

RESTORE: Smart Access to Digital Heritage and Memory

Emiliano Degl’Innocenti¹, Alessia Spadi² and Federica Spinelli³

¹ CNR-OVI, via di Castello 46, 50141 Firenze, Italia
Emiliano.deglinnocenti@cnr.it

² CNR-OVI, via di Castello 46, 50141 Firenze, Italia
spadi@ovi.cnr.it

² CNR-OVI, via di Castello 46, 50141 Firenze, Italia
spinelli@ovi.cnr.it

Abstract. This paper presents the first results of the RESTORE (smaRt accESs TO digital heRitage and mEmory) project. RESTORE is a 2 years project - co-financed by the Regione Toscana - whose main goal is the recovery, integration and accessibility of data and digital objects produced in the last twenty years by the partners involved, in order to build a knowledge base on the history of the city of Prato and its institutions, the development of its economic and entrepreneurial system, the role of women in the development of the city welfare network. The reference context for scientific and infrastructural development activities is the collaboration (through the CNR-OVI institute) with DARIAH-ERIC (ESFRI Landmark for the humanities and social sciences) and E-RIHS (ESFRI project for the heritage science) Research Infrastructures, as well as the European Open Science Cloud (EOSC).

Keywords: Data and information lifecycle, Annotated textual corpora, Collections, Data repositories and archives, Methodologies and technologies for the production, Representation, Preservation and enhancement of cultural heritage, Models and tools for semantic representation and Knowledge management in the humanities.

1 Introduction

The ecosystem of digital resources concerning cultural heritage is large and very rich, but it suffers from a very high fragmentation level, so high that it threatens to damage its value: quality content - produced by libraries, archives and research centers - is often difficult to find and scarcely accessible. Moreover, it is also characterized by a very low level of interoperability, since these subjects work in very specific operational and scientific contexts, or without any relation with other authorities.

During the last decades the trend to generate large amounts of digital information accelerates from year to year. Martin Hilbert identified the year 2002 [1] as a turning

point in the accumulation of knowledge on a global level (i.e. “the beginning of the digital age”¹: for the first time in the history of humanity there is more data produced and stored in the digital domain than in the analogue counterpart). A similar abundance and heterogeneity of information is also in Memory and Cultural institutions (i.e.: museums, libraries, archives, etc.). This situation indicates the need for a careful re-evaluation of the criteria to be used for the selection, structuring, enrichment and publication of high quality, scientifically validated data, services and tools which otherwise would remain trapped in the information systems of its creators: often poorly interoperable and, in some cases, made completely inaccessible by the lack of reference standards. A vast part of this huge amount of information available online is often fragmented and characterized by a data-silo logic. Given the high degree of specialization, these resources are often isolated from other scientific contexts and disconnected from other similar resources. A truly sustainable and interoperable digital ecosystem for cultural heritage resources should allow the reconstruction, at least in the digital domain, of a unified semantic dimension.

2 Lack of interoperability: the case of RESTORE

RESTORE² is a two years project, co-funded by the Regione Toscana, that features Research Centers (CNR-Opera del Vocabolario Italiano³), Cultural Institutions (State Archives of Prato⁴, Municipal Museum of Palazzo Pretorio⁵, Tuscany Archives and Libraries Superintendence⁶) and an SME (SPACE SpA⁷). The project aims at providing a multi-level reconstruction of the social reality of the city of Prato, starting from the Middle Ages and arriving - possibly - up to the XX Century. This will be possible through the creation of an information system populated with a relevant portion of the city's documentary heritage preserved in the memory institutions that joined the project. The RESTORE platform will enhance remote consulting of the partners' resources through the use of Linked Open Data (LOD)⁸ and by bringing to the users advanced tools to search and browse the project's knowledge base.

The State Archives of Prato preserves a vast asset of documentary resources about social, religious, economic life and history of Prato. These resources will constitute the first and more relevant building block of the RESTORE knowledge base, that will be integrated with other relevant assets, starting with the city's artistic heritage belonging to the Municipal Museum of Palazzo Pretorio.

¹ https://en.wikipedia.org/wiki/Digital_data#cite_ref-Hilbertvideo2011_3-0

² The project started in June 2020 and will end in September 2022. An overview of the project is available at: <http://restore.ovi.cnr.it>

³ <http://ovi.cnr.it>

⁴ <http://archiviodistato.prato.it>

⁵ <http://www.palazzopretorio.prato.it/it/>

⁶ <http://www.soprintendenzaarchivisticatoscana.beniculturali.it/index.php?id=2>

⁷ <http://www.spacespa.it/chi-siamo/>

⁸ <https://www.w3.org/DesignIssues/LinkedData>

The Datini⁹ and the Ospedale Misericordia e Dolce¹⁰ funds are among the oldest collections preserved by the State Archives that also produced their digital archival description, digital images of the documents, transcription of texts and a collection of lemmatized texts, in collaboration with CNR-OVI.

The Palazzo Pretorio Museum holds a collection of works of art coming from or strictly related to the Misericordia and Dolce Hospital and Datini's house.

Most of the archival heritage relevant for the project has been managed during the last 20 years with different digital platforms and tools. Due to the technological progress the interoperability and accessibility of such resources is now severely limited. As an example, the archival descriptions, images and transcriptions of the letters exchanged between Datini and his wife, published on CD-ROM about 20 years ago¹¹ is no more accessible with modern browsers and operating systems. Furthermore, the archival description of the Datini and that of the Misericordia e Dolce Hospital funds, are published on two different websites; the description of the other collections of the State Archives are published on a third, not interoperable platform. The lemmatization of Datini's correspondence is available on (yet) another stand alone website. To complete this landscape, no connection at all exists with the digital resources related to the Museum's items.

3 Building the RESTORE datastore

In order to improve the situation, fostering resources interoperability, increasing efficiency, transparency, rapid dissemination of resources and facilitating reuse, it is crucial to change the paradigm from (legacy) institutional data silos to linked open data. RESTORE aims at building a knowledge base populated with curated, scientifically reliable, information - made findable, accessible, interoperable and reusable, according to the FAIR principles¹² [2].

Resources can be exposed through a wide variety of channels, which differ according to the use of platforms and user interfaces with different qualities and characteristics. The panorama of tools for data management is extremely broad: several publishing

⁹ The Datini archive includes the administration documents and correspondence of the merchant Francesco di Marco Datini (1335-1410) that testifies, through his vast activity in the industrial, commercial and banking fields, a cross-section of the economy and social life of the entire Mediterranean basin. The digitization project began in 1999 and can be consulted on the website: <http://datini.archiviodistato.prato.it/il-progetto>

¹⁰ The Misericordia e Dolce Hospital is a charitable organisation that has cared for wayfarers, poor and abandoned children since the 13th century (cf. <https://sias.archivi.beniculturali.it/cgi-bin/pagina.pl?TipoPag=comparc&Chiave=432072&RicVM=indice&RicProgetto=as%2dprato&RicSez=fondi&RicPag=2&RicTipoScheda=ca>). Digital resources related to this fund can be consulted on the website: <http://www.archiviodistato.prato.it/accedi-e-consulta/asp005/tree>

¹¹ Per la tua Margherita: lettere di una donna del '300 al marito mercante: Margherita Datini a Francesco di Marco, 1384-1401; published in 2002.

¹² cfr. Fig. 1 (workflow diagram).

systems, including traditional institutional repositories (such as archives in Universities and research institutes), subject repositories (i.e. archives of academic publications in a particular subject area), or open access archives (designed to host miscellaneous resources)¹³ provide the means to upload and publish datasets with a general, disciplinary or institutional purpose but, in general, the support for the FAIR principles is still limited and can be improved.

Based on the specific needs of the project’s stakeholders, considered in different usage scenarios, we have identified a set of basic functional needs and selected a few core software components. As an example, in the following paragraph, we briefly analyze the selection process for the project’s data storage software. We initially considered CKAN, Dataverse, and Fedora as suitable candidates for the RESTORE data warehouse establishment: the aforementioned solutions were analyzed focusing on their ability to manage digital objects with related metadata according to the project needs and supporting basic FAIR requirements.

The candidate tools were evaluated considering some general but fundamental criteria, such as:

- supported file types;
- availability of metadata management tools;
- APIs availability, to foster data accessibility and reusability;
- User Experience, taking in consideration that the target users (GLAMs data providers) could have limited time, resources and/or know-how to spend on a new software;
- Community support;
- Available technical documentation;
- Open Source License

In principle, all the tools analyzed would have allowed us to achieve the basic goals we have set, allowing data and metadata management and dissemination through well documented APIs and being supported by a large community of users, as well as offering a series of extensible features for data processing and visualization, with open source licenses.

The reasons for choosing CKAN over Fedora and Dataverse are briefly explained in the following paragraph and summarized in Table 1, below.

Table 1. Comparative table of digital resource archiving tools.

	CKAN	Dataverse	Fedora
Filetypes supported		All file types	

¹³ Examples of these are Zenodo and Figshare, allowing all kinds of disciplines and different kinds of research content, such as tables, images, videos and datasets among their documents.

User support	Wide supporting community		
Documentation	Extensive and community maintained comprehensive documentation		
Metadata	Supports unrestricted metadata	Description/citation, domain-specific or custom fields	Follow with domain-level metadata schemas
API	Yes		
Development funded by	Government	Grants; Institutional	Organizations

CKAN is written in python and offers a simple, nice looking, and very friendly user interface. It comes packed with many features¹⁴, including tools for data harvesting, visualization and preview, RDF metadata representation for each dataset, as well as full-text, faceted and fuzzy search capabilities. A distinctive feature for CKAN is the availability of advanced APIs, allowing a variety of operations on the stored information, including: search, creation, delete and update of datasets, resources, and other objects¹⁵. CKAN is also widely extensible: custom themes and/or functionalities could be added through python packages containing one or more plugin classes, some provided by default, many others available in the CKAN extensions repository¹⁶. Furthermore, CKAN is widely used in the field of open government data¹⁷, where it has become a standard software package.

4 Creating the RESTORE knowledge base

In the first phase of the project the resources provided by the partners and the problems related to their integration were analyzed to elaborate an efficient data management workflow (i.e.: ingestion, mapping, normalization and integration, etc.) and to represent the partners' resources into the RESTORE knowledge base. The conceptual model is defined taking into account the characteristics of each partner's digital objects, as well as the main research questions and information production workflows and contexts [3, 4].

A preliminary model of the whole RESTORE digital data lifecycle is provided below (Fig. 1).

¹⁴ <https://ckan.org/features>

¹⁵ <https://docs.ckan.org/en/2.9/api>

¹⁶ <https://extensions.ckan.org>

¹⁷ <https://ckan.org/case-studies>

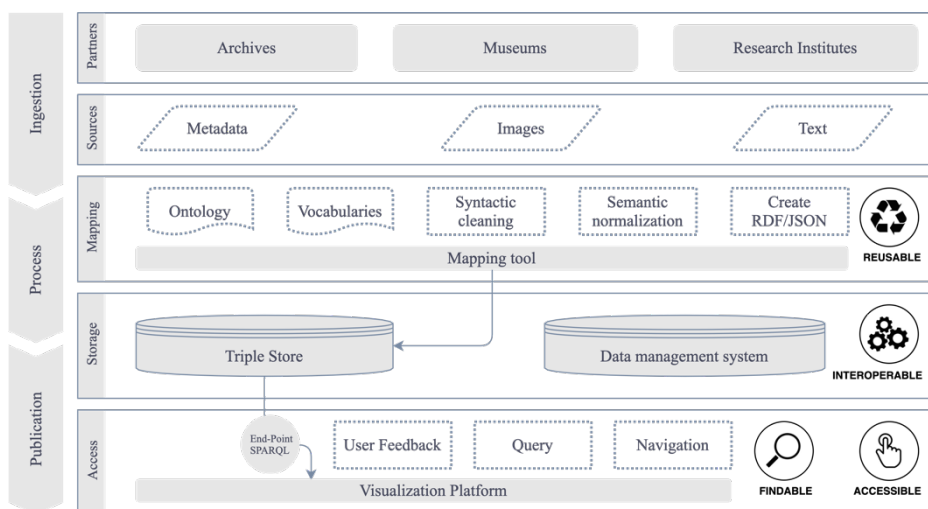


Fig. 1. Data lifecycle: from partner-supplied data ingestion to final publication, through data processing.

Based on the initial analysis of the partners' resources, an iterative methodology has been implemented to elaborate the RESTORE conceptual model (which is still under development). Here follows a brief outline of the most relevant steps in the process of identification of the classes and properties of the RESTORE ontology, based on the CIDOC Conceptual Reference Model¹⁸.

During the first phase of modeling, high-level entities and the relationships have been identified, through the analysis of existing data structures. Subsequently, other entities and relationships have been identified to support the implementation of test use cases coming from target users categories (i.e.: research communities, citizens, cultural and creative industry operators). This process has been performed in close cooperation with domain experts working in the project's partners institutions, that provided additional feedback and validated the model.

For the syntactic and semantic mapping of the most relevant schemas and standards used in each field (i.e.: TEI¹⁹ for texts; EDM²⁰ in addition to MAG²¹, MODS and METS²² for contents produced by libraries; EAD²³ and EAC²⁴ for archives; CIDOC CRM for content produced in other areas of the Digital Cultural Heritage, etc.) and to

¹⁸ CIDOC Conceptual Reference Model (CRM): <http://www.cidoc-crm.org/>

¹⁹ Text Encoding Initiative: <https://tei-c.org/>

²⁰ Europeana Data Model: <https://pro.europeana.eu/page/edm-documentation>

²¹ Administrative and management metadata: <https://www.iccu.sbn.it/export/sites/iccu/documenti/manuale.html>

²² <https://www.loc.gov/standards/mods/presentations/mets-mods-morgan-ala07>

²³ Encoded Archival Description: <https://www.loc.gov/ead/>

²⁴ Encoded Archival Context: <https://eac.staatsbibliothek-berlin.de>

perform the integration of the partners' datasets [5] we used the Karma Data Integration tool²⁵.

5 Plans for the future

Capitalizing on the work being done in the first phase of the project, the main goals for the second year of RESTORE will be represented by the consolidation of the ICT infrastructure and the extension of the project's strategy, in particular by: i) finalizing the architecture of the ICT components, ii) improving the strategies for the acquisition, mapping and modelling of the data provided by the partners; iii) developing procedures and tools for knowledge extraction and representation based on the information stored in the RESTORE knowledge base. On this latter point we will work in close collaboration with the SPACE SpA company, that will bring into the project extensive knowledge on the development of innovative tools and services for the Cultural Heritage sector (i.e.: virtual museums, smart thematic itineraries, etc.).

Eventually the RESTORE technological platform will be opened to host collections provided by other memory institutions seeking to preserve relevant digital assets and collections related to Prato and its society and extended in terms of functionalities and types of digital resources managed. The solutions produced by the project will possibly become a model to be replicated in the context of similar cultural institutions. From this point of view, the reference points will be constituted by the ESFRI²⁶ infrastructures active in the field of humanities (DARIAH-ERIC²⁷) and heritage science (E-RIHS²⁸), whose involvement will guarantee full compliance with the European guidelines on the subject. of accessibility, interoperability, reuse and creation of added value for society, starting from the results achieved by the research.

References

1. Hilbert, M.: How much information is there in the "information society"? *Significance* 9(4), 8–12 (2012). <https://doi.org/10.1111/j.17409713.2012.00584.x>.
2. Wilkinson, M.D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.
3. Gangemi, A., Presutti, V.: Ontology Design Patterns. In: Staab, S. and Studer, R. (eds.) *Handbook on Ontologies*. pp. 221–243 Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_10.
4. Gangemi, A., Presutti, V.: Content ontology design patterns as practical building blocks for web ontologies. In: *Conceptual Modeling-ER*, pp. 128-141.
5. Degl'Innocenti, E. et al.: Report on the Assessment of Data Policies and Standardization. Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization

²⁵ <https://usc-isi-i2.github.io/karma/>

²⁶ <https://www.esfri.eu/>

²⁷ <https://www.dariah.eu/>

²⁸ <http://www.e-rihs.eu/>

and Synergies - PARTHENOS Project Zenodo (2018).
<https://doi.org/10.5281/ZENODO.2662530>.