centeno Ortiz / Laura Inés Furlong / Núria Queralt-Rosinach / Christine Chichester / Juan M. Banda / Egon Willighagen / Friederike Ehrhart / Chris Evelo / Tareq B. Malas / Michel Dumontier: Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. In: 2018 IEEE 14th International Conference on e-Science (Amsterdam, 29.10.–01.11.2018). New York 2018, pp. 83–92. DOI: 10.1109/eScience.2018.00024

Anne Luther: Digital Provenance, Open Access, and Data-Driven Art History. In: Kathryn Brown (ed.): The Routledge Companion to Digital Humanities and Art History (= Routledge Art History and Visual Studies Companions). New York 2020, pp. 448–458. DOI: 10.4324/9780429505188-38

Fabio Mariani: Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data. In: Proceedings of the Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022, Lausanne, 09.–10.06.2022). Aachen 2023, pp. 63–84. PDF. [online]

Fabio Mariani / Lynn Rother / Max Koss: Teaching Provenance to AI: An Annotation Scheme for Museum Data. In: Sonja Thiel / Johannes C. Bernhardt (eds.): AI in Museums. Reflections, Perspectives and Applications (= Edition Museum, 74). Bielefeld 2023, pp. 163–172. DOI: 10.14361/9783839467107-014

Arcangelo Massari / Silvio Peroni: HERITRACE: A User-Friendly Semantic Data Editor with Change Tracking and Provenance Management for Cultural Heritage Institutions. In: Umanistica Digitale 9 (2025), no. 20, pp. 317–340. 2025. DOI: 10.6092/issn.2532-8816/21218

David Newbury: Art Tracks: Using Linked Open Data for Object Provenance in Museums. In: Museums and the Web 2017 (MW17, Cleveland, 19.–22.04.2017). Silver Spring, US-MD 2017. [online]

David Newbury: LOUD: Linked Open Usable Data and linked.art. In: CIDOC 2018 Conference (Heraklion, 29.09.–05.10.2018). 15.09.2018. PDF. [online]

David Newbury / Louise Lippincott: Provenance in 2050. In: Jane Milosch / Nick Pearce (eds.): Collecting and Provenance. A Multidisciplinary Approach. Lanham, US-MD 2019, pp. 101–109. [Nachweis im GVK]

Dominic Oldman / Diana Tanase: Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In: Denny Vrandečić / Kalina Bontcheva / Mari Carmen Suárez-Figueroa / Valentina Presutti / Irene Celino / Marta Sabou / Lucie-Aimée Kaffee / Elena Simperl (eds.): The Semantic Web – ISWC 2018. 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II (= Lecture Notes in Computer Science, 11137). Cham 2018, pp. 325–340. DOI: 10.1007/978-3-030-00668-6_20

Lynn Rother / Max Koss / Fabio Mariani: Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums. In: Emily Lew Fry / Erin Canning (eds.): Perspectives on Data. Chicago 2022. DOI: 10.53269/9780865593152/06

Lynn Rother / Fabio Mariani / Max Koss: Hidden Value: Provenance as a Source for Economic and Social History. In: Jahrbuch für Wirtschaftsgeschichte / Economic History Yearbook 64 (2023), no. 1, pp. 111–142. DOI: 10.1515/jbwg-2023-0005

Lynn Rother / Fabio Mariani / Max Koss: Interpreting Strings, Weaving Threads: Structuring Provenance Data with AI. In: Katharina Günther / Stefan Alschner (eds.): Sammlungsforschung im digitalen Zeitalter. Chancen, Herausforderungen und Grenzen. Göttingen 2024, pp. 93–103. DOI: 10.15499/kds-005-008

Jeffrey Smith: Toward »Big Data« in Museum Provenance. In: Giovanni Schiuma / Daniela Carlucci (eds.): Big Data in the Arts and Humanities. Theory and Practice (= Data Analytics Applications). New York 2018, pp. 41–50. [Nachweis im GVK]

Sarah Binta Alam Shoilee / Victor de Boer / Jacco van Ossenbruggen: Polyvocal Knowledge Modelling for Ethnographic Heritage Object Provenance. In: Maribel Acosta / Silvio Peroni / Sahar Vahdati / Anna-Lisa Gentile / Tassilo Pellegrini / Jan-Christoph Kalo (eds.): Proceedings of the 19th International Conference on Semantic Systems (SEMANTICS 2023, Leipzig, 20.–22.09.2023). Amsterdam 2023, pp. 127–143. DOI: 10.3233/SSW230010

Ruben Taelman / Joachim Van Herwegen / Miel Vander Sande / Ruben Verborgh: Comunica: A Modular SPARQL Query Engine for the Web. In: Denny Vrandečić / Kalina Bontcheva / Mari Carmen Suárez-Figueroa / Valentina Presutti / Irene Celino / Marta Sabou / Lucie-Aimée Kaffee / Elena Simperl (eds.): The Semantic Web – ISWC 2018. 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II (= Lecture Notes in Computer Science, 11137). Cham 2018, pp. 239–255. DOI: 10.1007/978-3-030-00668-6_15

United States Department of State, Office of the Special Envoy for Holocaust Issues: Washington Conference Principles on Nazi-Confiscated Art. 03.12.1998. [online]

Ruben Verborgh / Ruben Taelman / Jesse Wright / Simon Van Braeckel / Laurens Rietveld / Danny Hurlburt / Kuno Woudt / Thomas Bergwinkl / Vincent / Thomas Tanon / Samuel Roze / Noel De Martin / Jim Smart / Martin Maillard / Martin / Mohammad Fathi / Pieter Colpaert / Pieter Heyvaert / Ruben / Shawn / Wes Turner / alxflam / elf Pavlik / Ludovic Roy / Jitse De Smet / Jacopo Scazzosi / Iddan Aaronsohn / Howard Zuo / Gregor Middell: rdfjs/N3.js. In: Zenodo. Version 1.17.3: 24.03.2024. DOI: 10.5281/zenodo.10866356

Mark D. Wilkinson / Michel Dumontier / IJsbrand Jan Aalbersberg / Gabrielle Appleton / Myles Axton / Arie Baak / Niklas Blomberg / Jan-Willem Boiten / Luiz Bonino da Silva Santos / Philip E. Bourne / Jildau Bouwman / Anthony J. Brookes / Tim Clark / Mercè Crosas / Ingrid Dillo / Olivier Dumon / Scott Edmunds / Chris T. Evelo / Richard Finkers / Alejandra Gonzalez-Beltran / Alasdair J. G. Gray / Paul Groth / Carole Goble / Jeffrey S. Grethe / Jaap Heringa / Peter A. C. 't Hoen / Rob Hooft / Tobias Kuhn / Ruben Kok / Joost Kok / Scott J. Lusher / Maryann E. Martone / Albert Mons / Abel L. Packer / Bengt Persson / Philippe Rocca-Serra / Marco Roos / Rene van Schaik / Susanna-Assunta Sansone / Erik Schultes / Thierry Sengstag / Ted Slater / George Strawn / Morris A. Swertz / Mark Thompson / Johan van der Lei / Erik van Mulligen / Jan Velterop / Andra Waagmeester / Peter Wittenburg / Katherine Wolstencroft / Jun Zhao / Barend Mons: The FAIR Guiding Principles for Scientific Data Management and Stewardship. In: Scientific Data 3 (2016). DOI: 10.1038/sdata.2016.18

Nancy H. Yeide / Konstantin Akinsha / Amy L. Walsh: The AAM Guide to Provenance Research. Washington 2001. [Nachweis im GVK]

# List of Figures and Tables