Archival Science (2025) 25:16
https://doi.org/10.1007/s10502-025-09485-x

ORIGINAL PAPER

Overcoming data siloes in cultural heritage crime research: a consolidated OSINT-derived dataset on art, antiquities, and the trade in cultural goods

Madison Leeson[1] · **Riccardo Giovanelli**[1] · **Sara Ferro**[1] · **Michela De Bernardin**[1] · **Arianna Traviglia**[1]

Accepted: 19 April 2025

Abstract

The current landscape for provenance researchers, cultural heritage crime analysts, and law enforcement working in the culture sector is characterised by the siloing of data across dozens of databases, resulting in fragmented and incomplete resources that must be manually correlated and validated to provide insights into cultural heritage crime. The European Union-funded Research, Intelligence, and Technology for Heritage and Market Security (RITHMS) project is developing a platform to assist law enforcement agencies across Europe in tackling the illicit trafficking of cultural goods by aggregating open, specialised, and police data from a range of sources. This article outlines one of the initial phases of data collection, which has developed 30 tailored web scrapers for the collection of data from existing databases of stolen, missing, protected, and unprovenanced cultural goods. This has resulted in the largest known non-police dataset of these and associated data objects addressing the real-world challenge of data siloes in heritage crime and provenance research. This article details the multi-step process of data collection and pre-processing that has produced the novel consolidated dataset. The mechanism for knowledge discovery developed during this project has immediate applications and has resulted in actionable intelligence for the investigation of cultural goods crimes, highlighting the value of consolidated data as a resource. This research also offers a cursory analysis of the resulting dataset, demonstrating how mining of this resource can enable new scientific insights and offer promising opportunities for intelligence-led policing of cultural heritage crimes.

Keywords Data management · Web scraping · Provenance · Python · Looted art · Data mining

Extended author information available on the last page of the article

Published online: 05 June 2025

Introduction

Dozens of databases exist for investigating stolen and looted cultural goods, from Interpol's Stolen Works of Art Database to private police repositories and specialised archives of Nazi-looted artworks and other wartime losses. As a result, cultural heritage crime analysts, provenance researchers, and law enforcement agencies (LEAs) are obligated to painstakingly identify and sift through numerous databases managed by a broad number of actors, navigating different user interfaces and access restrictions, to acquire limited information, which is often isolated from other available open data on these objects and the networks that facilitate their circulation. This situation is inefficient and deeply flawed, as it limits the ability for actors in the field of cultural heritage crime research and recovery to acquire the information they need. Companies such as the Art Loss Register have emerged to address this climate, with dedicated full-time staff who will take on the task of manually searching these databases (as well as their own) for any available information on objects of interest. Despite admissions in scholarship that a unique, consolidated dataset is needed to address this problem, none has yet been produced (Foley 2014; Levahi 2023).

To meet this need, within the framework of the European Union-funded Research, Intelligence, and Technology for Heritage and Market Security (RITHMS) project,[1] we have developed a suite of Python-based web scrapers to gather, aggregate, and minimally post-process data from 30 of the largest and most frequently accessed open databases of missing, stolen, unprovenanced, and protected cultural objects. The present study first outlines the development of these scrapers and the logic, conventions, and methodology that guided their creation. While the development of scrapers to feed the RITHMS platform was our initial objective, this development process produced a consolidated dataset of over two million entities (encompassing artefacts as well as associated actors, locations, dates, and events) related to the circulation of problematic cultural goods, producing the largest known non-police dataset of these items. Although the RITHMS platform is being developed for exclusive use by European LEAs, we anticipate publication of a large portion of the consolidated dataset—depending on data reproduction rights granted by database managing authorities—in the fall of 2025, for use by researchers and the public.

The focus of the present article is this dataset, which represents an unanticipated (but very welcome) asset produced by the data collection task of RITHMS project. To our knowledge, the dataset is only surpassed in size by the Leonardo database of the Italian Carabinieri, which holds records on nearly 7 million artworks, of which approximately 1.3 million have been reported stolen and are currently being investigated (Carabinieri Command for the Protection of Cultural Heritage, n.d., 2021; Rapicavoli, n.d.). This article presents the multi-step process of data collection and reviews the methods of data pre- and post-processing that

[1] RITHMS is developing a software platform which leverages social network analysis and open data (as well as specialised data from private sources) to track the circulation of cultural goods, with the aim of tackling illicit trade and trafficking activities. It is outside the scope of the present article to present the platform in more detail, but more information can be found at www.rithms.eu.

are implemented to prepare the datasets for mining and analysis. The mechanisms for knowledge discovery realised during this phase of the project have immediate applications, consolidating dozens of open-source intelligence (OSINT) resources into one uniformly structured dataset. An example of the insights that can be gleaned from mining the consolidated data is offered in the discussion, based on an initial examination of an eighteenth-century painting looted during the Second World War.

The creation of this dataset offers a significant contribution to archival science while addressing key themes such as access, preservation, ethics, and social justice (Colavizza et al. 2021; Duff et al. 2013). By integrating data from multiple sources, this initiative promotes the principle of interoperability and demonstrates the value of consolidation for archival practices (Romein et al. 2020). It overcomes the siloed nature of many archival collections, enhancing discoverability and providing a comprehensive resource for researchers, law enforcement, and cultural heritage professionals. It also aligns with the continuum model of archival science, emphasising the ongoing lifecycle of records and their role in addressing contemporary issues, for example restitution and decolonisation (Anderson 2024).

From an ethical and social justice perspective, the dataset responds to the growing need for transparency and accountability in the cultural heritage sector (Duff et al. 2013; O'Neill & Stapleton 2022). By structuring and normalising available information on looted and stolen items, it can support efforts to repatriate artefacts to their rightful owners or countries of origin by bringing to light historical injustices which are often obscured by hard-to-navigate web interfaces and opaque or deceptively worded provenance texts (Anderson 2024; Savoy 2015). It also offers opportunities for research into provenanced but still contested objects, for example those acquired during colonial contexts or within times of armed conflict. Further, the dataset addresses the problem represented by fragmented provenance by providing a centralised resource for tracing the ownership history of cultural goods, an important part of combatting illicit trafficking (Savoy 2015). The focus on ownership of cultural goods also reinforces the role of archives in shaping and preserving collective narratives, ensuring that marginalised voices and histories are represented and empowered (Duff et al. 2013). Technologically, the dataset contributes to the aims of digital archiving by integrating data from diverse formats and sources into a single unified resource, leveraging the automation of recordkeeping and the transformation of the archive from a storage tool to a resource to be mined (Colavizza et al. 2021; Moss et al. 2018). This also reflects advances in metadata standardisation, ensuring that semantic information is both interoperable and durable (Colla et al. 2022; O'Neill and Stapleton 2022).

Overall, this research offers a solution to the pressing problem facing provenance and cultural heritage crime research represented by the siloing of data across the field. Through the novel, consolidated dataset, we aim to support provenance research and highlight new opportunities for the investigation of cultural goods crimes through the leveraging of OSINT sources.

Methodology and requirements for data collection

This section will first outline the high-level decisions that informed the data collection process, including the initial definition of data sources (which are outlined in greater detail in Sect. "Databases of cultural goods") and legal and ethical requirements for the project. It then provides a detailed overview of the technical development process for the web scrapers, explaining the decision-making process and discussing specific Python packages used. Lastly, it discusses the post-processing steps that were taken to ensure all the collected data were of a uniform structure and standard, that certain properties (e.g. dates) were normalised, and that data fit into one of the pre-defined ontology classes, all of which was intended to support correlation of the data across data sources.

The development of modules for data collection from databases of cultural goods began with the identification of target sources. For the purposes of the present research, 'cultural goods' refers to any object of cultural, artistic, or historical significance, particularly those which have been commodified and are therefore susceptible to criminal conduct such as theft, undocumented import and export, forgery and counterfeiting, and other illicit activities. In total, 37 promising databases of such objects were identified. Of these, seven were found to not be viable for data collection as they were either private, internal police databases that could not be accessed publicly, or because the managing authority did not have the ability to grant us permission for data collection due to data reproduction rights. Within this group of non-viable sources are the Interpol Stolen Works of Art database and the Art Loss Register (ALR), which are perhaps the two most well-known databases of stolen and missing cultural goods among both the general public and cultural heritage experts. Future improvement of the consolidated dataset produced during this project can focus on the integration of the Interpol database as well as other police and LEA resources, which would further support the objectives set out by the present research. However, it is unlikely that the data of the ALR could be aggregated in the final dataset considering it is a proprietary resource of a for-profit company whose business model relies on siloed data. Excluding these non-viable sources, the remaining 30 databases are presented in greater detail below, in Sect. "Databases of cultural goods".

Before starting data collection, a set of reports assessing the legal and ethical considerations of data collection was developed by the Universidade da Coruña (UDC), a partner in the consortium managing the RITHMS project, to ensure the data collection tools adhere to responsible ethical standards and ensure data privacy and protection.[2] As the technical partner that led the current data collection task, the Italian Institute of Technology (IIT) conducted an extensive analysis of the sources identified to determine whether authorisation was necessary to gather data from the target databases, to respect intellectual property rights and relevant copyright laws, and to ensure the protection of personal data. First, it was determined that individual authorisation from each database managing authority was unnecessary as the project

[2] Reports classified at an open or public security level can be viewed at www.rithms.eu/results.

concerns open, publicly available data, which is solely being used in the scope of scientific research. Despite this, communications were initiated between the authors and each database managing authority to convey the project's intentions and determine whether any concerns existed around the parameters of data collection. This also served to create a network of database managing authorities and actors in the field to further a shared aim of promoting available resources on cultural goods.

The procedure of data collection outlined in this article relied on techniques of web scraping, which are legal for scientific research purposes within the European Union but could have limitations in the context of copyright law, the terms of use of the databases considered, and the General Data Protection Regulation (GDPR). By default, however, copyright law protects not the information itself but the structure and creative investments of the database managing authority. In cases where the information is explicitly under copyright, Italian Legislative Decree 177/2021 exempts scientific research institutions from limitations on data collection and Article 102-3 further clarifies that the activities of extraction or reuse of the database's contents are not subject to the authorisation of the managing authority if the database is already publicly available. In the one case of a database which requires user authentication to access, written permission was received from the managing authority to scrape the contents of the repository using credentials granted to one of the authors. A review was also conducted of the terms of use for each data source, all of which (if available) permitted data collection for non-commercial research purposes, provided that proper credit is given to the original source. Lastly, in terms of the GDPR, data processed include 'personal data' such as names, surnames, birth dates, and locations of artists, buyers, owners, donors, and service providers. The processing of such publicly available data was conducted by IIT under the auspices of RITHMS and on the basis of its *legitimate interest* of carrying out research.

Module development

The web scraping programmes created for this project were written in the Python programming language. Each followed an extract-transform-load (ETL) pattern where data were first collected from the source using either the Requests and BeautifulSoup packages, or Selenium, or a combination of the three. These different approaches depended on whether a database had implemented JavaScript in its source code, and if so, to what degree. In cases where dynamic web content required the use of the Selenium package, we also developed a function to automatically update the web driver (in our case, Chrome) to ensure the tool continues functioning as regular browser updates are rolled out.

While collecting data, original page metadata were retained and added as corresponding properties to the output, to safeguard retention of creation date, author, and other valuable information. To ensure scraped data can always be linked to the source from which it was taken, two additional properties to the output data were introduced: 'datasource_id' and 'tool_id'. Both are unique values and identify, as their names would suggest, the source from which data were

gathered and the web scraper used to collect it. In the case of the Obiecte Furate database, for example all scraped data includes the properties:

'datasource_id': 'obiecte_furate_database'
'tool_id': 'obiecte_furate_scraper'

After scraping, the data were structured in a Pandas DataFrame for initial pre-processing, the removal of superfluous characters, limited translation of non-English data (using the Google Translator library of the Deep-Translator package), and deterministic or rule-based cleaning to normalise the data. Thus transformed, each dataset was then mapped to a common structure which distilled individual entities, namely people, artefacts, dates, locations, and events, and saved in JSON format for consolidation and mining. To ensure the resilience and efficiency of the system, a series of parallel scrapers was implemented with each customised for a unique data source and following the ETL pattern described above. This system evenly distributed the processing load across multiple tools to avoid straining any one individual scraper, to allow expansion as new targets are integrated, and to prevent a single error from causing a system-wide disruption.

To mimic normal visitor behaviour and not overload the host, the decision was made to limit the rate of requests sent to the target servers. This was implemented using time.sleep() and driver.implicitly_wait(). Further, each module includes a function to randomise the user-agents of the browser sending requests, thereby protecting the real user-agent, which may contain sensitive information on the browser or operating system of the requestor, ensuring adherence to security best practices. To further safeguard data integrity and avoid man-in-the-middle attacks (where communications may be interrupted or intercepted), a thorough manual examination was conducted of data source Root Certificates, which validated each server's SSL certificate (where necessary) against a list of trusted Certificate Authorities using the Certifi package.

Additional functions were implemented in each tool to ensure the standardisation of data collection and pre-processing and to facilitate monitoring and maintenance of the tools. The first of these was a status log that periodically prints time-stamped updates to ensure the tool was running as planned and had not met any errors. Further, a checksum method was developed to determine whether any changes had been made to a data source since the last time it was scraped, to avoid redundant data collection and processing. This used the Hashlib package to generate a hashed string of the page's source HTML, which can be consulted again in future operations to ensure no page is processed twice, thus minimising memory usage and the strain on the host server. The Hashlib package was also employed to generate universally unique identifiers (UUIDs) for each scraped data object based on a hash of the object's permanent URL; this facilitated the correlation of entities and ensures that each object remains linked to the database record page from which the original information was gathered. Each UUID comprises a 15-digit numeric hash appended to a source-specific prefix used to identify the database. For example, the UUID of an object retrieved from the Federal Bureau of Investigation (FBI) National Stolen Art File may have the UUID 'FBI_146329830573183'.

Data post-processing

Entities identified during the data collection phase were mapped to a domain-specific ontology that was developed for the purposes of RITHMS by the European Software Institute (ESI), the technical partner overseeing the development of the platform. This leveraged existing cultural heritage ontologies, for example the International Committee for Documentation's (CIDOC) object-oriented Conceptual Reference Model, as well as the scope of the gathered data, to classify entities such as actors, organisations, locations, and events according to a limited set of object classes and subclasses. These allow for the conceptual organisation of entity subtypes in a constructive way, enabling, for example, the grouping of all objects of type 'painting', all organisations of type 'museum', and all locations of type 'city'. It should be noted that the ontology is a work in progress and will continually be updated and expanded as additional sources are identified and gathered. Current ontology classes have been defined and are added as necessary based on the needs of the data. Super classes of the ontology, within which more specific child classes are defined, have been listed in Table 1.

A set of common properties has been defined for each class to aid correlation of objects across data sources and support research into the dataset. For objects of type 'Activity', the only common property is 'description', which describes the event in a free text string (e.g. providing information about an exhibition or restoration campaign). Entities of type 'Actor' have considerably more properties, depending on what has been made available in each data source. Further, while person entities might have a date and place of birth, for example this would not be common to the entire 'Actor' category as it would not be informative for organisations. Rather, the common properties for all actors are: name, alias (representing alternative names, including colloquial names and multi-language aliases), location, contact (email or phone number, if available), and domain (representing subject matter expertise). Additional properties are made available on a case-by-case basis depending on the original data source. 'Artefact' and 'File' classes have the same common properties, namely: title, date of creation, external URL (if an image or digital record is available), artist/author, object type, and description. Conversely, 'Location' entities have just three properties: name, longitude, and latitude—though in most cases only a name is available. Location entities are enriched further through links to other location entities,

Table 1 Ontology classes defined for the RITHMS project

Super class	Sample child classes
Activity	Destruction, Exhibition, Reconstruction, Repair, Restoration, Theft
Actor	Organisation: Auction House, Company, Gallery, Library, Place of Worship Person: Artist, Collector, Expert, Government Official, Seller, Student
Artefact	Book, Clothing, Drawing, Furniture, Jewellery, Painting, Sculpture, Vessel
File	Document, Email, Record
Location	Address, Archaeological Site, City, Country, Region
Temporal Definition	Date, Period

which represent hierarchies of place; for example, 'Paris' (#City) '#hasCountry' (is related to) 'France' (#Country). Lastly, entities of 'Temporal Definition' have just one common property: date. This represents the original date as provided in the original data source. These are of variable structure, for example '10-05-2020', 'XVII sec.', 'August 16th 1970'. For entities of type 'Date', a property was also added to convey the normalised value in ISO date time format (YYYY-MM-DD), but this was often not possible for 'Period' entities, which range from temporally vague (e.g. Bronze Age) to extremely precise (10:15–10:45 on 12th April 2005).

Lastly, a set of data cleansing functions, consisting entirely of rule-based statements, performed normalising transformations of the data to optimise further the comparison of objects from disparate sources and the potential for mining the resulting consolidated dataset. For example, information determined to be most valuable for correlation, namely object provenance, type, and materials, were translated into English (if not already) and standardised with uniform spelling conventions—in our case, conforming to UK English. Similarly, individual values were extracted from the object dimensions, using keywords to detect, extract, and tag length, width, depth, diameter, and weight values, if provided. This enabled the sorting of artefacts by each of these dimensions, for example to return all objects larger than 15 cm, heavier than 10 kg, and so on. Lastly, a function was developed that leveraged the Geopy package to identify and normalise city, region, and country names and identifiers. For countries, in addition to the official and common names (e.g. 'Republic of Armenia' and 'Armenia'), the function also returns the two- and three-digit country codes (e.g. 'AM' and 'ARM'), multi-language aliases of the country name (e.g. 'Armenia', 'Armenië', 'Αρμενία'), and the country's numeric calling code. This consolidated the most common referents of each location entity, enabling the correlation of these values as referring to the same data object. Together, these processing steps enriched the final consolidated dataset by improving the compatibility of information gathered from the previously isolated sources.

Databases of cultural goods

All data collection conducted for this research occurred within the scope of RITHMS project, and while it is outside the scope of the present paper to discuss the project in further detail, it should be recognised that its geographic focus on Europe was the leading factor in the decision to prioritise European databases. That being said, select non-European databases (e.g. art and artefacts in the United States, Chile, and Iraq) were included because they hold objects which are at particularly high risk of being trafficked into Europe. Further, in many cases, they contribute additional information on objects also recorded in European databases, adding non-European perspective to aid in the triangulation and enrichment of a 'ground truth', or evidenced history, for these problematic items.

Table 2 lists the 30 databases targeted by the present study and the scope (geographic and thematic) of each. Of these, 14 are national and international repositories of stolen objects from Romania, Bosnia and Herzegovina, Spain, the United States (US), Chile, Ukraine, and Iraq, and unprovenanced or protected objects

Table 2 Databases targeted by the data collection modules

Category	Database (managing authority)	Scope of the data
Stolen, unprovenanced, or protected objects	Obiecte Furate (Romanian Police)	[Romania] Artefacts stolen from Romanian individuals and museums, including descriptive data about the object, artist if known, and circumstances of theft if known
	Securius (Federal Criminal Police Office)	[Germany] Objects that have been confiscated by German police within the scope of other police activities or investigations. Basic object information is provided as well as the location of the seizure
	Index of Damaged, Stolen, Missing or Illegally Exported Movable Cultural Objects (National Heritage Institute)	[Romania] Stolen and missing art and artefacts from Romania, including information on the original museum and in some cases on the theft event
	Database of Stolen/Missing Art in Bosnia and Herzegovina (Center Against Trafficking in Works of Art)	[Bosnia and Herzegovina] Art and artefacts reported stolen from individuals and organisations in Bosnia and Herzegovina. Basic descriptive object information provided, in some cases also the date of theft
	Works of Art–Illegal Art Trade (Ministry of Internal Affairs)	[Bosnia and Herzegovina] Art and artefacts reported stolen from individuals and organisations in Bosnia and Herzegovina. Basic descriptive object information provided, in some cases also the date of theft
	Stolen Works of Art* (Guardia Civil)	[Spain] Artworks reported as having been stolen from individuals and organisations in Spain
	National Stolen Art File (Federal Bureau of Investigation)	[USA] Art and artefacts valued at over $2,000 USD and reported stolen in the United States. Object information includes title, description, and material/object type
	SURDOC* (Heritage Assets Documentation Centre)	[Chile] Art and artefacts reported stolen from museums and organisations in Chile. Descriptive information is provided on the object and the organisation from which it was stolen

Table 2 (continued)

Category	Database (managing authority)	Scope of the data
	Stolen Cultural Assets Database (National Cultural Heritage Service)	[Chile] Art and artefacts reported stolen from individuals and organisations in Chile. Descriptive object information provided, and in most cases, details are also provided on the theft event, including date and location
	Stolen Heritage* (National Agency on Corruption Prevention)	[Ukraine, Russia] Art and artefacts that have been looted from Ukrainian museums, galleries, and archaeological sites by Russian soldiers during the invasion and occupation. Detailed object information is provided as well as the circumstances of the theft
	War & Art* (National Agency on Corruption Prevention)	[Ukraine, Russia] Information on artworks owned by individuals who have been sanctioned for their support of the Russian invasion of Ukraine. Object information includes artist, date, material, and size as well as details on the owner and circumstances of acquisition (including date and facilitator, e.g. Sotheby's auction house)
	Registry of New Acquisitions of Archaeological Material and Works of Ancient Art (Association of Art Museum Directors)	[North America] Art and artefacts acquired by North American museums and galleries that lack sufficient pre-1970 provenance. Information provided includes artist (if known) or country/culture of origin, exhibition and publication history, provenance, and descriptive data
	Register of Protected Cultural Property and Collections (Ministry of Education, Culture, and Science)	[The Netherlands] Cultural objects registered as protected, meaning they cannot be sold without national/official approval. Information includes descriptive data, provenance if known, current ownership/location, and artist or country/culture of origin
	Iraq Museum Database (University of Chicago Oriental Institute)	[Iraq] Basic object information provided about artefacts that were registered in Iraqi museums before 2003 (and thus which cannot appear legally on the international market)
WWII-looted goods	Lost Art Database (German Lost Art Foundation)	[Europe] Artworks and artefacts looted during WWII. Records include descriptive object information, provenance, and contact information for either the current location or the person who filed the missing object request

Table 2 (continued)

Category	Database (managing authority)	Scope of the data
	Central Registry of Information on Looted Cultural Property, 1933–1945 (Commission for Looted Art in Europe)	[Europe] Artworks and artefacts looted during WWII. Records include descriptive object information, provenance, and contact information for the current location/organisation who filed the record
	Rose Valland MNR-Jeu de Paume (Ministry of Culture)	[Europe] Artworks and artefacts looted during WWII. Descriptive information is provided about each object as well as provenance if known and circumstances of the confiscation and later transportation
	Degenerate Art Database (Free University of Berlin, Germany)	[Europe] Artworks confiscated by the Nazis as 'degenerate art'. Descriptive information is provided on the art and artefacts that were not destroyed, including provenance data, context on the original confiscation, exhibition history, and biographical information on the artist
	Catalogue of Wartime Losses (Ministry of Culture and National Heritage)	[Poland] Objects that were looted from Poland during WWII. Descriptive information is provided about the objects as well as the current rightful owner/heir. If the item has been recovered, context is provided on the recovery/restitution
	Cultural Goods of the Second World War (Ministry of Education, Culture, and Science)	[The Netherlands] Objects in Dutch collections that were looted during WWII or have gaps in their wartime provenance. Descriptive information is provided on art and artefacts, including a detailed and structured provenance history
	Reichskunstdepot Kremsmünster–Paintings* (Commission for Provenance Research)	[Europe] Artworks and artefacts looted during WWII; descriptive information includes object type, artist name, title, description, provenance, observations/comments, and inventory numbers
	Museum Acquisitions from 1933 Onwards* (Dutch Museums Association)	[The Netherlands] Objects in Dutch museums with gaps in wartime provenance. Descriptive object information is provided as well as information on the current location
	Nazi-Era Provenance Internet Portal* (American Alliance of Museums)	[USA] Objects in American (US) museums with gaps in wartime provenance. Descriptive object information is provided as well as information on the current location

Table 2 (continued)

Category	Database (managing authority)	Scope of the data
	Einsatzstab Reichsleiter Rosenberg Collection (Institute for War, Holocaust, and Genocide Studies)	[Europe] Artworks and artefacts looted during WWII. Descriptive information is provided about each object as well as provenance if known, circumstances of the confiscation and later transportation, and details on individuals who owned major collections which were subject to seizure
	Linz Collection (German Historical Museum)	[Europe] Artworks confiscated by the Nazis during WWII. Detailed object information includes provenance history and context on the circumstances of the confiscation and subsequent transport
	Herman Goering Art Collection (German Historical Museum)	[Europe] Artworks confiscated by Nazi officer Hermann Goering during WWII. Detailed object information includes provenance history and context on the circumstances of the confiscation and subsequent transport
	Tableau et Dessin Database (Commission for the Compensation of Victims of Spoliation)	[France] Artworks and artefacts looted during WWII; descriptive information includes object type, artist name, title, description, provenance, observations/comments, and inventory numbers
	Max Stern Art Restitution Project (Concordia University)	[Europe] Artworks confiscated or sold under duress during WWII by dealer and gallerist Max Stern. Available data covers object description, exhibition and publication history, and known provenance
Provenance	Proveana (German Lost Art Foundation)	[Europe] Individuals, events, organisations, literature, and other entities associated with the looting of cultural goods during WWII and their subsequent circulation. Available information includes biographical details (e.g., date and place of birth, work, death), related entities, and insights from provenance research projects conducted by the German Lost Art Foundation
Sanctioned individuals	War & Sanctions* (National Agency on Corruption Prevention)	[Ukraine, Russia] Information on individuals and organisations either sanctioned for their support of the Russian invasion of Ukraine or identified for future sanctions. Detailed biographical information is provided about each person as well as their known connections (familial and professional, referring to other people in the database)

from Germany, the Netherlands, Russia, and North America. An additional 14 are national and international databases of cultural goods that were looted or sold under duress during the Second World War (hereafter 'WWII-looted goods'). To these 28 databases are added Proveana and War & Sanctions, two repositories of provenance and individuals, respectively, managed by authorities that also manage databases of cultural goods; considering the structure of these sources closely mirrors that of the other databases already integrated, it was deemed prudent to include these targets as well.

Regrettably, the databases marked with an asterisk in the second column of Table 2 have either been taken down or archived since the initial data collection. In these cases, the relevant web scrapers encountered errors during their most recent data collection attempts, and a manual investigation determined that the tools had become obsolete due to the decommissioning of the original data sources. In some cases, such as with the three databases managed by Ukraine's National Agency on Corruption Prevention, the repositories are being moved to another host server and are anticipated to be relaunched in future, at which point the original tool will be modified to collect data from the new source. With others, for example the American Alliance of Museums' Nazi-Era Provenance Internet Portal that was archived in mid-2024, the managing authority has deemed the database obsolete, given recent developments in museum practices and new, more effective technological resources for researching museums' object provenance (American Alliance of Museums, n.d.). Regardless of the reason, this highlights an additional valuable application for the RITHMS consolidated dataset: as an archive of public data that could outlast the websites where the databases were originally hosted. To ensure this potential is realised, we anticipate the publication of a majority of the consolidated dataset following the end of the project development period (in or shortly after September 2025), to promote open access to the data and encourage mining of its contents. It should be noted, however, that the full dataset cannot be reproduced due to limitations on the reproduction permissions granted by some database managing authorities.

While each database target has a unique thematic focus—including stolen art, police-confiscated items, missing objects, and more—one common theme is the lack of sufficient provenance, or information pertaining to the ownership history of these items. In the context of cultural heritage, 'provenance' refers to documented or verifiable information about their origin, including where they were discovered or produced, their historical context, and their subsequent chain of ownership (Anderson 2024; Sweeney 2008). Unprovenanced artefacts present ethical and legal challenges for research on cultural heritage because their unknown histories may conceal illicit activities such as looting, smuggling, or transfer of ownership under duress within a context of colonialism or conflict (Brodie 2011b; Savoy 2015). In contrast, *under*-provenanced artefacts have some documentation, but the information is incomplete, inconsistent, or insufficient to establish a clear and legitimate chain of custody (Gerstenblith 2019; Ruiz Romero 2020). These artefacts may have ambiguous origins or gaps in their known ownership history, making it difficult to assess their legality and cultural significance.

Even artefacts that are completely provenanced, meaning their ownership and transfer history is fully documented, can still be contested. This often occurs when

legal ownership does not align with ethical or cultural considerations (Bach 2024; La Follette 2017). For example, artefacts acquired during colonial periods may have clear records of transfer but remain contentious due to the exploitative contexts of their acquisition (Breske 2018; Brodie 2018; Green 2017). Similarly, some artefacts may be sold or donated legally yet still provoke claims for restitution or repatriation by their communities or nations of origin (Graham et al. 2023; Lowenthal 1998). As a result, provenance is not just a matter of recordkeeping but has implications for governance, cultural identity, and stewardship. By consolidating all available sources of provenance on objects particularly susceptible to looting and illicit trade, we anticipate that the consolidated dataset can contribute to a better understanding of the historical ownership of these objects, supporting claims for restitution and recovery.

The following section briefly presents and discusses the databases identified for this project as well as the extent of information provided in each. Sources are roughly grouped by region of scope, comprising South America, North America, the Middle East, and Europe, in that order.

Only two relevant databases were identified for stolen objects in South America, both holding records of objects stolen from individuals and institutions in Chile. The two registries—SURDOC and the Stolen Cultural Assets Database—are managed by the National Cultural Heritage Service, which also promotes collaboration and communication among a network of museums, libraries, and archives in the country. The two databases provide detailed information for each object, and the former also includes records on individual theft events.

In North America, the Association of Art Museum Directors (AAMD) operates two databases of art objects in museum collections that lack sufficient post-1970 provenance (referring to the date of the landmark UNESCO Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property). A single web scraper was developed that gathers data from both repositories, collecting all the records in the New Acquisitions of Archaeological Material and Works of Ancient Art, and Resolutions of Claims for Nazi-Era Cultural Assets registries. These artefacts lack sufficient provenance which, relying on the guidelines of the 1970 UNESCO Convention, could demonstrate that they were either already outside of their countries of origin by 1970 or that they were exported legally after this date. Considering these objects are under-provenanced, their integration in the consolidated dataset offers promising opportunities for correlation with databases of missing artefacts elsewhere in North America as well as in South America, Europe, and the other regions represented in the dataset. The AAMD also managed, in collaboration with the American Alliance of Museums, the Nazi-Era Provenance Internet Portal, which, before it was archived in mid-2024, held records on art objects in American museum collections 'that changed hands in Continental Europe during the Nazi era' and lack provenance from this period (American Alliance of Museums n.d.). The integration of this data source is valuable for the same reason mentioned above: it constitutes a dataset of under-provenanced cultural goods that can be compared with looted art databases to foster knowledge discovery on illicit cultural items. Another relevant database in this category is the National Stolen Art File operated by the United States' FBI. While the

database scope is somewhat limited because it only holds artworks with a reported value greater than $2,000 USD, it is an important resource for identifying stolen cultural goods that are at risk of being trafficked out of the country and into European markets.

In the same vein, Middle Eastern and North African cultural goods have been recognised to be of particularly high demand in Europe, especially from countries that have experienced extended periods of armed conflict such as Syria, Iraq, Yemen, Egypt, and Libya (Brodie 2011a, b; Brodie and Manivet 2017). While no national databases of looted or missing cultural goods could be identified for these countries,[3] a database of Iraqi cultural goods developed by researchers at the University of Chicago's Oriental Institute was located. Although seriously outdated (the database has not been maintained since 2008), the Lost Treasures from Iraq database is still a useful reference for artefacts such as cylinder seals, ancient lamps, and cuneiform tablets that were known to be in Iraqi museum collections as of 2003. Their presence on an international market would, as a result, immediately indicate illicit activity.

Indeed, cultural heritage is particularly vulnerable during extended periods of conflict. Following the start of Russia's 'special military operation' in Ukraine in February 2022, Ukraine's National Agency on Corruption Prevention (NACP) established three databases to complement the implementation of sanctions on individuals and legal entities that have endorsed the conflict (Leeson et al. 2024). The first, 'War & Art', stored records on cultural objects known to be in the collections of sanctioned individuals, to prevent their being sold in violation of economic restrictions. The second, 'Stolen Heritage', held records of cultural goods illegally removed from Ukrainian museums or archaeological excavations in occupied territories. The third, 'War & Sanctions', was a comprehensive database of individuals and legal entities under sanctions or who had been identified by the NACP for the future imposition of sanctions. It should be noted that, at the time of writing, all three databases have been taken down and are currently in the process of being transferred to the server of Ukraine's Main Directorate of Intelligence, which will be the managing authority of the databases going forward. In the meantime, cached versions of the databases can be accessed using the Internet archive Wayback Machine.

Other databases integrated in this phase of data collection include national repositories of stolen and looted cultural goods from Romania (the National Heritage Institute's Index of Movable Cultural Goods and the Obiecte Furate database of the Romanian Police), Spain (the Guardia Civil's Stolen Works of Art database), and Bosnia and Herzegovina (the Database of Stolen/Missing Art, managed by the Center Against Trafficking in Works of Art, and the Illegal Art Trade registry, managed by the Ministry of Internal Affairs in Sarajevo). Not only are these crucial data sources for the construction of a consolidated dataset of

[3] In the case of Yemen, the General Authority of Antiquities and Museums in Sana'a issues periodic lists of antiquities that have been looted from the country and consigned at international auctions (Yemen News Agency 2024). However, these lists are in the form of documentary reports rather than web repositories, so while they are envisioned as a future data source for RITHMS platform, they fell outside the scope of the present task to integrate.

looted objects, but they also represent resources of nations widely regarded as source countries for cultural goods trafficking (Brodie et al. 2019; Hardy 2021). Consequently, the integration of these databases is of immediate relevance to LEAs in those countries.

A web scraper was also built for the Collection Netherlands Register of Protected Cultural Property. Although this is not a database of stolen goods, it comprises a set of cultural objects whose unauthorised presence in an auction or sale likely indicates illicit activity. Similarly, German police have established the Securius database of cultural goods confiscated by police across the country which have also been linked to illicit activities. This database includes the locations of the seizures and the locations of each police department, provided opportunities for geospatial intelligence provisioning and mining.

Last are databases of WWII-looted goods. These hold records of items in museums that lack WWII-era provenance (the Dutch Museums Association's Museum Acquisitions from 1933, Collection Netherlands' Cultural Goods of the Second World War, and French Ministry of Culture's Rose Valland database), artworks that are known to have been looted and are still missing (Lost Art Database, Degenerate Art Database, Hermann Goering and Linz Collections, Reichskunstdepot Kremsmünster list ('K-List'), databases of the Max Stern Art Restitution Project, Tableau et Dessin database, Division of Looted Art Catalogue of Wartime Losses, Einsatzstab Reichsleiter Rosenberg (ERR) Collection, and Central Registry of Looted Cultural Property), and provenance records for cultural goods that circulated in Europe during and after the war (Proveana). The consolidation of these datasets represents one of the most significant prospects for knowledge discovery as records on the same artefact from multiple databases can be aggregated, with fragmentary information from each contributing to a more accurate reconstruction of provenance than any single source is able to provide. Further, variations on the available properties of the data objects from each source provide additional aliases (for artist or object title, for example) and object details (namely materials or dimensions) to broaden the scope of investigations into the recoveries of missing cultural goods.

While this research boasts numerous advancements and strengths, it also comes with certain limitations. First, as with all OSINT, research is dependent on the extent and nature of publicly available information. As mentioned above, Middle Eastern and North African cultural goods are particularly vulnerable to trafficking into Europe and North America because they often originate in contexts that lack stringent mechanisms for the enforcement of cultural property protection. Further, these objects are frequently looted to meet increased demand in western art markets. At the same time, there are very few databases on looted or missing goods from these countries, due partly to the fact that significant resources are needed to develop these databases, but also because many of the looted items are taken directly from archaeological sites without being formally reported. Even within Europe, the geographic focus of the project, it should be noted that the dataset prepared for this study skews towards countries that possess the resources and inclination to prepare such databases.

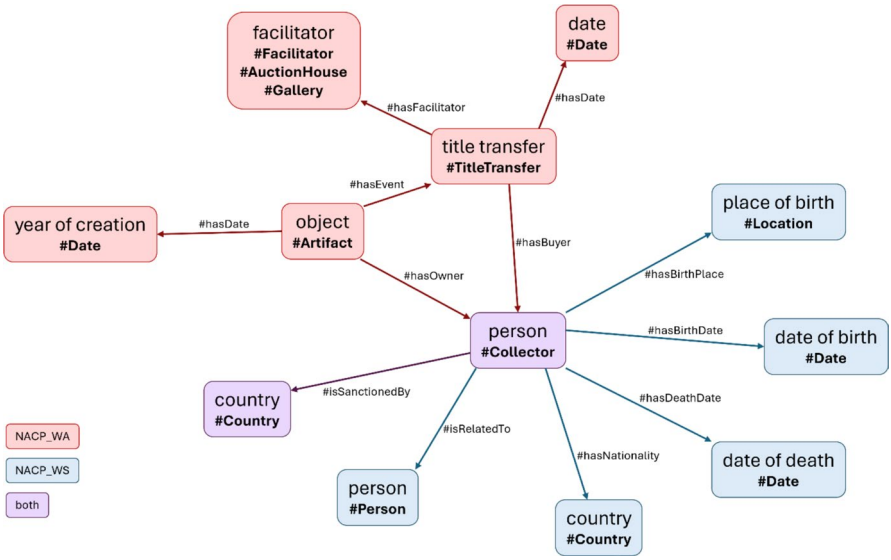

Fig. 1 Overlaps in scope between the NACP's War & Art ('NACP_WA') and War & Sanctions ('NACP_WS') databases

Results

The final dataset derived from scraped data consists of 518,353 cultural objects, including paintings, drawings, sculpture, textiles, books, furniture, coins, and other items. However, the dataset is valuable for more than just its coverage of artworks; as outlined above, entities associated with artworks (including artists, dates, and locations) were also identified, extracted through deterministic means, and integrated in the dataset as distinct data objects. This resulted in over 2 million total entities related to the circulation of cultural goods.

Following the initial period of data collection and processing, we opted to integrate additional existing open resources (already made available through an application programming interface (API) or as an open dataset on GitHub) to enrich the consolidated dataset further. The choice of which datasets to include was based primarily on which had the most significant overlaps in temporal and geographic scope with the scraped data. As a result, data were consolidated with existing open records from European museum collections, historic sales of artworks, and the largest art museum in North America. These consisted of:

- The Mona Lisa database, managed by the French Ministry of Culture, holding 673,137 records of objects in French museum collections[4]

[4] Available at https://data.culture.gouv.fr/explore/dataset/base-joconde-extrait/information/?disjunctive.region&disjunctive.departement&disjunctive.manquant.

Table 3 Categories of data and sample values present in the consolidated dataset

Data category	Example values
Artwork / artefact name	Attic Black-Figure Kassel Cup
Object description	Depicts two swans in flight, Greek origin, sixth century BCE
Artist / producer / culture of origin	Greek (Attic, sixth century BCE)
Medium / materials	Oil on canvas
Dimensions	Height: 8 cm, Diameter: 18 cm
Date / period	Archaic Period (sixth century BCE)
Condition	Good condition
Provenance (ownership history)	Previously part of a private European collection, acquired in the nineteenth century
Current location	Example Museum, New York, United States of America
Inventory /accession number	G12345-34-1975a
Exhibition history	Displayed at XYZ Exhibition in 2022
References	Cited in "Ancient Greek Pottery Studies, 2021"
Price / value	$12,000
Related objects	Similar black-figure pottery from Attica
Image URL	High-resolution artefact photographs
Details of theft	Stolen in the evening of 30 March 2005 from Example Museum, NYC
Status	Recovered
Record created / modified	3 April 2005
Institution contact	Department of Numismatics, Example Museum, info@museum.eu
Person contact	Dr Firstname Lastname, Director, firstname.lastname@museum.eu

- The Getty Provenance Index, produced by the Getty Research Institute, holding 1,843,234 records of art world transactions derived from historic European sales catalogues[5]
- The Open-Access dataset of the Metropolitan Museum of Art, holding 484,957 records of objects in the museum's collection[6]

The final dataset produced from these open resources aggregated with the scraped data comprises over 3.5 million artworks. However, it should be noted that the final dataset does not comprise 3.5 million *unique* artworks, but rather 3.5 million records that now must be mined and examined to identify and consolidate data on specific objects. Overlaps between sources are of particular interest, because the goal was not to create simply a large dataset, but to leverage intersections among the data and produce something more informative than the sum of its parts.

Figure 1 shows a conceptual visualisation of the entities identified from two scraped data sources, both formerly managed by Ukraine's NACP and described in Sect. "Databases of cultural goods" above. Data scraped from the War & Art

[5] Available at https://www.getty.edu/research/tools/provenance/search.html.

[6] Available at https://github.com/metmuseum/openaccess.

database are visualised as red nodes and edges; data scraped from the War & Sanctions database are visualised as blue nodes and edges; data provided by both databases are coloured purple to indicate their shared provenance.

This represents one of the most valuable (though labour-intensive) applications of this data: to enrich our understanding of objects which appear multiple times within the dataset, leveraging diverse information provided by multiple sources. In the case of the NACP data, correlating entities between the two datasets was relatively straightforward because that person's web page in the War & Art database linked directly to their record in War & Sanctions (Leeson et al. 2024). Further, a unique ID number was assigned to each individual by the NACP, greatly reducing the risk of accidentally linking two individuals with a shared name. Correlation across other datasets was somewhat more complicated, however, and needed to leverage other pieces of data.

Considering the fragmented nature of the individual data sources, it is valuable to conclude by briefly reviewing the scope of the resulting consolidated dataset in terms of the categories of data that are present across the data sources. Table 3 provides a summary of the key categories of data, which provide information on the artworks/artefacts as well as any related dates (e.g., date of

Fig. 2 Blumendekoration mit Quellennymphe (Flower Decoration with Spring Nymph), oil painting, François Boucher, 1756. Wikimedia Commons

Table 4 Information on Boucher's "Blumendekoration mit Quellennymphe" in the databases under examination

	Linz Collection	ERR collection	K-list
Image	Yes	Yes	
Inventory numbers	Munich and Linz	ERR, Rothschild, and Linz	K-List and Linz
Artist name	Boucher, Francois	François Boucher	Boucher F
Artist lifespan	1703–1770	1703–1770	
Artwork title	Blumendekoration mit Quellennymphe [Flower Decoration with Spring Nymph]	Brunnendekoration mit Quellnymphe [Fountain Decoration with Spring Nymph]	Brunnendek. m. Quellnymphen [Fountain Dec. w. Spring Nymphs]
Date		1756	
Signature		signed and dated on the right edge of the fountain, "F. Boucher 1756"	signed
Material / technique	Canvas	Oil on canvas	Lwd. [Canvas]
Dimensions	267×218	H. 214.5 cm, W. 162.5 cm	267/220
Conservation status			Blasenkrank (drgd.)
Previous owner	Rothschild Collection	Maurice de Rothschild	Rothschild
Previous location	Paris R 252	Paris, France	Paris 252
Provenance	Confiscation: Reichsleiter Rosenberg operational staff / Paris (confiscation France) Consignment: Reichsleiter Rosenberg operational staff (NS office)	This item was set aside for Goering at the Jeu de Paume to be handed over to Hitler. Maurice de Rothschild recovered in 1951 a painting by Boucher entitled 'Venus recevant la pomme d'or.' The theme of this painting-R 252-is similar to the title of the painting that Mr. de Rothschild [lost]. Hence, we will presume that it is one and the same painting	Transfer date: 1 March 1944 Transfer destination: Thürnthal
Restitution	France (via Thürntal)	repatriated to France 22 May 1951	

creation or date of theft), people (e.g., artist or point-of-contact for the institution from where the object was stolen), organisations, and locations. While not all categories of data are available in every source (depending on the scope of the source, as indicated in Table 2), these pieces of information are all present in the consolidated dataset. The next section provides a specific example of an artwork identified in multiple sources using the consolidated dataset, demonstrating the value of this OSINT-derived resource.

Discussion

The resulting dataset has applications for cultural heritage crime research as well as provenance research. For both, correlation of data across the dozens of diverse sources is a crucial step for identifying and aggregating overlaps in coverage. As the simple example of the NACP shows, basic correlation was accomplished through the linking of ID numbers across datasets. The following paragraphs will discuss a specific object that was initially correlated using other, though still relatively straightforward, techniques for data mining of the consolidated dataset.

In the case of WWII-looted art, the consolidation of data from over a dozen databases presents one opportunity for overcoming the challenge of incomplete, missing, and in some cases even fabricated post-war provenance records. Consequently, analysis of these scraped datasets 'can fill gaps on the routes through which looted objects travelled, complementing investigations and demonstrating the value of integrating traditional provenance research with novel technologies' (Giovanelli et al. 2025). As introduced above, there is significant overlap among databases of WWII-looted goods, which often hold fragmented data on the same artefacts, representing a promising opportunity for research and analysis. Combining object records pertaining to the same item enables reconstruction of the object's ownership history through triangulation across sources as well as the identification of influential related entities for further investigation.

For example, consider an object that is detailed in multiple databases, identified and linked using data mining techniques applied to the consolidated dataset: 'Blumendekoration mit Quellennymphe' (Flower Decoration with Spring Nymph), a 1756 painting by the French artist François Boucher (see Fig. 2). A record can be found for this artwork in three databases: the Linz Collection of objects acquired for the unrealised 'Fuhrer-Museum' in Linz, Austria, the ERR Collection of artworks seized by a Nazi task force under the authority of ideologue Alfred Rosenberg, and the K-List archive of Nazi-looted goods that were stored at the Kremsmünster Benedictine monastery in Austria (German Historical Museum n.d.; ERR Project n.d.; Leonhard Weidinger n.d.). Correlation among these datasets was achieved through the leveraging of common inventory numbers used in multiple databases: in this case, the numbering system used by Nazi officials for artworks intended for the proposed museum in Linz (so-called 'Linz numbers'). As illustrated in Table 4, different data are provided about the object in each data source, meaning the information gathered by researchers and law enforcement agents would be fragmented if they only consulted a single database.

First, the procedures of data cleansing and pre-processing (as outlined in Sect. 'Data post-processing') were followed, which normalised certain fields to enable linking of recurring entities. These were performed using regex statements on the data during the transformation phase, while structured as a Pandas DataFrame. For example, this allowed the correlation of 'François Boucher' and 'Boucher, Francois' to determine that these labels refer to the same person. Further, the values provided in the 'dimensions' property for each object were extracted and normalised, with the larger of the two assigned the 'height' and the smaller classified as the 'width'. In the two cases where the unit of measurement was not provided, manual inspection of the data source was used to determine the conventions used, which in the present case found that all three used the metric system of measurement with values provided in centimetres. Extraction and normalisation of these values facilitated comparison of the resulting properties, which highlighted significant variations in the dimensions provided by the three sources for the painting:

{'height': '267 cm', 'width': '218 cm'}
{'height': '214.5 cm', 'width': '162.5 cm'}
{'height': '267 cm', 'width': '220 cm'}

First, it should be remarked that this paper does not claim or attempt to provide evidence of criminality. Rather, it seeks to highlight variations in data provided on the same object by multiple sources, demonstrating the value of a consolidated dataset in providing holistic object information rather than limited details from disparate data silos. In the case of the Boucher painting, the item recorded in the ERR Collection is approximately 50 cm less in height and 50 cm less in width than the object records of the Linz Collection and K-List. Further, the note in the provenance field that casually remarks 'we will presume that it is one and the same painting' despite the recovered painting having a different title ('Venus recevant la pomme d'or') and substantially different dimensions, suggests this case may actually concern two separate paintings that have been erroneously mistaken as the same piece. In other words, significant variations in the data suggest that the painting recovered by authorities in the post-war campaign for looted art may not be the same object originally reported stolen.

Variations among the dimensions reported for the original work were only observed while correlating data from multiple sources for inclusion within the consolidated dataset. During normal (i.e., manual) provenance or cultural heritage crime research, when investigators are obligated to consult a single source at a time and can only hope to locate all available information on a given object, the likelihood of this case being flagged is low. This is especially the case considering the work is no longer regarded as missing, as the ERR Collection reported 'we will presume it is one and the same painting' and deemed the matter closed, despite troubling variations in the descriptive data provided on the object.

This example involves just one of hundreds of Nazi-looted artworks that were linked across multiple sources using the consolidated dataset, representing a major opportunity for wartime provenance researchers to examine, at a glance, conflicting details provided by multiple sources and determine, based on their subject matter expertise, promising works for further analysis.

Limitations

While this research represents a significant step forward in consolidating data on stolen, missing, and unprovenanced cultural goods, it is important to acknowledge its limitations, which have been mentioned throughout this article and will be summarised as closing remarks here. As with all research that relies on open data, a fundamental constraint lies in the variability of data availability across different regions. In other words, the comprehensiveness of the resulting dataset is ultimately determined by the extent and accessibility of public records, which tend to disproportionately represent certain countries and regions. Additionally, the dataset reflects disparities in documentation efforts across different jurisdictions. Even within Europe, the primary geographic focus of this study, the availability and quality of data vary widely. Countries with strong mechanisms for cultural heritage protection and extensive national resources are more likely to maintain detailed and regularly updated databases, leading to a disproportionate representation of certain regions in the dataset. Future work should aim to mitigate this by advocating for more standardised reporting practices and promoting documentation efforts in less-represented regions. Further, while the final RITHMS platform is intended solely for use by LEAs, we anticipate publication of a part of the consolidated dataset for use by the public. In the interest of compliance with existing legislation and ethical requirements, we will only include data for which we have been given express permission to reproduce; all other information will only be retrievable through the original source (which is already publicly accessible and can still be manually consulted by researchers and the public). Lastly, there is a risk that databases' source HTML may change in future, rendering the scrapers obsolete or non-functional. While this may present an obstacle for future data collection, it is one that can be overcome through regular maintenance of the developed tools and frequent manual verification that the data collection and post-processing pipeline is continuing as intended.

Conclusions

While the value of a consolidated dataset of missing, looted, and unprovenanced cultural goods has been emphasised in scholarship in the disciplines of heritage management and cultural policy, none has yet been developed. The creation of the dataset by the present study therefore meets the real-world demand for a single unified solution to address the current inefficient reality of research into stolen and unprovenanced cultural goods. While not originally the purpose of the data collection tools developed within the scope of RITHMS project, the consolidated dataset produced by this development task offers significant opportunities for research into object provenance and investigations into cultural heritage crime. Further, the mechanisms of data collection and analysis outlined in this study demonstrate that known techniques (specifically web scraping, deterministic data cleaning, and data mining) have pragmatic applications that have so far been underutilised. The anticipated publication of the dataset created by this project offers additional opportunities for

researchers and investigators to conduct data analysis and contribute to the knowledge discovery made possible by these resources.

The multi-step process for data extraction, transformation, and analysis that has been detailed in this study has produced the largest known non-police dataset of looted, missing, protected, and unprovenanced cultural goods. The creation of this dataset has resulted in something that is truly greater than the sum of its parts, as fragmented data from multiple sources are consolidated to reveal new insights on the illicit trade of cultural goods. Initial mining of this dataset has resulted in promising leads for the investigation of cultural goods crimes, particularly the recovery of looted cultural objects in Europe. This research has the potential to significantly impact the process with which researchers investigate unprovenanced and underprovenanced artworks, reinforcing the value of a consolidated dataset built from OSINT sources.

Acknowledgements We are indebted to the efforts of the foundations, research institutes, LEAs, NGOs, and government ministries that have prepared the databases on which this research relies. We are also grateful to Patricia Faraldo Cabana and her team (UDC) and Pavel Varbanov, Georgi Koykov, and George Sharkov (ESI) for the work they have done on their respective tasks, which has enabled the present study, and to the other partners of the RITHMS Consortium for their support and participation. Finally, we would like to thank the IIT Legal Affairs Directorate for their thorough support in navigating the GDPR and copyright issues relevant for this task.

Author contributions Conceptualisation was performed by M.L., R.G., S.F., M.D., and A.T.; methodology was conducted by M.L., R.G., S.F., M.D., and A.T.; software and data collection was developed by M.L.; formal analysis was carried out by M.L. R.G., and S.F.; writing—original draft preparation was prepared by M.L.; writing—review and editing were done by R.G., S.F., M.D., and A.T.; visualisation was provided by M.L.; supervision was provided by M.D. and A.T.; funding acquisition was secured by A.T. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Istituto Italiano di Tecnologia within the CRUI-CARE Agreement. This research was conducted under the auspices of RITHMS project (G.A. 101073932), funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission; neither the European Union nor the granting authority can be held responsible for them. This research was carried out at the Center for Cultural Heritage Technology of the Italian Institute of Technology.

Declarations

Competing interests The authors declare that they have no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

American Alliance of Museums (n.d.) The Nazi-Era Provenance Internet Portal Project. Nazi-Era Provenance Internet Portal. http://www.nepip.org/. Accessed 11 Dec 2024

Anderson B (2024) Kindred contexts: archives, archaeology, and the concept of provenance. Arch Sci 24:761–781

Association of Art Museum Directors (n.d.) Captive figure – object information. New acquisitions of archaeological material and works of ancient art. https://aamd.org/object-registry/new-acquisitions-of-archaeological-material-and-works-of-ancient-art/424. Accessed 12 Apr 2024.

Bach J (2024) Provenance-centered reckoning. Berl Bl 89:43–60

Breske A (2018) Politics of repatriation: formalizing indigenous repatriation policy. Int J Cult Prop 25:347–373

Brodie N (2011b) The market in Iraqi antiquities 1980–2009 and academic involvement in the marketing process. In: Manacorda S, Chappell D (eds) Crime in the art and antiquities world: illegal trafficking in cultural property. Springer, New York, pp 117–133

Brodie N (2018) Problematizing the encyclopedic museum: the benin bronzes and ivories in historical context. In: Effros B, Lai G (eds) Unmasking ideologies: the vocabulary and symbols of colonial archaeology. Cotsen Institute, Los Angeles, pp 61–82

Brodie N, Manivet P (2017) Cylinder seal sales at Sotheby's and Christie's (1985–2013). J Art Crime 17:3–16

Brodie N, Yates D, Slot B, Batura O, van Wanrooij N, Hoog G (2019) Illicit trade in cultural goods in Europe—characteristics, criminal justice responses and an analysis of the applicability of technologies in the combat against the trade: final report. Publications Office of the European Union. https://publications.europa.eu/en/publication-detail/-/publication/d79a105a-a6aa-11e9-9d01-01aa75ed71a1/language-en/format-PDF/source-search. Accessed 14 June 2024.

Brodie N (2011a) Scholarship and insurgency? The study and trade of iraqi antiquities. In: Illicit traffic of cultural objects: law, ethics, and the realities. An Institute of Advanced Studies Workshop, 4–5 August 2011, Perth, Australia. University of Western Australia, Perth, pp 1–28

Carabinieri Command for the Protection of Cultural Heritage (2021) Press release. https://www.carabinieri.it/docs/default-source/cittadino_doc/informazioni/tutela/press-release_english.pdf. Accessed 16 Dec 2024.

Colavizza G, Blanke T, Jeurgens C, Noordegraaf J (2021) Archives and AI: an overview of current debates and future perspectives. J Comput Cult Herit 15:1–15

Colla D, Goy A, Leontino M, Magro D, Picardi C (2022) Bringing semantics into historical archives with computer-aided rich metadata generation. J Comput Cult Herit 15:1–24

Carabinieri Command for the Protection of Cultural Heritage (n.d.) SWOADS. https://tpcweb.carabinieri.it/SitoPubblico/home/informazioni/swoads. Accessed 16 Dec 2024.

Duff WM, Flinn A, Suurtamm KE, Wallace DA (2013) Social justice impact of archives: a preliminary investigation. Arch Sci 13:317–348

ERR Project (n.d.) Card ID 16018. Cultural plunder by the einsatzstab reichsleiter rosenberg: database of art objects at the Jeu de Paume. https://www.errproject.org/jeudepaume/card_view.php?CardId=16018. Accessed 12 Apr 2024.

Foley T (2014) Art loss and databases: the quest for a free single unified system (Order No. 10185640). Publ. Avail. Content database. (2266447058). https://www.proquest.com/dissertations-theses/art-loss-databases-quest-free-single-unified/docview/2266447058/se-2. Accessed 18 Apr 2025.

Gerstenblith P (2019) Provenances: real, fake, and questionable. Int J Cult Prop 26:285–304

Giovanelli R, Leeson M, De Bernardin M, Ferro S, Traviglia A (2025) Social network analysis on the proveana database: insights on the circulation of Nazi-looted cultural goods during and after WWII. Soc. Netw. Anal. Min., accepted pending revisions

Graham S, Yates D, El-Roby A, Brousseau C, Ellens J, McDermott C (2023) Relationship prediction in a knowledge graph embedding model of the illicit antiquities trade. Adv Archaeol Pract 11:126–138. https://doi.org/10.1017/aap.2023.1

Green J (2017) Museums as intermediaries in repatriation. J East Mediterr Archaeol Herit Stud 5:6–18

German Historical Museum (n.d.) Data sheet LI001590. Linz collection. https://www.dhm.de/datenbank/linzdbv2/queryresult.php?obj_no=LI001590. Accessed 12 Apr 2024

Hardy SA (2021) It is not against the law, if no one can see you: online social organisation of artefact-hunting in former Yugoslavia. J Comput Appl Archaeol 4:169–187

Hashemi L, Waddell A (2022) Investigating the online trade of illicit antiquities. In: Hashemi L, Shelley L (eds) Antiquities smuggling in the real and virtual world. Routledge, New York, pp 218–239
La Follette L (2017) Looted antiquities, art museums and restitution in the United States since 1970. J Contemp Hist 52:1–19
Leeson M, Giovanelli R, De Bernardin M, Traviglia A (2024) War, art, and sanctions: social network analysis on the NACP's databases of sanctioned russian individuals and art collectors. Int J Digit Humanit. https://doi.org/10.1007/s42803-024-00089-y
Levahi A (2023) From global databases to global norms? The case of cultural property law. Univ Pa J Int Law 44:359–416
Lowenthal D (1998) The heritage crusade and the spoils of history. Cambridge University Press, London
Leonhard Weidinger (n.d.) Reichskunstdepot Kremsmünster-Painting: "K-List." Leonhard Weidinger Wien. https://leonhard.weidinger.wien/daten/kremsmuenster-gemaelde. Accessed 12 Apr 2024.
Moss M, Thomas D, Gollins T (2018) The reconfiguration of the archive as data to be mined. Archivaria 86:118–151
O'Neill B, Stapleton L (2022) Digital cultural heritage standards: from silo to semantic web. AI Soc 37:891–903
Rapicavoli S (n.d.) Innovation and technology with artificial intelligence to explore the web in search of stolen works of art—the S.W.O.A.D.S. Project. https://rm.coe.int/international-conference-the-nicosia-convention-a-criminal-justice-res/1680abb733. Accessed 16 Dec 2024
Romein CA, Kemman M, Birkholz JM, Baker J, De Gruijter M, Meroño-Peñuela A, Ries T, Ros R, Scagliola S (2020) State of the field: digital history. History 105:291–312
Ruiz Romero Z (2020) Lawfulness and ethics around cultural property auctions: the case of the barbier-mueller pre-columbian collection. Int J Cult Prop 27:397–416
Savoy B (2015) Plunder, restitution, emotion and the weight of archives: a historical approach. In: Rotermund-Reynard I (ed) Echoes of exile: Moscow archives and the arts in Paris 1933–1945. De Gruyter, Berlin, pp 27–44
Sweeney S (2008) The ambiguous origins of the archival principle of "provenance." Libr Cult Rec 43:193–213
Yemen News Agency (2024) Antiquities authority issues list of 50 antiquities that smuggled abroad. Saba Net. https://www.saba.ye/en/news3314279.htm. Accessed 12 Apr 2024

Dr Madison Leeson is a historian of the modern Middle East, focusing on the ways in which cultural heritage is leveraged for political legitimacy. She is currently a Postdoctoral Researcher at the Center for Cultural Heritage Technology (CCHT) of the Italian Institute of Technology, based in Venice, Italy. She holds a PhD in Archaeology and History of Art from Koç University (2022), specialising in Cultural Heritage Management and focusing on cultural governance in modern Iraq. Her research into cultural heritage crime emphasises the historical conditions that have enabled the trafficking of cultural goods. Other research interests include developments in museum studies, particularly concerning restitution and repatriation, and the role of objects in the manifestation of power. Her current position with CCHT combines many research interests—heritage management, antiquities protection, archival research, and innovative technologies—in support of the RITHMS project to tackle the illicit antiquities trade.

Dr Riccardo Giovanelli, an archaeologist specializing in Egyptology and Roman Provinces, directed his research towards investigatingthe looting and illicit trafficking of cultural artefacts in 2011, prompted by the widespread looting of museumsand archaeological sites in Egypt during the Arab Spring. He is now a Postdoctoral Researcher at the Center forCultural Heritage Technology (CCHT) in Venice. He researches practical applications of advanced digital methodsand Knowledge Graph frameworks towards the challenges associated with preserving cultural heritage andcurbing the trade of looted artefacts. He is the President of the civil society organisation "Art Crime Project", andactively engages in initiatives aimed at raising awareness, involving the broader public and fostering collectiveaction to prevent and counter art-related crimes across various dimensions.

Dr Sara Ferro, a Control and Automation Engineer specialising in Machine Learning and Deep Learning models applied to sequential data, has focused her research on developing innovative models for the automatic digitalisation of historical documents written in Western Latin-based languages. She earned her PhD in Computer Science from Ca' Foscari University of Venice, where her doctoral programme was conducted in collaboration with the Center for Cultural Heritage Technology (CCHT) of the Italian Institute of Technology, based in Venice, Italy. Her thesis introduced new procedures, methods, and models to accurately transcribe historical texts by leveraging contextual information within the data, an important concept in Computer Vision and Natural Language Processing research. She is a Postdoctoral Researcher at CCHT, working under the EU Project RITHMS to utilise and define new Natural Language Processing models for extracting information from online sources to help create a Knowledge Graph database for the RITHMS platform.

Dr Michela De Bernardin is an ancient historian and classical archaeologist with honours from the University of Pisa and the Scuola Normale Superiore, where she also obtained her PhD in Ancient History and Archaeology within a joint programme with the Ruprecht-Karls-Universität Heidelberg. After a specialised postgraduate course at the Centro Studi Criminologici in Viterbo, she completed a two-year postgraduate programme at the University of Roma Tre addressing the illicit trade of Palmyrene funerary portraits and the impact of numismatic forgery on the art market. She co-founded and is now the secretary of the *Art Crime Project APS* organisation. Since 2020, she has been collaborating with the Center for Cultural Heritage Technology (CCHT) of the Italian Institute of Technology (IIT). Her research stretches from provenance, looting, and forgery studies to the dynamics of illicit trafficking. Currently, she is a Postdoctoral Research Fellow at the CCHT and the Scientific Project Manager of the HE RITHMS project.

Dr Arianna Traviglia serves as the Coordinator of the Center for Cultural Heritage Technology (CCHT) within the Italian Institute of Technology (IIT). Her research concentrates on integrating technology-based practices into the examination, protection, and conservation of cultural heritage. Leveraging her expertise in both archaeology and multi-spectral and hyperspectral imaging, she oversees research at the Center in Artificial Intelligence, Nanotechnologies, and Robotics applied to antiquities and ancient landscape studies. With the CCHT team, she leads or contributes to various projects aimed at the preservation and protection of cultural heritage. Notably, she currently coordinates the RITHMS project, focusing on advanced technologies based on social network analysis and knowledge graphs to combat cultural property trafficking. This initiative involves collaboration with 7 European police forces and partners from academia and industry. She also spearheads the European Space Agency (ESA)-funded ALCEO project, utilising satellite imagery and remote sensing to identify looting evidence in archaeological areas.

Authors and Affiliations

Madison Leeson[1] · Riccardo Giovanelli[1] · Sara Ferro[1] · Michela De Bernardin[1] · Arianna Traviglia[1]

✉ Arianna Traviglia
arianna.traviglia@iit.it

Madison Leeson
madison.leeson@iit.it

Riccardo Giovanelli
riccardo.giovanelli@iit.it

Sara Ferro
sara.ferro@iit.it

Michela De Bernardin
michela.debernardin@iit.it

[1] IIT–Center for Cultural Heritage Technology, Via Torino 155–Epsilon Building, Venice, Italy