The RESTORE project: a final review

Emiliano Degl'Innocenti^{1,*}, Leonardo Canova¹, Francesco Coradeschi¹, Carmen Di Meo¹, Maurizio Sanesi¹, Alessia Spadi¹ and Federica Spinelli¹

¹ Consiglio Nazionale delle Ricerche, Istituto Opera del Vocabolario Italiano, Firenze, Italy

Abstract

This paper provides a complete assessment of the research activities and the results achieved by the RESTORE project (smaRt accESs TO digital heRitage and mEmory), started in June 2020 and recently ended. The project consortium, coordinated by the Istituto Opera del Vocabolario Italiano - National Research Council of Italy (OVI-CNR), included the State Archives and the Museum of Palazzo Pretorio in Prato, the Archival and Bibliographic Superintendency of Tuscany as well as the SPACE SpA software company. The project - cofinanced by the Region of Tuscany - had as its main purpose the recovery, integration and accessibility of data and digital objects produced in the last twenty years by its partners, in order to build a knowledge base gathering information on the history of the city and of its institutions, on the development of its economic and entrepreneurial system, and on the role of women in the development of a welfare state and network, starting from the figure of the merchant Francesco di Marco Datini, his family and his entourage. Different aspects of the RESTORE project, including data mapping and modelling, have already been presented in previous IRCDL editions; this paper will focus on the latest updates related to the RESTORE platform final release, focusing on the services offered to support research in heritage science and humanities.

Keywords

Semantic Web Technologies, Linked Data for DLs, Standards and Interoperability, Data and Information Lifecycle, Research Infrastructures, FAIR, ontologies, mapping, modelling, social sciences and humanities, GLAM, Knowledge Base Construction for cultural heritage.

1. Introduction

RESTORE (smaRt accESs TO digital heRitage and mEmory²) is a project coordinated by the Istituto Opera del Vocabolario Italiano of the Italian Consiglio Nazionale delle Ricerche (CNR-OVI) located in Florence. The project's aim was to support the participating institutions in the process of applying the FAIR principles to their digital resources providing access to a rich and interconnected data space with resources produced by leading institutions in decades of work, fostering the interpretation, contextualization and comprehension of cultural heritage assets by representing their semantic depth in order to reconstruct the complex network of interrelations with other entities (individuals, objects, places, concepts) and restore the connections existing between tangible and intangible aspects of Cultural Heritage Objects. The data integration platform design and implementation has been guided by the *Findability*, *Accessibility*, *Interoperability* and *Reuse* (FAIR³) approach [1]: a set of principles for the integration of resources and information encoded with different standards and with different

EMAIL: emiliano.deglinnocenti@cnr.it (E. Degl'Innocenti); canova@ovi.cnr.it (L. Canova); fr.coradeschi@gmail.com (F. Coradeschi); dimeo@ovi.cnr.it (C. Di Meo); maurizio.sanesi.eng@gmail.com (M. Sanesi); spadi@ovi.cnr.it (A. Spadi); spinelli@ovi.cnr.it (F. Spinelli) ORCID: 0000-0002-3839-9024 (E. Degl'Innocenti); 0000-0002-4087-6504 (L. Canova); 0000-0003-0808-2736 (F. Coradeschi); 0000-0001-9953-7753 (C. Di Meo); 0000-0003-3917-5935 (M. Sanesi); 0000-0002-9670-3426 (A. Spadi); 0000-0002-2195-3930 (F. Spinelli)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings (CEUR-WS.org)

^{*} Corresponding author

¹⁹th IRCDL (The Conference on Information and Research science Connecting to Digital and Library science), February 23-24, 2023, Bari, Italy

² The website is consultable at: http://restore.ovi.cnr.it

³ Findability, Accessibility, Interoperability and Reuse - FAIR: https://www.go-fair.org/fair-principles/

workflows. In doing this, the RESTORE development team adopted a modular architecture, developing customized software components when needed and integrating existing solutions when available, to build the final RESTORE platform. Following the same approach, RESTORE also developed a set of data management and integration workflows based on the requirements expressed by the project's partners, to support the creation of a LOD with information (i.e. textual corpora, documents descriptions and related images) describing tangible and intangible aspects of cultural heritage objects (documents, books, works of art) coming from different contexts (i.e. domains). This paper presents the results achieved in the two years work of the RESTORE project with the aim to share the solutions implemented and the studies that drove the project team choices. In the initial phase of the project, the resources provided by the partners were analyzed and the requirements related to their integration were considered, in order to create the knowledge base, taking into account the research and production contexts (e.g. libraries, archives, museums, etc.). It was therefore necessary to develop tools for mapping the standards used in each domain of reference (TEI⁴ for texts [5]; EDM⁵, MAG⁶, MODS and METS⁷ for resources produced by libraries; EAD⁸ and EAC⁹ for archives [2][3]; ICCD¹⁰ as a cataloguing system for works of art [4]; other formats and standards used for different disciplines in the field of cultural heritage science such as EDF¹¹, HDF5¹² etc.). Afterwards, the process of converting the resources provided by the partners towards the reference ontology of the RESTORE project (CIDOC -Conceptual Reference Model¹³ [6]) has been outlined and undertaken. To achieve this, the team developed a workflow that included several phases, from collecting and storing the data in the Comprehensive Knowledge Archive Network (CKAN)¹⁴ data store, to mapping and aligning the concepts expressed in the different datasets and eventually populating the RESTORE ontology with the partners' information and making it available for search, navigation, visualisation and reasoning services through the Virtuoso¹⁵ triplestore.

2. State of the art

2.1. Context

The State Archives of Prato holds a vast number of sources created by civil, religious and charity institutions that, integrated with those of other cultural and memory institutions of the city of Prato, including the Museum of Palazzo Pretorio, allows to reconstruct the history of the city's community, territory and society, from the 14th century to the present day. For the two oldest and most rich in historical information collections, namely the "Datini¹⁶" and "Ospedale Misericordia e Dolce¹⁷" funds, a computerised archival description, digital images of the documents¹⁸, transcription of the texts and a lemmatized text corpus produced by OVI already existed. In the past, digital research tools were also created (i.e.: the CD-ROM *Per la tua Margherita*¹⁹, the *Datini online* website²⁰, etc.), but as of today they are poorly interoperable, not easily accessible and - in some cases - almost completely unusable. To a large extent, these resources are the result of unrelated projects and operations, lacking the extra value given by the project coordination and the integration of data produced by various actors on the

¹⁹ Project website: http://datini.archiviodistato.prato.it/margherita/Root/Servizio/proj_fr.htm

⁴ Text Encoding Initiative - TEI: https://tei-c.org/

⁵ Europeana Data Model - EDM: https://pro.europeana.eu/page/edm-documentation

⁶ Administrative and Management Metadata - MAG: https://www.iccu.sbn.it/export/sites/iccu/documenti/manuale.html

⁷ Metadata Object Description Schema - MODS e METS: https://www.loc.gov/standards/mods/presentations/mets-mods-morgan-ala07

⁸ Encoded Archival Description - EAD: https://www.loc.gov/ead/

⁹ Encoded Archival Context - EAC: https://eac.staatsbibliothek-berlin.de

¹⁰ Istituto Centrale per il Catalogo e la Documentazione - ICCD: http://www.iccd.beniculturali.it/

¹¹ European Data Format - EDF: https://www.edfplus.info/

¹² Hierarchical Data Format - HDF5: https://www.hdfgroup.org/solutions/hdf5

¹³ CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org/

¹⁴ Comprehensive Knowledge Archive Network - CKAN: https://ckan.org/

¹⁵ Virtuoso Open Link: https://virtuoso.openlinksw.com

¹⁶ Archive of administrative documents and correspondence of the merchant Francesco di Marco Datini (1335-1410) attesting a cross-section of the economy and social life of the entire Mediterranean area, through his extensive activities in industry, trade and banking.
¹⁷ Charitable organisation that cared for wayfarers, the poor and abandoned children since the 13th century. Digital resources related to this

¹⁷ Charitable organisation that cared for wayfarers, the poor and abandoned children since the 13th century. Digital resources related to this fund can be accessed on the Archives website: http://www.archiviodistato.prato.it/accedi-e-con-sulta/aspoSt005/tree

¹⁸ The digitization project began in 1999 and can be accessed on the Archives website: http://datini.archiviodistato.prato.it/il-progetto/

²⁰ State Archives of Prato website: http://datini.archiviodistato.prato.it

same cultural heritage artefacts. RESTORE proceeded to structure these resources according to the FAIR approach and to integrate them with new data and digital objects from other institutions, including the OVI-CNR Institute and the Museum of Palazzo Pretorio. The latter institution, preserves a significant part of the city's artistic heritage, including works of art linked to the iconography of Francesco di Marco Datini and to the "Ospedale Misericordia e Dolce", the institution to which a consistent part of the civic collection of works of art belongs. Up to now, the published bibliography considered as works of art belonging to the Ospedale Misericordia and Dolce most of those held by the civic Museum from the second half of Nineteenth Century, when the works of art left the Ospedale to be exhibited in the "Galleria Comunale", opened in 1858, while instead only a part of them were effectively coming from the Ospedale Misericordia and Dolce. Recently, in order to precisely define the great nucleus of works of art by their provenance, the Museum staff has undertaken a research work consulting entries, repositories, inventories, entry/exit records, identifying three main groups: works of art commissioned by the Ospedale and belonging to it since the Fourteenth Century; works of art gifted to the Ospedale by donations; works of art only temporarily stored at the Ospedale and then held by the Comune of Prato for the civic collection. As stated by the Head Conservator and Curator of the Museum of Palazzo Pretorio²¹, thanks to the RESTORE project the result of such research work is now digitally published, and it is clearly possible to search for works of art certainly belonging to the Ospedale Misericordia e Dolce, and those belonging to the so-called "Galleria Martini". A thorough research work on this specific subject is still in fieri. In addition to the already identified digital resources, the Museum of Palazzo Pretorio has selected a number of sources - which were considered relevant to the purposes of the project - that are still to be digitised: the creation of an integrated knowledge base for these resources makes it possible to break down existing disciplinary, institutional and technological barriers and to make them accessible in innovative ways. A rich documentary section, otherwise poorly known but potentially able to shed light on themes (e.g., the commercial and welfare networks in the Fourteenth Century Prato, including abandoned childhood hospitals, the role of women in both these contexts, etc.) that are strongly rooted in the history of Prato and - more broadly - of Tuscany, is now publicised, valorized and made easier to navigate and understand. The information thus accessible closely touches all layers of citizenry and aims to engage the population in a cultural journey towards reappropriating spaces of knowledge and local histories that are significant for increasing and preserving collective memory. RESTORE analysed the critical issues identified by the cultural institutions involved, in the context of the collection, aggregation, semantic enrichment and availability of high value-added information - related to the mentioned cultural heritage objects - in order to build a knowledge base populated with scholarly reliable information, made accessible, interoperable and reusable. To this end, models and solutions have been developed for the management, processing and fruition of cultural and documentary heritage that can be reproduced in a sustainable manner (both in terms of technology and of resources required to maintain them) by other institutions. The knowledge base resulting from these activities is then interoperable and reusable in the context of other national and international initiatives such as the H2IOS \hat{C}^{22} and the SSHOC²³ projects, both connected to the EOSC²⁴.

2.2. Data: Archival sources, museum collections, textual corpora

In detail, the resources ingested and processed by the project were:

- the Datini Fund, world's largest mercantile archive for the Middle Ages, consisting of 150.000 letters and about 600 registers from which we can get information about the people involved, the costs and typologies of goods, the places designated for trade, etc.;
- the Ospedale Misericordia e Dolce fund, with its 7.000 archival units, presenting all the articulations of the functions of a welfare institution: from the support of the traveler, to the care of the poor and

²¹ Statements from Conservator and Curator of the partner institutions regarding the selection of the original resources shared for the project, are to be found here: http://restore.ovi.cnr.it/about.html [date of access: nov. 29, 2022].

²² Humanities and cultural Heritage Italian Open Science Cloud – H2IOSC: https://www.h2iosc.cnr.it/

²³ Social Sciences & Humanities Open Cloud – SSHOC: https://sshopencloud.eu/

²⁴ European Open Science Cloud - EOSC: https://eosc-portal.eu/

the sick, to the reception of the "gettatelli", abandoned children raised, thanks to the Hospital itself, by the entire community of Prato;

• a selection of works of art taken from the collection of the Museum of Palazzo Pretorio, catalogued according to the ICCD standard ("OA" template - Work of Art).

CNR-OVI institute also made available a textual corpus representing a selection of edited letters belonging to the extensive Datini correspondence, selectively lemmatized. The corpus records the graphic and morphological varieties of many terms which are relevant for the reconstruction of aspects related to the economic and legal life, as well as the social relations of the time. The lemmatization includes: anthroponyms, including any nicknames and specific positions (if the indication refers to an identified precise historical personage); toponyms, including names of cities, towns, districts, localities, streets, squares, gates, churches, monasteries, palaces, hospitals, organisations, institutions, etc. Lemmatized tokens also include terms pertaining to the religious and agricultural fields, body parts, time scan etc., and are also distributed within 22 conceptual categories (called *hyperlemmas*) including: clothing and furnishings, food, animals, arts and crafts, calendar, economics law and politics, construction and architecture, medicine, coins, navigation, kinship, leather and textiles, etc. The lemmatized corpus of the Datini correspondence, created by CNR-OVI between 2003 and 2005, is one of the many corpora²⁵ that the institute set up through the Gestione degli Archivi Testuali del Tesoro delle Origini (GATTO²⁶) software in a specially dedicated version that can be queried via Web²⁷. The corpus consists of: 2.511 texts; 45.259 types; 977.034 tokens of which 126,663 are lemmatized; 6.510 lemmas and 22 hyperlemmas. These corpora and the data ingested and provided by the project partner were compliant to the following format:

- EAD and EAC standard for the codification of the archival data and of the information connected to the "CFP" extension for subjects (corporate entities, families and bodies);
- ICCD (OA "Opera d'Arte", 3.00, 2018) standard for the codification of each entry related to works of visual art;
- custom TEI codification of the GATTO software, for the representation of the digital textual material of the OVI-CNR Institute.

This description is meant to emphasize the diversity and inconsistency of the original data before our standardization and processing efforts.

3. Data integration

The goal of the project - to achieve full integration of the datasets provided by the project partners - was a complex issue, as the data produced by different institutes refer to scientific and disciplinary domains that often follow widely differing schemes. This complexity was reflected in the use of standards and cataloguing systems diverging in structure and purposes. It was therefore necessary to set up a workflow to achieve semantization of data within a single reference domain and to develop a shared data management model that is valid for multiple metadata schemas and multiple standards. The project workflow describes the transformation of the data provided, regardless of the original format, into semantic triples (triplification)²⁸. This result was achieved through the mapping and modelling processes. Data triplification involves further data analysis, carried out in collaboration with domain experts, consisting of specific additional steps: i) choosing the most relevant information to be modeled; ii) specifying the implicit meaning of the information to be represented; iii) aligning information of the same kind, contained in different datasets, expressed through different semantic schemes and/or structures (mapping); iv) structuring data into triples according to the rules of the reference ontology (CIDOC-CRM), choosing the appropriate entities and properties (modelling).

²⁵ Among these are the "Tesoro della Lingua Italia delle Origini" (TLIO) and the "Opera del Vocabolario Italiano" corpuses, which constitute the reference database for the realisation of the Tesoro della Lingua Italiana delle Origini vocabulary: http://tlio.ovi.cnr.it/TLIO/. A complete list of the corpuses can be found here: http://www.ovi.cnr.it/Interroga-il-Corpus.html.

²⁶ Gestione degli Archivi Testuali del Tesoro delle Origini - GATTO: http://www.ovi.cnr.it/Il-Software.html

²⁷ Lemmatized corpus of Francesco Datini's (1335-1410) correspondence. It includes almost 150.000 letters and it is freely accessible through the GattoWeb platform: http://aspweb.ovi.cnr.it/(S(qmmiy5m0sybb4lao4qqmexyo))/CatForm01.aspx

²⁸ By "triple" it is meant here a transformation of data into a logical-conceptual construct, consisting of three properties (entities), i.e. a subject, an object, and a predicate (Subject, Object, Predicate). The reference model for this kind of semantisation operation is the RDF format.

3.1. Mapping and modelling

The data mapping phase involves the alignment of information expressed through different semantic schemes and/or structures through the combination of elements belonging to one dataset with those of another dataset if both express the same concept (eg: shelfmark, author, place, etc.). This workflow facilitates the migration of data to be converted into logical constructions based on the CIDOC-CRM model. The ontology provides the developers with a model of management and semantic representation of knowledge bases, allowing the definition of a set of relevant concepts and the description of the fundamental relationships between them. The goal of data modelling is to formalise the knowledge (i.e.: translate it into a formal language), to minimise its ambiguity (typical of natural language) and make it "computable", or machine readable, through a module called reasoner, so that the calculator can make inferences about it. Through the modelling, data becomes structured according to the rules of the reference ontology, which will form the basis for the transformation of the datasets into triples.

3.1.1. Mapping and Implementation details

The conversion of the original data into triples is done using custom parsers, implemented in the Python programming language. The first parser in the workflow takes as input the original data sources in XML format (e.g., according to the EAD, EAC, ICCD or TEI standards) and returns them in the form of tables encoded in CSV format, following a 1:1 ratio (a corresponding CSV file is produced for each original file). The next stage of data processing involves transforming the information contained in the CSV tables into triples, in TTL format. This stage is also accomplished using a parser implemented *ad hoc*. The triplification of data, i.e., the conversion of the data from the table format (CSV) into triples (TTL), involves further analysis of the data, which allows the information to be modelled according to the ontology and a set of reference vocabularies [7].

The conversion of the original ingested data transforms the reference standard to the CIDOC CRM model. The process has been performed for EAD and EAC, on the basis of the technical documentation of reference²⁹, and each XML EAD element conversion to the CIDOC CRM has been also shared by the Archive domain experts. First of all, the mapping involved the overall table-structure of the hierarchy of the elements and their attributes as it is expressed by the EAD standard for the entire archival resource, and each of those parts were converted into a CIDOC CRM Class system where Elements, Attributes and their relationships are expressed by a CIDOC Class, either a Property ("P"+n) or an Entity ("E"+n) or even a Subclass ("E" or "P"+n) of correspondence into the CIDOC CRM system. A full and complete description of the selected CIDOC Classes used for the RESTORE Mapping and Modeling (4) is available open access in a precedent devoted contribution, and also made visible in the integrated double view table under the "Mappatura" section published on the RESTORE platform at the following link: http://restore.ovi.cnr.it/standard.html, where it is possible to click on the two tables "EAD TO CIDOC", and "ICCD TO CIDOC" to access the overview of Mapping.

3.2. Vocabularies and thesauri

Besides the ontology, vocabularies and thesauri were also important elements of the data integration process. The common entities that occur in the different datasets will therefore have to share the link to the reference vocabulary. Through the references to vocabularies, it was possible to identify and index toponyms and anthroponyms, two of the entities identified as points of contact between the various datasets. In this way, the integration of data related to entities that refer to the same anthroponym or toponym is achieved. In fact, if entities belonging to different datasets refer to the same URI, they will be automatically linked also through the same vocabularies. Another point of contact was represented

²⁹ Stasinopoulou, T. et al. (2007). Ontology-Based Metadata Integration in the Cultural Heritage Domain. In: Goh, D.HL., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds) Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. ICADL 2007. Lecture Notes in Computer Science, vol 4822. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77094-7_25.

by people, for which the ULAN³⁰ and VIAF³¹ vocabularies were used. Then the Iconclass³² system was already a documented field of the work of art records. See Table 1 for details on vocabularies used in the RESTORE project.

Table 1

Vocabularies and thesauri

Vocabulary	Description
Art & Architecture Thesaurus (AAT)	Controlled vocabulary for describing art historical and demo-ethno- anthropological resources (artworks, architecture and material culture). It refers to broad, generalised concepts (e.g., painting, book, cathedral), but not to specific places, people, events.
Union List of Artist Names (ULAN)	Controlled vocabulary containing mainly biographical type references related to artists. It includes proper names, art names and pseudonyms, lexical, spelling and temporal variants, and translations of biographical data into multiple languages.
Getty Thesaurus of Geographic Names (TGN)	Controlled vocabulary containing names and information associated with places. It includes references to political-administrative entities and their physical features, current and historical places.
Virtual International Authority File (VIAF)	International Authority File constituting a database of controlled authority entries from different national catalogues.
Iconclass system	Iconographic classification system, structured by themes and motifs related to the Western artistic tradition designed for art and iconography. It contains "classes" of concepts organised hierarchically.

3.3. Browsing resources

Once the data modelling based on the CIDOC-CRM ontology is completed, the data are converted in triples (TTL format) and uploaded to the Virtuoso triplestore. Data access is performed through some features that have been implemented for RESTORE:

- SPARQL endpoint (http://dev.restore.ovi.cnr.it:8890/sparql/): a service that lets the users write SPARQL query to send their requests and obtain results as a response to the queries.
- Virtuoso Faceted Browser (http://dev.restore.ovi.cnr.it:8890/fct/): a facility service for browsing the resources in the Triplestore. It lets the users type keywords in a search bar to start navigation.
- LodLive instance of RESTORE (http://dev.restore.ovi.cnr.it/lodlive/): a visual resource browser enhancing RDF data quality by making it efficiently visualised through navigable graphs.

The data can be downloaded from the websites of the original providers at their discretion and according to their licenses³³. On the RESTORE side the data modeled according to the CIDOC-CRM ontology can be accessed and downloaded from the SPARQL endpoint in the preferred format such as TTL, CSV, etc. Moreover, the platform allows the user to generate a citation snippet or to acquire the hyperlink of the resource to disseminate the results of their queries. These three tools are very versatile and allow in-depth semantic searches; however, they are unlikely to be used by a vast part of our target audience. In fact, the selected target is very heterogeneous in terms of digital literacy, ranging from unexpert citizens interested in research activities to experts working in a memory, research or conservation organization. It is therefore necessary to have more or less specialised search options,

³⁰ Union List of Artist Names - ULAN: https://www.getty.edu/research/tools/vocabularies/ulan/

³¹ Virtual International Authority File - VIAF: https://viaf.org/

³² Iconclass: https://iconclass.org/

³³ See <u>http://restore.ovi.cnr.it/termini.html</u> for more details

depending on the targeted user's familiarity with surfing the net, and on the level of expertise in the topic object of the webquest.

4. Visualisation Tools

In this section, the latest version of the RESTORE visualisation platform will be presented, focusing on the features that support research in heritage science and humanities. Updates will also be provided compared to the works presented in previous IRCDL editions, where different aspects of the stilldeveloping RESTORE project were presented, including the mapping and modeling of data and the developed workflow. The data visualization solutions proposed in this contribution are the result of collaboration with domain experts for an effective presentation of semantic data. Providing a userfriendly interface, where data consultation and browsing does not require high level digital expertise or knowledge of the SPARQL language, becomes a crucial requisite for the project. The RESTORE platform thus offers a wide range of data visualisation solutions, bringing into focus the description and visualisation of the relationships made between the various entities mentioned in the collections, each intended to meet the needs of a certain target audience.

4.1. Custom visualisation records

RESTORE developed a user-friendly interface to let users browse the data regardless of their web literacy. The interface is, nonetheless, one of the means, though the more intuitive and visually appealing, by which the users interact with the data in the Virtuoso triplestore. The interface manages the users' queries and allows them to navigate the semantic relationships existing between different items and concepts, represented in the data CIDOC-CRM ontology. The aim of the platform is to have different ways of visualising and browsing the data, so that the user can find the resource of interest and access the information directly on the institution website that provided it. Infact, RESTORE doesn't want to build a mere visualisation tool focused on archival entries or artwork entries, given that those are already made available by the partners' interfaces. What the project intended to achieve is the integration of semantic data in an integrated environment that will activate new ways to reveal information through the relationships mapped during the stages of the work done. In this perspective, the purpose of the RESTORE interface was to make this browsing activity simple for all users, and to effectively link the resource in a multifaceted context. First of all, a search interface connected to the SPARQL endpoint of the project, managing the user queries and the interaction with the underlying triplestore was implemented. The results of the search consume the data loaded into the Virtuoso triplestore in the form of triples. As for the appearance of the search results, a classic view in tabular form of the occurrences of what is sought, with related information and links to the resource, will appear after the search is launched. What was missing were pages where resource information was aggregated to gather all connected data and their links to empower the navigation into a semantic environment. Therefore, for each resource resulting from the search, it was made possible to access a specialised view. Each page, describing a specific type of resource or context, reports:

- A label describing the resource;
- The indication of the domain of reference;
- Four action buttons to:
 - o access the original resource through its link,
 - \circ save the hyperlink of the resource,
 - o save a citation,
 - access the navigation graph (LodLive tool).

Aside from these common elements each page shows information related to the item, together with links to other resources, original domains and vocabularies. With the interconnections between resources being indicated, the triplified resources enable the transition from a conventional display (i.e. lists of results) to a representation of items in a data graph with their semantic implications. This process helps to explore the graph by creating a network of paths that, starting from an entity (single data), allows the user to browse every data connected to it. It's always possible to switch to a simpler data

navigation system and, eventually, return to the original data provider website. As previously mentioned, specific views were provided for each resource class, including Object (primarily archival documents), Artwork, Letter, Person, and Place. Some various representation options were introduced depending on the type of data.

4.2. Time and space visualisation: timeline and maps in a semantic environment

In typical Kantian fashion, RESTORE considered time and space as reliable dimensions for the sake of data aggregation and representation, since they could be used to consider different resources regardless of their nature, type and context of creation. Moreover, using vocabularies, we have shared identifiers that can be used also outside the context of the RESTORE project. From the technical point of view, the maps were built on top of the Leaflet.js library³⁴, while the timelines were custom made using javascript and jquery and styled with CSS3. From the data perspective, in RESTORE data related to places has been integrated to provide the following information:

- Geographical localization on a map using Global Positioning System (GPS) coordinates: in the "place" custom view the user can see the exact location of the selected place on a map, given that coordinates were provided;
- Connections with other places: in the "place" custom view it is also possible to browse additional information about places such as province, municipality, country, micro toponym, etc. as well as relationships between different places.
- Localization of events related to objects and people (e.g.,: production, creation and transformation of objects, movements of artefacts, place of birth and death for individuals, workplace, travel place, etc.). Through the use of CIDOC it was possible to define and model a great variety of events. For each event (if available) RESTORE will present information such as: chronology, place where the event happened, people involved in the event etc.

When both spatial and chronological data is present for the same type of resource, an aggregated visualisation - including both a map and a timeline - is available: the map shows the location of the items involved throughout their coordinates, while the timeline shows their distribution in the chronological scheme. Both the map and the timeline are browsable and change dynamically when an element is selected. To keep track of the elements involved there is also a section with information about the resources in textual form that dynamically changes according to a time or place being selected by the user. When a particular place is selected only the resources produced in that specific place are shown. As for the timeline, it starts with granularity set to "years"; by selecting a specific year, the result set is filtered down to show only the relevant dates for that year. By selecting a date only, the corresponding resources are shown, and the markers on the map are also filtered to match the data. By shuffling between these parameters, it is possible to visualise only the resources of interest on the time and space axes. Each resource comes with a link to the original website or information system of provenance, so that the user can check it.

Within RESTORE, resources of the "document" type have been processed to extract contextual information needed to implement the style of visualisation described before, such as:

- Temporal information: complete date, day, month, year;
- Spatial information: placename, coordinates;
- Textual information: subject, text.

Since both temporal and spatial data have been collected it was possible to manage them at the same time, implementing dynamic timelines and maps based on the same data array, exploiting the connection to the SPARQL endpoint to launch specific queries. The needed information is gathered via the SPARQL endpoint collected into a multidimensional array and then used to feed the visualisation tools. See below an example of an element of the array:

```
[1] {
    coordinate: "43.716667 , 10.383333"
    data: "13.12.1384"
```

```
34 Leaflet library:https://leafletjs.com/
```

```
evento: "Invio"
giorno: "13"
mese: "1"
anno: "1384"
luogo: "Pisa"
soggetto: "DATINI FRANCESCO DI MARCO E COMP. a BASCIANO DA PESCINA"
testo: "[...]"
uri_info: "http://datini.archiviodistato.prato.it[...]/ASP000163311/E73_OVI"
uri_luogo: "http://dev.restore.ovi.cnr.it/vocabularies/places/273"
uri_soggetto: http://datini.archiviodistato.prato.it/[...]/ASP000163311
```

(1)

Spatial and temporal information have been integrated in the RESTORE knowledge base regardless the context of provenance therefore the results obtained by searching for these data encompass different domains: if a domain is not specified, users can see all the resources connected to a given date or place regardless of the source. For example, the lemma "vino" appears in 156 documents among the resources collected into the RESTORE dataset. Each document was produced at a specific time and in a specific place. So, for the lemma "vino", the documents that have it in their text present a chronological indication between 1384 and 1411. For the geographical distribution we can count a total of 12 different places where the documents have their origin. These data are represented in the time-space visualization interface: the 12 places are distributed on the map and a timeline with range 1384-1411 is generated. At the side of the map are reported the texts of the documents with the lemma "vino". All this information is connected (text to place and to time) so, by clicking a place, only the documents with the lemma "vino" with origin in that place remain and the timeline also shuffles to have only the dates connected to the remaining documents. At the same time, if a year on the timeline is selected the timeline changes to show a more detailed view of dates within that year so that the user can select only the chronological reference of interest. While doing so the text are reorganized to show only the texts with the lemma "vino" with that chronological reference.

4.3. Collection of works of art (painting collections, image catalogues, image documentation, historical inventories)

Since the data ingestion process related to works of art and the related documentation included the need to display original images of each collection piece, it was necessary to implement a tool to provide a smooth visualisation experience. The choice fell on the OpenSeaDragon³⁵ viewer for its highly customizable properties, the IIIF³⁶ support, and eventually the possibility to implement it within the EVT³⁷ tool, used by RESTORE for the visualisation of textual material. In the RESTORE project most of the images were scanned images and photographic records of documents: images of documental material comprising historical guidebooks dated back to 1858 and 1888 (the 'Guida 1858', 'Guida 1888' labels); digitised inventories entries that goes from 1970s to 1990s (the 'Scheda Storica' or the 'Inventario 1956' labels); digitised Museum catalogues complete of critical texts and tables, so that the documents are in direct relationship with the information that image convey; those images are defined as visual representation of documents in our mapping schema. To summarize, within the RESTORE project, data about works of art and images were provided for the most part by the Museo di Palazzo Pretorio of Prato. The Museum provided data for: i) the original record in historical inventories for each collection piece, ii) 1858 guidebook³⁸, iii) 1888 guidebook³⁹, iv) 1912 guidebook⁴⁰, v) 1956 inventory, vi) 1958 guidebook⁴¹. For each piece of this documentation, the related images of the pages documenting it, were provided by the Museum, and also by the Archivio di Stato di Prato with regard to user's guide sources. In the template we created for works of art ('Scheda opera d'arte'), the written records are organised into the following four sections: i) General information ('Informazioni'); ii)

³⁵ OpenSeaDragon - OSD: https://openseadragon.github.io/

³⁶ International Image Interoperability Framework - IIIF: https://iiif.io/

 ³⁷ EVT is distributed as open-source software, all current code base is available on GitHub: https://github.com/evt-project/evt-viewer-angular
 ³⁸ Guasti, G., Alcuni quadri della Galleria Comunale di Prato, Prato, Tipografia Guasti, 1858.

³⁹ Guasti, G., I quadri della Galleria e altri oggetti d'arte del Comune di Prato, Prato, Tipografia Giachetti, figlio e c., 1888.

⁴⁰ Papini, R., Catalogo della Galleria Comunale di Prato, Bergamo, Istituto Italiano d'Arti Grafiche, 1912.

⁴¹ Marchini, G., La Galleria Comunale di Prato, Firenze, Arnaud, 1958.

Visualisation of the image of the work of art ('Immagine Opera'); iii) Historical records, analyses, and notes of art criticism ('Notizie Storico-Critiche'); iv) Bibliography ('Bibliografia'). The General Information section includes many different "labels", through which the user can navigate resources connected to the documentation about the work of art and its available original sources and digitised materials. First of all, in this first section it is possible to read the following information: Author (name, dates related); Datation; Subjects; Iconclass codes; Type; Material and Technique; Measures; Custodian Institution; Conservation Status; National catalographic ID; Image Provenance and License. Data in this section is useful to provide a general understanding of the work of art according to the format in which it is usually catalogued following the ICCD (Istituto Centrale per il Catalogo e la Documentazione) guidelines: it documents the unique identifier (NCT) number of the work of art (i.e. the ID connecting the work of art to the National database), its title, subject, chronology and other information useful to physically locate the piece and to get at its actual status. The section is dynamically populated with the information associated with the resource in the semantic environment. The second section displays an instance of the aforementioned OpenSeaDragon viewer. This tool was integrated in order to have a high-definition and zoomable environment for the visualisation of the images connected to the resource. These images include the work of art's digital reproduction, which is shown in the main section, as well as the digitised pages of historical guidebooks and inventories, made accessible by a navigation menu at the top of the General Information section, allowing to browse between the different tabs and therefore providing a solid scholarly contextualization of the artwork. By selecting a menu item, the General Information section and the Image Viewer section change dynamically showing the requested data: for instance, when a user gets through the General Information section, choosing a specific guidebook or inventory to navigate, the section will show, if available, information about: Author; Date; Subject; Bibliographic ID; Description; Notes; Provenance and Image Licensing. All the documentation, both in textual and image form, is provided in an integrated interface, allowing the users to visualise different kind of resources and get comprehensive information about the artwork. This customised visualisation system allows contextual consultation of different documents, together with advanced image support, gathering all the documents in one place. Since it is highly customizable it is possible to continue adding related documentation and set the viewer for the most effective fruition experience. Furthermore, the "historical and critical notes" section provides additional information, originally only available in handwritten form. The last section is dedicated to bibliography and collects all the documented publications, books, essays, magazines, articles connected to that resource.

4.4. Letters, integrated views for documents

Letters, as a type of document, have a custom view. Data about a set of letters (3012 documents) came from two different domains, namely Archivio di Stato di Prato and Opera del Vocabolario Italiano. In terms of metadata (items descriptions), the two sources are reporting mostly the same kind of information, with the latter (CNR-OVI) offering extensive lemmatization of the texts. RESTORE wanted for both the domains to be represented into the template without confusing the sources of information, so that the user could comfortably compare them. Therefore, an alternating views visualisation mode has been implemented, allowing the different sets of information - one for the Archivio di Stato di Prato and one for Opera del Vocabolario Italiano - to be browsed through the buttons on the top of the page. This kind of visualisation also makes it possible (if necessary) to add other domains, easily browsable by the user. Below the information section there is the transcription section (when available). In this case too it is possible to switch between different versions/editions if provided⁴². The third section contains a table with the list of lemmas marked in the document. Some of the lemmas came with a link to the Tesoro della Lingua Italiana (TLIO⁴³) Vocabulary, also managed by OVI. For lemmas some information to ease their identification (type, part of speech, hyperlemma) is also provided, together with links to the TLIO vocabulary's entry (when available) and another custom view specifically realised for lexicographic entries. Lastly all information is conveyed to the Edition Visualization Technology (EVT) tool, that provides a custom view for critical editions.

⁴² For the RESTORE project we have only the transcription provided by Opera del Vocabolario Italiano.

⁴³ Tesoro della Lingua Italiana delle origini - TLIO: http://tlio.ovi.cnr.it/TLIO/

4.5. EVT, a tool for digital scholarly editions

Developed by a team of students at the University of Pisa which includes a researcher currently part of the RESTORE development team, Edition Visualization Technology (EVT) is an open-source software for publishing digital editions based on the Text Encoding Initiative (TEI) encoding scheme. From its origins, it aims to enable production and dissemination of digital scholarly editions by requiring minimal computer knowledge, so that it is accessible to the philologist and does not require the intervention of the computer scientist. Due to the features described just above, the user-friendly interface and the applicability to different kinds of editions, the software gave rise to several collaborations extending and enriching it, and finally leading to the redefinition of its infrastructure [8][9]. EVT was chosen because it meets the following requirements: it is a lightweight, open-source tool specifically designed to create digital editions from encoded texts in the XML format, allowing the user to browse, explore and study the digital editions through a user-friendly interface. The work on implementing EVT for OVI resources is part of the effort to develop new methodologies for text visualisation and analysis. The goal of EVT implementation is to offer advanced navigation methods that allow users to explore in detail the lexicographic aspects of a text.

4.6. State of the art

The latest version of the software, EVT 3, is currently in active development - by the EVT development team, and as already indicated by one of the authors of the article who is a part of it - and differs from the previous one for increased flexibility in handling the many, sometimes very complex, features, and for adopting the Angular44 framework, which incorporates most of the good software design patterns and encourages best practices related to web application development, updating standard technologies. A remarkable consequence of the adoption of the new Angular framework is reflected in the new XML/TEI parser implemented in EVT 3. The parser is charged with transforming one or more valid XML/TEI documents into a Typescript Object Model that represents the actual subset of encoded information that EVT can handle and visualise. The architecture is designed to be as modular as possible in order to allow the arrangement of small building blocks that cooperate in the construction of new views that, in turn, will handle the display of different levels of edition and features.

4.7. Features and improvements

Distributed Facsimiles

Compared to the previous version, EVT 3 implements a component that can display digital facsimiles described in a IIIF manifest.json file. The manifest can be local to the displayed edition, or it can be available remotely as an URI. The image viewer is based on the well-established OpenSeaDragon viewer. In addition, support for viewing images in static mode (local storage) has been added. The mode of document image input (text-image display mode) can therefore be done with a IIIF manifest or, alternatively, by indicating the local folder path. In order to implement these two image management modes, the following components were added:

- dynamic mode image display support (IIIF framework): component for displaying digital facsimiles described in an IIIF manifest.json file. The manifest can be local to the displayed edition, or it can be available remotely as a URI;
- support for displaying images in static mode (local or remote storage).

Interpretive edition with lexicographical apparatus

A new feature implemented in EVT 3 is the support for the management and visualisation of a lexicographic apparatus; this functionality allows the user to visualise and analyse lemmatised texts in detail. As part of the RESTORE project, the resources provided by OVI represent the use case for the development of an integrated component to support the visualisation of an interpretive edition with lexicographic apparatus. EVT has been implemented in order to accommodate lexicographic elements

⁴⁴ Angular: https://angular.io/

(lemmas, *hyperlemmas* and additional associated information) using XML-TEI to carry specific encoding for lemmas and detailed descriptions of them. Such an implementation makes it possible to examine all the different occurrences of lemmas in the same text, their normalised form, grammatical category etc. In addition, the newly implemented display mode allows the user to have a text-image view, facilitating the consultation of digital facsimiles and text transcription in the same browsing environment.

Selection of lexicographic elements

By using the appropriate selectors, it is possible to select the grammatical categories and *hyperlemmas* to be highlighted in the text. Particular emphasis has been placed on the possibility of customising the elements of the final interface: therefore, the tools for selecting lexicographic elements have also been managed in order to allow their complete configuration by the editor. The available grammatical categories and *hyperlemmas* are defined in the edition configuration file "edition_config.json" In the file, lemmas to be displayed in the selector are declared, entering values in the "lemsSelectItems" parameter. The *hyperlemmas* to be displayed in the selector are declared in the file by entering the values in the "iperlemsSelectItems" parameter. The tool for displaying grammatical categories and *hyperlemmas* in the text is accessible through special selectors placed at the bottom within the frame containing the text, so that the user can choose which elements to highlight in the text. Clicking on the drop-down menu will show a list of existing grammatical categories within the configuration file, each represented by its own identifier.

Once the lists are open, it will be possible to select the desired items using the check box provided. In the window containing the text, each selected category will be highlighted with its corresponding color.

Detailed view on lexicographic elements

This feature is available due to the newly implemented visualisation mode, which allows to compare the text with detailed lexicographic information within a single interface. Support for detailed visualisation of lexicographic elements provides for each lemma all available information reported in the encoding:

- Normalised form: as shown in the TLIO vocabulary (e. g. singular masculine or feminine for nouns; singular masculine for adjectives; present infinitive for verbs etc.);
- *Hyperlemmas*: conceptual categories identifying specific semantic fields such as economy and finance, buildings and architecture, animals, medicine, sailing etc.;
- Grammatical category: parts of speech, named entities (person and place names) etc.;
- Notes: editorial notes, if useful for understanding lemmatization choices, and observations on the uses of some place names;
- Link to external vocabularies and resources: with reference to the specific lexicographic entry in the TLIO vocabulary

In addition, the number of occurrences and their position in the text can be displayed for each lemma.

5. Conclusions and future outcomes

The RESTORE project [10] officially ended on 31 August 2022. The final deliverable is a full set of tools for transforming data from the original format to semantic triples, navigable in an integrated platform that provides specific visualisation modes. Due to its architecture composed of highly integrated and customizable components for a variety of data encoded with different standards, the workflow can be replicated and applied to other resources in other contexts. Even if the project is closed its technology continues to be used.

6. Acknowledgements

The RESTORE project was co-financed by the Region of Tuscany and coordinated by the Istituto Opera del Vocabolario Italiano of the Italian CNR (National Research Council of Italy). The project collaborated with national Cultural Heritage institutions, such as the State Archives and the Museum of Palazzo Pretorio of Prato, the Archival and Bibliographic Superintendency of Tuscany, and the SPACE

SpA software company. The data displayed on the platform as a result of the search are provided under a Creative Commons License 1.0 Universal Public Domain Dedication (CC0 1.0^{45}).

Original images provided by partners and visualised on the platform in their original, unprocessed format, (such as images) have an attribution Non-Commercial Share Alike (CC BY-NC-SA⁴⁶).

7. References

- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). doi: https://doi.org/10.1038/sdata.2016.18
- [2] Wisser, K. M.. Describing entities and identities: the development and structure of encoded archival context—corporate bodies, persons, and families. Journal of library metadata, 11(3-4), 166-175 (2011). doi: https://doi.org/10.1080/19386389.2011.629960
- [3] Pitti, D. V.. Enhancing Access to Contextual Information on Individuals, Families, and Corporate Bodies for Archival Collections (2013). doi: http://dx.doi.org/10.17613/M61T0Q
- [4] Ministero dei beni e delle attività culturali e del turismo. Istituto per il Catalogo centrale e la documentazione. Normativa OA – Opere e oggetti d'arte, Versione 3.00. Struttura dei dati delle schede di catalogo ICCD (2018). URL: http://www.iccd.beniculturali.it/getFile.php?id=7508
- [5] TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.3.0. Last updated on 31st August 2021. TEI Consortium. URL: http://www.teic.org/Guidelines/P5/
- [6] CIDOC CRM Special Interest Group. 2022. Definition of the CIDOC Conceptual Reference Model Version 7.2.2. URL: https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.pdf
- [7] Francesco Coradeschi, Emiliano Degl'Innocenti, Carmen Di Meo, Maurizio Sanesi, Alessia Spadi, and Federica Spinelli. RESTORE: Opening Data in Digital Humanities and Cultural Heritage, in: G. M. Di Nunzio, B. Portelli, D. Redavid. G. Silvello, Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 2022, pp. 24-25. doi: URL: http://ceurws.org/Vol-3160/
- [8] Roberto Rosselli Del Turco, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti and Jacopo Pugliese. "Edition Visualization Technology: A Simple Tool to Visualize TEIbased Digital Editions". In: Journal of the Text Encoding Initiative [Online], Issue 8, December 2014 - December 2015. doi: https://doi.org/10.4000/jtei.1077
- [9] Rosselli Del Turco, R., Di Pietro, C. and Martignano, C. 2019. Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo. Umanistica Digitale. 3, 7 (2019). doi: https://doi.org/10.6092/issn.2532-8816/9322
- [10] RESTORE web-site: http://restore.ovi.cnr.it

⁴⁵ CC0 1.0: https://creativecommons.org/publicdomain/zero/1.0/deed.en

⁴⁶ CC BY-NC-SA: https://creativecommons.org/licenses/by-nc-sa/2.0/