RESTORE: Opening data in Digital Humanities and Cultural Heritage

Francesco Coradeschi[x], Emiliano Degl'Innocenti[x], Carmen Di Meo[x], and Maurizio Sanesi[x], Alessia Spadi[x], Federica Spinelli[x]
[x] *Consiglio Nazionale delle Ricerche, Istituto Opera del Vocabolario Italiano, Firenze, Italy*

Abstract
The RESTORE project (smaRt accESs TO digital heRitage and mEmory) started in June 2020 with a duration of 2 years. The project consortium, coordinated by the Istituto Opera del Vocabolario Italiano of the Italian CNR (National Research Council of Italy), includes national Cultural Heritage institutes, such as the State Archives and the Museum of Palazzo Pretorio in Prato and the Archival and Bibliographic Superintendency of Tuscany, and the SPACE SpA software company. The project - co-financed by the Regione Toscana - has its main purpose in the recovery, integration and accessibility of data and digital objects collected by partner, in order to build a knowledge base made of information regarding the history of the city and of its civic institutions, the development of its economic and entrepreneurial system, the role of women in the development of a welfare state and network. Starting a local history approach, it is nonetheless possible to broaden the focus from the local dimension to reconstruct a significant part of the history of European and Mediterranean cities of the 14th century, including commercial and economical aspects. This paper presents a focused overview on the mapping and modeling of archival and museum digital resources encoded with different standards, such as XML-EAD, XML-EAC, XML-TEI and ICCD-OA, and covered in the CIDOC-Conceptual Reference Model, the ontology that was chosen as "common language" for semantic data integration. The long term sustainability for the project's results will be fostered by the collaboration with key players in the EU RIs environment, such as DARIAH-ERIC (ESFRI Landmark for the Humanities and Social Sciences) and E-RIHS (ESFRI project for the Heritage Science), as well as other actors within the EOSC Framework.

Keywords
FAIR, collections, archives, GLAM, ontologies, CIDOC-CRM, conceptual modeling, mapping, data lifecycle, Linked Open Data, semantic Web, metadata integration, cultural heritage, semantic data, knowledge management, social sciences and humanities.

1. Introduction

RESTORE (smaRt accESs TO digital heRitage and mEmory[2]), is a project coordinated by the CNR-OVI[3] (Istituto Opera del Vocabolario Italiano, Consiglio Nazionale delle Ricerche) based in Florence. The goal of the RESTORE project consists in the creation of a digital environment for data integration, with a user-friendly graphic interface focused on datasets coming from the domains of

[1]Francesco Coradeschi, Emiliano Degl'Innocenti, Carmen Di Meo, Maurizio Sanesi, Alessia Spadi, Federica Spinelli, Month 12, 2021, Firenze, Italy
EMAIL: fr.coradeschi@gmail.com (A. 1); emiliano.deglinnocenti@cnr.it (A. 2); dimeo@ovi.cnr.it (A. 3); maurizio.sanesi.eng@gmail.com (A. 4); spadi@ovi.cnr.it (A. 5); spinelli@ovi.cnr.it (A. 6)
ORCID: 0000-0003-0808-2736 (A. 1); 0000-0002-3839-9024 (A. 2); 0000-0001-9953-7753 (A. 3); 0000-0003-3917-5935 (A. 4); 0000-0002-9670-3426 (A. 5); 0000-0002-2195-3930 (A. 6)

CEUR Workshop Proceedings (CEUR-WS.org)

[2] The project website is available at: http://restore.ovi.cnr.it
[3] OVI-CNR: http://ovi.cnr.it

Digital Humanities and Heritage Science. The data integration platform will also allow users to extract relevant entities and other valuable information from datasets originally encoded by the partners in a variety of standards in a new integrated digital environment. The platform relies on a modular architecture - integrating custom components realized by the RESTORE dev team with existing solutions already available - to support the conversion of the digital resources provided by the project partners, to the CIDOC-CRM ontology. Among the supported formats:: TEI[4] for texts; EDM[5] and MAG[6], MODS, METS[7] for content produced by libraries; EAD[8] and EAC[9] for archives; ICCD-OA[10] for museum collections; formats in use in the heritage science like EDF[11], HDF5[12]. The final version of the RESTORE platform will also include components for data ingestion and storage (i.e. CKAN[13]) data management, access and integration (based on the Virtuoso[14] triplestore).

Starting from a keen analysis of the original datasets, taking into account the specific production context (e.g. libraries, archives, museums, etc.) and the integration requirements expressed by the partners, the team designed and implemented a set of custom solutions, including tools for the normalization, syntactic alignment and semantic modeling of data. using the CIDOC-CRM[15] as the reference ontology, allowing the representation of both tangible and intangible aspects of cultural heritage artifacts.

In this contribution, the integration methods so far implemented are presented, with a focus on the sources provided by the Archivio di Stato di Prato[16], the Museo di Palazzo Pretorio[17] and the Istituto Opera del Vocabolario Italiano (cfr. note 3). The target audiences for the project are represented by experts and researchers specialized in the scientific domains of reference; active citizen scientists involved in cultural institutions' activities; average people who want to get access to the consultation of data.

The project follows a FAIR[18][1] approach to data integration, and is developed in line with the guidelines for data management provided by EU projects and other international actors promoting Open Science for the SSH domains, such as the H2020 SSHOC[19]- Social Sciences and Humanities Open Cloud thematic cluster, the DARIAH-EU[20] research infrastructure, ESFRI[21] Landmark for Digital Humanities; and the E-RIHS[22] research infrastructure, ESFRI Project for Heritage Science.

---

[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 - HDF: https://www.hdfgroup.org/solutions/hdf5

[13] Comprehensive Knowledge Archive Network - CKAN: is an open source tool that allows the creation of websites based on open data and the publication of datasets, making them accessible to multiple users. To learn more, see the dedicated website: https://ckan.org/

[14] Virtuoso Open Link: https://virtuoso.openlinksw.com

[15] CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org/

[16] Archivio di Stato di Prato: http://archiviodistato.prato.it

[17] Museo di Palazzo Pretorio: http://www.palazzopretorio.prato.it/

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

[19] Social Science and Humanities Open Cloud - SSHOC: https://www.sshopencloud.eu/

[20] Digital Research Infrastructure for the Arts and Humanities - DARIAH-EU: https://www.dariah.eu/

[21] ESFRI: https://www.esfri.eu/esfri-roadmap-2021

[22] European Research Infrastructure for Heritage Science - E-RIHS: http://www.e-rihs.eu/

2. Original resources and their data structure

In this paragraph, a synthetic overview of material and sources that the team can ingest and process is given. It comprises, as samples of the workflow, archival fonds, museum collections, textual material from libraries and research centers.

Archival fonds and resources:

- the "Datini[23]" collection is the largest mercantile archive for the Middle Ages available, consisting of 150,000 letters and about 600 registers from which it is possible to extract information on: i) people; ii) costs and types of goods involved; iii) places mentioned, etc.; the "Ospedale Misericordia e Dolce[24]" collection, with its 7000 archival units, presents all the articulations of the functions of an assistance institution: from support to the wayfarer, to the care of the poor and the sick, up to the reception of the children abandoned and reared, thanks to the hospital itself, by the entire community of Prato;

Museum collections:

- a selection of metadata regarding the works of art with a connection to the "Ospedale Misericordia e Dolce" (see above). These are preserved as part of the artworks collection held by the Museo di Palazzo Pretorio di Prato, and cataloged according to the ICCD standard (i.e. "OA" - "work of art" entries);

Textual corpora and text based resources:

- lemmatized correspondence made available by the CNR-OVI Institute. The dataset contains 3012 letters from the Datini Archive and selectively lemmatized. The lemmatized elements are distributed within 22 categories (hyperlemmas) which include, in addition to personal and place names, terms and verbs pertaining to the religious field, agriculture, parts of the body, names of the week, the generic terms month and year, including: clothing and furnishings, food, animals, arts and crafts, calendar, economics, law and politics, construction and architecture, medicine, coins, navigation, kinship, leather goods and textiles, etc.[25]

3. Workflow

The aim of the project is to achieve full data integration for the project's partners datasets - produced following various procedures, normative schemas, standards and cataloging systems - and with different disciplinary concerns. It is, therefore, necessary to set up a workflow that not only allows the semantization of data within a single reference domain, but also proposes a shared data processing model, valid for multiple metadata schemes and for multiple standards. On the technical side, it was necessary to clean the data before processing, to remove elements resulting from incorrect data processing (i.e. use of different management tools, standards and styles) undermining its understanding.

---

[23] Collection of administrative documents and correspondence of the merchant Francesco di Marco Datini (1335-1410) which 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.

[24] A charitable organization that has cared for wayfarers, the poor, and abandoned children since the 13th century. The digital resources related to this fund can be consulted on the website of the Archives: http://www.archiviodistato.prato.it/accedi-e-con-sulta/aspoSt005/tree

[25] The lemmatised corpus of the Datini correspondence, produced by CNR-OVI between 2003 and 2005,has been developed with the software Gestione degli Archivi Testuali del Tesoro delle Origini - GATTO (http://www.ovi.cnr.it/Il-Software.html), the same programme that manages the corpus Tesoro della Lingua Italia delle Origini - TLIO (http://tlio.ovi.cnr.it/), in a specially dedicated version that can be queried online via the GattoWeb software (http://aspweb.ovi.cnr.it/(S(qmmiy5m0sybb4lao4qqmexyo))/CatForm01.aspx)

The technical phase is accompanied by a periodic focus with domain experts, to carefully analyze the materials to be treated and the effectiveness of the solutions provided by the team. Following the feedback provided, the team aims to implement a navigation environment that holds in itself the semantic meanings discussed with domain experts. This is possible by structuring the data as a semantic dataset and by defining the relationships that exist between entities.

As a first step taken in data management, the original resources were uploaded to a data store, for storage and access: the team has activated an instance of the CKAN tool (cfr. note 14) in which the original XML files, i.e. the sources provided by the partner institutions, are stored and described. In fact, CKAN allows for the description of datasets by means of an association of the original metadata files to its own metadata system (so, it essentially results as a description of the data collection). The project workflow involves the transformation of the data provided, regardless of the original format, into semantic triples[26]. A "triple" is a statement on the semantic data, expressed using a set of three elements: Subject, Object and Predicate; the triples are the smallest building blocks for the RDF[27] format. A triple, therefore, is the result of an arrangement of the data itself in the form of subject-predicate-object expressions, where the different components of the expression can be associated with a single and unique URI[28], or sequences of characters specifically identifying the resources that have been mapped and "transferred" in a semantic context. The conversion of the original data into triples was carried out through the use of custom parsers[29].

The first parser of the workflow takes in input the original data sources, and outputs them in the form of tables results encoded in a CSV[30] format, following a 1:1 ratio (the corresponding CSV file is produced for each original file). So, the program processes the original data by ordering them in a predetermined way, corresponding to the display of the same aggregated data in tabular form. The next step in data processing involves transforming the information contained in the CSV tables into triples, in the TTL[31] format. This phase is also achieved through the use of a parser implemented *ad hoc* in Python. Data triplication of the data, or the conversion of the same from the table format (CSV) into triples (TTL), implies a further analysis of the data, for it is based on the nature of the information to be conveyed. It is, therefore, necessary to make the following operations: i) choosing the most relevant information (with the help from experts in the relevant scientific domain); ii) specifying the meaning of the information to be represented (depending on the context of use, etc.); iii) providing for the actual modeling in CIDOC-CRM, choosing the appropriate entities and properties.

In the context of the activities described, the mapping indicates a process of aligning information of the same kind, though contained in different datasets, and 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: signature, author, place, etc.).

The data mapping phase facilitates the migration of data to be converted into logical constructs articulated on the basis of the CIDOC-CRM model. The ontology provides the developers with a model of management and semantic representation of knowledge bases allowing for defining a set of concepts and describing fundamental relationships between them. The finality of the process is to formalize the knowledge (i.e.: translate it into a formal language[32]), so as to minimize its ambiguity[33] (typical of natural language) and make it "computable", machine readable, through a module called reasoner, so the calculator can make inferences about it. Through the modeling, the data becomes

---

[26] See: https://www.w3.org/TR/rdf11-primer/#section-triple

[27] Resource Description Framework - RDF: https://www.w3.org/RDF/

[28] Universal Resource Identifier - URI: https://www.w3.org/wiki/URI

[29] For the generic definition of parser, see the Treccani online: https://www.treccani.it/enciclopedia/parser/. In this specific case, parsers are programs in the Python programming language for transforming data from one format to another.

[30] Comma Separated Values - CSV: https://datatracker.ietf.org/doc/html/rfc4180

[31] Turtle - TTL: https://www.w3.org/TR/turtle/

[32] cfr: Gottlob Frege, Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens, Halle a. S.: Louis Nebert, 1879: https://gallica.bnf.fr/ark:/12148/bpt6k65658c

[33] cfr: Bertrand Russell, On Denoting, Mind 14, no. 56 (1905): 479–93: http://www.jstor.org/stable/2248381

structured according to the rules of the reference ontology, which will form the basis for the transformation of the datasets into triples.

To proceed with the mapping and modeling it was necessary to conduct an in-depth study of the EAD and EAC standards archival data encoded with such a standards [2][3]; an in-depth study of the ICCD regulations (OA forms, regulation 3.00, 2018) for the entries regarding the collections of works of art [4]; an in-depth study of the TEI standard for the representation of texts in digital format [5]; the study and selection of the reference ontology which could better be used as a semantic basis to build the triples. Through the use of the common ontology, the data were integrated and aligned to the CIDOC-CRM entities and properties. Of particular interest are the entities that in the mapping phase were identified as possible points of alignment between the different datasets, present in all the resources: persons, places, objects. The effective integration of the data, is based on the reference to a unique identifier (ie: a unique and possibly persistent URI) for the entities common to multiple datasets (ie: the toponym Prato, or the anthroponym Francesco di Marco Datini), which constitutes the point of contact between competing or ambiguous descriptions, and/or different information relating to the same objects (i.e.: people, places, etc.) coming from different information contexts (archives, museums, libraries, etc.).

4. Mapping and Modeling

After a deep analysis of the standard used by the different institutions that provided the database of the project we started the modeling activity from the elements identified as conjunction points between the collections at disposal, following the chosen ontology, CIDOC-CRM. In addition to the ontology (a scheme) it was also necessary to select a series of reference vocabularies, for each domain of interest (authors, artists, places, etc.). 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, and general concepts both materials and immaterials. The common entities that occur in the different datasets will therefore have to share the link to the reference vocabulary. In this way, the integration of data on 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.

4.1. Defining objects and identifiers

The standards studied and described in this paper concern the description of handmade things. From the CIDOC-CRM perspective, a handmade thing is a physical object purposely created by human activity. For this reason, the *E22 Man-Made Object* CIDOC-CRM class has been used for representing the object, which the information refers to. Actually, the alignment to the CIDOC-CRM ontology made it necessary to distinguish between two levels of understanding in regards to the description of the objects treated. The artifacts (man-made objects, such as texts, works of art, written records of any kind) were, in fact, considered both as a product of intellectual work - as an “idea” (the stages of its commission and design by the author/artist) - and as physical, tangible manifestation of the same type of work (as elements forming part of a physically identifiable collection and, therefore, placed, musealised or archived, protected and even kept by organizations and “personas" throughout history).

The first level designates the immaterial (intangible) aspect of the work from the moment of its creation (even if only conceptual), while the second level describes its production and transposition onto a physical (tangible) support. The distinction of the two aspects is indeed important, because those different points of view equally contribute to the realization of the object as a whole; for the purposes of data mapping, this fact imposes a reflection on the information to be circumscribed and described, requiring a correct, mutual and concrete understanding. The two levels just distinguished are therefore expressed through the CIDOC-CRM ontology, as follows:

- E73 Information Object: information object (idea)
- E22 Man-Made Object : physical object (reification)

The physical object refers to all the physical and tangible characteristics of the work, while the information object refers to all the conceptual information.

As a physical object with a physical mass the resource has a physical location, expressed through unique identifiers described by catalogs and archives and assigned to resources. These identifiers are identified respectively in the signature for documentary materials preserved in archives and libraries and in the Unique Property Code (NCT) for artworks.

Traditionally, the shelfmark of archival documents is composed of the name of the fonds and the identifier of the archival unit, possibly divided into several sections (envelope, insert or file, code of the documentary unit). On the other hand, artworks are cataloged according to the ICCD (OA schema [34]) regulations with a unique identification code (NCT): this is a number given by the sequence of values assigned to the subfields Region Code (NCT+R) and General Catalog Number (NCT+N), which is assigned by the ICCD.

These two values are both to be understood as identifiers of the resources to which they refer and are therefore mapped with the CIDOC-CRM *E42 Identifier* class, which designates precisely the identifiers, to be associated with the entity that describes the physical object (*E22 Man-Made Object)* through the property *P1 is identified by*. Moreover, in order to distinguish the origin of the identifiers of the two entities, the identifier of the entity has been associated with a type, expressed by the class *E55 Type*, which is expressed respectively in "signature" and "unique asset code (NCT)". The information object (*E73 Information Object*) is also associated with its own identifier, which is the title assigned to the resources, expressed by the CIDOC-CRM class *E35 Title*. The title is in fact the most immediate reference to the information content of the object described.

4.2. People and roles

Person entities are represented by the class *E21 Person*. If persons act in groups or within complex organizations, the class *E74 Group* will be used instead. The class *E21 Person* refers to all persons who exist or existed in the past and of whom reliable historical information can be obtained. The description of the person is linked to anthroponyms with references to possible vocabularies, dates and places of birth and death, membership of groups, possible identifiers, qualifications and roles. As far as birth and death of the person are concerned, these are described by the classes *E67 Birth* and *E69 Death*, which express respectively the event of birth and the event of death, to which it is possible to link the date and place of the event.

Persons present identifiers in the datasets we treated, which could be used to associate the context of different collections to the entity. For archives the EAD elements that contain the name of the producing entity are *persname*, *famname* and *corporateBody*: these are associated with an attribute, *authfilenumber*, whose value represents the identifier that allows the link to the authority files in the EAC-CPF format. For museums the ICCD standard provides for a separate authority file (AUT) in which all the information relating to the authors of the works described in the OA files is given, to which it is linked via the author identification code.

An interesting problem in the modeling of people's information concerns the definition of the role played by the actors in an event. In fact, in a context of integration of datasets coming from different institutions when (e.g.) an anthroponym is common to both, the same person may have covered a different function or role during his involvement with different resources (i.e.: Francesco di Marco

[34] OA: http://www.iccd.beniculturali.it/it/ricercanormative/29/oa-opere-oggetti-d-arte-3_00

Datini is a merchant and a banker when he deals with his business, a commissioner when he commissions his portrait to an artist, a sender or a receiver of a letter, etc.). CIDOC-CRM ontology provides the *P14 carried out by* property to define the person responsible for an event. However, using only this property it is not possible to define the role of the author with respect to the event, so all roles are modelled as qualifications of the same person without distinction of the role carried out with respect to the different events in which he may have participated. It is therefore necessary to distinguish between the different roles played by the people involved in the creation and production of the object. To do this it is necessary to extend the CIDOC-CRM ontology with the CRM-PC schema, which introduces, among others, the property *P14.1 in the role of*. Along with that the new entity *PC14 carried out by* acts as a glue and makes it possible to define the role of an actor with reference to a specific event. *PC14 carried out by* is defined together with the properties *P01 has domain* and *P02 has range*, indicating respectively the domain (event) and the range (actor) of the entity, and *P14.1 in the role of*, defining the role of the actor in the event.

4.3. Location and places

Another important connection element in the modeling is defined by the class *E53 Place*. This class comprises physical extents in space, independent from temporal phenomena and matter. The instances of Place are referred to by toponyms that effectively represent the different appellation of places through time and linguistic variations. Moreover most events have a place associated through the property *P7 took place at* that indicates where the event happened. As far as it was possible to pinpoint the exact location of a place coordinates (latitude and longitude) where associated to it with the *P168 place is defined by* property, as to determine the position into the physical space of Earth

There is also the necessity to define the physical location of the resource. It requires, in addition to resource-specific identifiers, the indication of the current geographical and administrative location.

According to the EAD standard, the institution or agency responsible for providing intellectual access to the described materials is described through the repository element. The OA schema also includes the PVC field to document the current location (geographical and archival) of the resource.

In CIDOC-CRM such locations are also described through the *E53 Place* class, the *P54 has current permanent location* property that has been chosen to define the current permanent location of the physical object. The need to specify the current location of the resource arises because, although the repository providing intellectual access usually also has physical custody of the materials, this is not always the case. For example, a repository may assume responsibility for long-term intellectual access to electronic documents, but the actual electronic files or data systems may continue to reside in the office where they were created and maintained, or they may be held in long-term custody by entities that can provide the appropriate technical facilities for preservation.

4.4. Object Description

As far as the modeling of technical data is concerned, these are to be considered in relation to the physical object as they describe tangible characteristics of the object itself. We consider technical data associated with the physical object to be information on:

- Material: comprises the concepts of materials that compose the objects (e.g: paper, canvas, wood, ecc). It is mapped through the class *E57 Material* and it is associated with the object with the *P45 consists* of property.
- Technique: techniques used during the production of the object, described using the *P32 used general technique property* associated with the instances of the *E12 Production* event.
- Dimensions: comprises the information about all dimensions associated with the object (e.g: width, length, height but also monetary value). Dimensions are described as instances of the *E54 Dimension* class that can be specified with other information through related properties such as *P90 has value* and *P91 has unit*.

- State of preservation: the state of preservation is the information about the condition of the object at a certain time; if there are no time references, the information is considered to be relative to the time the object was cataloged and it is described with the *E3 Condition State* class linked to the physical object through the *P44 has condition property*.

- Features carried by the object: comprises all the features that can be found on the object (e.g: inscriptions) that are mapped through the *P56 bears feature* property.

In the EAD standard, the *physdesc* (physical description) element identifies the texture and all information about the physical characteristics of the material described. The information is given as free text directly within it or, alternatively, is subdivided into the elements *dimension*, *extent* (consistency of the material described), *genreform* (characteristics of the physical format), *physfacet* (information on the external/physical appearance of the material). The EAD standard also allows encoding the information used to describe the state of preservation of the document: the specific information is provided within the phystech element.

In the OA schema, the MTC field includes both the material and the technique used during the production of the work, while the MIS field concerns the measurements of the physical support of the work. The STCC field describes whether the state of the object is 'good', 'fair', 'mediocre', 'bad' or whether the data is not available. Given the nature of the information in the STC field and the phystech element, the condition is understood to be at the time of cataloging, if no other specification is present.

4.5. Events

An event is defined as a set of delimited and coherent processes and interactions, which bring about changes in cultural, social or physical systems, and which modify the state of the entity they refer to. In general, an event is qualified as any defined process that results in a change of the state of the entity it refers to. The effects of an event are not necessarily permanent and incontrovertible, but the event itself must be able to be documented. In order to document the event, some trace of it must remain in the form of the elements that participated in it. In general, an event is understood as an interaction between:

- One or more participants
- Venue of the event
- Temporal extremes

In fact, although an event may appear as having an "instantaneous" effect, in reality any material process or interaction has a spatio-temporal extension (this is why events are referred to as temporally defined entities or "temporal entities[35]", while other entities, such as people etc., are considered as persistent entities or "persistent entities[36]"). We describe below the events that have the widest use within the modeling performed so far.

4.5.1. Production

The CIDOC class *E12 Production* comprises activities that aim to create one or more new objects (i.e.: characterized by spatial extension[37]) from other materials. Production can be understood as a specialization of the activity of modification, where "modification" means a reworking of the same object, which does not change in substance, while "production" implies the creation of something "new" from the initial materials, which will be different from the final result of production. New means something that does not resemble in substance and form the materials involved in its

[35] cfr.: http://www.cidoc-crm.org/Entity/E2-Temporal-Entity/Version-7.1.1

[36] cfr.: http://www.cidoc-crm.org/Entity/E77-Persistent-Item/version-7.1.1

[37] cfr.: René Descartes Principia Philosophiae Amsterdam 1644, LXIII.

production, or something that acquires a new meaning at the level of documentation in the light of a modification, different from what it had before.

In the case of works of art, the creation of a work of art is the production of a new physical object, with different characteristics and bearing different meanings than the materials it is made of. One or more participants (author(s)) take part in the production of the work in a given place and time. On the other hand, as far as archival materials are concerned, production is to be understood as referring to the object as a whole with its physical characteristics, made up of paper, wood, ink, etc., in order to create a work of art. If we want to document the activity of writing the manuscript, therefore the conception of the content, we refer to the class *E65 Creation*, which describes the creation event.

4.5.2. Creation

The class *E65 Creation* describes an event that leads to the creation of conceptual objects or intangible products, such as legends, poems, texts, music, images, films, laws, etc.. Therefore, we chose to associate the information content of the described material with the creation event, made by an author (person, family or group), in a specific place and time.

The event "creation" has been associated with the writing of the documents of the Archivio di Stato di (and can be used - in the same way - in similar contexts, linked to the manuscript heritage etc.), as it was considered that the information content expressed by the text was to be considered separately from the support on which it is reported. As a conceptual object the text therefore falls within the scope of class *E65 Creation*.

Among the documents in the archive, it was established to map the creation event only for the register document type, while the exchange letters event, specifically created for the modeling of these resources, was chosen to be associated with the correspondence.

4.5.3. Transfer of custody

The *E10 Transfer of Custody* class describes the transfer of physical custody or legal responsibility for a physical object (so the reference is always to the *E22 Man-Made Object* entity). The use cases of the entity *E10 Transfer of Custody* are:

- taking over custody (if no previous responsible person exists);
- end of custody (there is no subsequent keeper;
- transfer of custody from one person to another;
- taking over custody from an unknown subject (previous responsible subject is unknown, not documented);
- loss of the work (the current custodian is unknown).

Note that the indication of the donor or the receiver of the custody is optional, as they may not be documented. In addition, the legal responsibility may lie with a party separate from the physical custodian(s) of the object, in which case the type of responsibility in place with respect to the transfer of custody can be documented with the *P2 has type* property. In the case of physical custody, the object must be physically in the possession of the keeper (in whole or in part in the case of compound objects).

In the OA schemas there is the field LA (other geographic-administrative locations), which basically contains the changes of hand of the work documented over time: the beginning and end of the custody at each institution is documented in a more or less precise way through the field PRD, which contains the information on the time in which a work is documented at a particular institution. The relationship between these places has been mapped with CIDOC through instances of *E10*

Transfer of Custody: the chronology of the different passages can be reconstructed and thus the actual passage of custody can be mapped[38]. If the temporal data had not been available, it would have been possible to use only the property *P49 has former or current keeper*, which associates the physical object with its current and/or previous custodians and which is also a "shortcut" to the more detailed path represented by the event described by the class *E10 Transfer of Custody*.

Instead, it was decided to map the event in order to be able to keep track of the history expressed in the OA tab. The last documented custodian of the work is actually reported in LC (current geo-administrative location), which describes the current custodian of the work. So the last custody step is between the last geo-administrative location reported in LA (based on the chronology) and the location reported in LC. All the entities described are or have been keepers of the work. In order to express the chronology (*E52 Time-span*), the PRDI field has been used (custody start date at the institution to which it refers), to indicate the date on which the custody transfer event from one institution to another took place.

4.5.4. Exchange of letters

The *E9 Move* event is used to describe changes to the physical position, i.e. the movement of the physical object. The properties *P27 moved from* and *P26 moved to* connect to instances of the class *E53 Place* that describe - respectively - the starting and ending point of the move. This class, focusing on the movement of the object, proved not to be the best solution for the representation of information related to correspondence: a typology that involves the sending and receiving of a document and therefore needs to express also essential information on the recipient and the sender. It was necessary, therefore, to implement a new class, *EL1 Exchange of letters*, having as superclass *E7 Activity*, for the definition of the exchange of letters. The two subclasses *EL2 Send letter* and *EL3 Receive letter*, indicate more specifically the activity of sending and receiving a letter. By associating to these classes the property *P14 in the role of*, it is possible to specify the actors involved and their respective roles of participation (sender and receiver).

5. Browsing the data

As stated in the workflow section, once the data modeling based on the CIDOC-CRM ontology is completed, data are converted in triples (TTL format) and uploaded to the Virtuoso triplestore. It is therefore necessary here to describe how data access is performed. The methods implemented for accessing the resources are:

- 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 Facet 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[39] instance of RESTORE (http://dev.restore.ovi.cnr.it/lodlive/): it's a visual "navigator" of resources, a tool that enhances the RDF quality data with the effectiveness of visualization through navigable graphs

Those thus listed, are very versatile tools that allow in-depth semantic searches; however, they are not really user-friendly. Therefore, having a user-friendly interface where data consultation and browsing does not require knowledge of the sparql language becomes a must, and it should allow for easy navigation, independent from the users' web proficiency. In fact, the hypothesized target is very

[38] Also in this case, the modeling takes into account the need of domain experts to reconstruct the history (i.e. origin and provenance) of documents, as part of the philological-historical investigation etc.

[39] http://lodlive.it/

heterogeneous in terms of digital literacy, ranging from citizens interested in research activities to the experts in a research infrastructure. It is therefore necessary to have more or less specialized search options, depending on the targeted user's familiarity with surfing the net, and on the level of expertise in the topic object of the webquest. For this reason, the platform offers different data visualization solutions (in practice, each target segment must have its own specific offer). The team wants to stress particularly on the definition and visualization of the relationships established between the several entities described in the data.

In brief, the platform should make the data visible, modellable, savable and downloadable by all users. All these operations must be made available within the platform.

RESTORE develops 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 simulates the queries and allows the user to navigate the semantic relationships established between data using CIDOC-CRM ontology. In fact, the aim of the platform is to have different ways of visualization and browsing of information, so that the user can find the resource of interest and read the information on the institution website that provided the information. For instance, a search made on "Francesco di Marco Datini" gives records as results standing for letters, registries or person's information; the user can scroll and refine the results and then, by selecting one among the entris, be linked to the institutions' website. In fact, RESTORE doesn't want to build a pure visualization tool of archival entries or artwork entries, given that those are already made available by the partners' interfaces. What the project achieves is the integration of data in a semantic environment that will activate new ways to reveal information through the relationships mapped during the previous stages of the work. In this perspective, the purpose of the RESTORE interface is to make this browsing activity simple for all users, and to effectively link the resource in a multifaceted context.

At the moment, an alpha version of the search interface was deployed and a testing session is set. The search interface implemented consumes the data loaded into the Virtuoso triplestore in the form of triples and simulates some simple sparql queries that can be performed on the sparql endpoint of the project.

The triples have been loaded in the triplestore into distinct graphs (the distribution is based on the original data providers) and are then modeled following the CIDOC-CRM.

First of all a search bar is made available in the search interface. The search bar is an intuitive graphical element in which the user can type keywords to launch a search operation that goes through all the data stored. The textual search by keywords corresponds to a query in which the results are filtered on the basis of the keywords entered, searched through all the resources, without further constraints. In order to refine the search results, two filters have been enabled:

- Collections: the user can select the collection to browse, narrowing the scope of the search. The collections correspond to the graphs associated with the data in the triplestore. So, selecting a collection is equivalent to defining the graph in a sparql query.
- Types of entity: the user can narrow the search by choosing the type of entity of interest. Almost each entity in the triplestre has an associated type following the CIDOC ontology. So, choosing a type equals specifying the type of entity to search in the sparql query.

The collection and the reference category are selected through the choice menu available within the search bar. The user should type his request in the search bar and select the search options of interest, if the search is successful, the resources corresponding to the selected terms will appear on the screen.

As for the appearance of the search results, a classic view in tabular form of the occurrences of what is sought with relative information and links to the resource will appear after the search is launched. Every record reports the label assigned to the resource, the name of the graph that contains

it and the type of entity; the information provided will also help the user to refine the search if needed. Moreover a series of actions are enabled starting from the resource: it will be possible to i) access a specialized view for the type of resource, ii) save the hyperlink of the resource, iii) save a citation and to access the iv) lodlive view.

The triplified data allow the move from a traditional visualization of lists of data to their semantic extensions with the relationships between resources highlighted. This mode provides for the exploration of the graph through the construction of a series of paths which, starting from the entity sought, explores other entities going through the relationships that bind them; it is always possible, nonetheless, to get back to a simpler data navigation and to get to the original websites that provided them.

The team is implementing custom views for the different types of source, and specific tools are going to be used according to the characteristics of the resources. The resources with a dedicated view will be: Persons, Places and Documents. For letters, intended as a type of a document, a custom view has been created, showing the information (metadata) of the letter, graphically placed next to the transcription of the same letter, so that the user can check the information while reading. Moreover, a customized view is being implemented with a specialized software: EVT (Edition Visualization Technology) is an advanced software tool for digital scholarly editions. EVT was chosen because it respects 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. Currently, EVT 3, the latest version to be released in January 2022, is being used for current development. (in nota: EVT is distributed as open source software, all current code base is available on GitHub https://github.com/evt-project/evt-viewer-angular).

6. Conclusions and future works

We are planning to extend the platform with other elements and ways of visualization of the data. A study on toponyms was predisposed and each toponym is being associated with the corresponding places and its coordinates. As the study goes on we are deploying maps to investigate the movements of the entities described in the Triplestore.

We are also extending the study of anthroponyms with the help of domain experts in order to have a comprehensive list of the name variants of the same person. This will facilitate further searches on actors and enhance the definition of relationships.

When the studies on these elements are concluded it will be a great help for the researchers and we'll be able to create new ways of visualization that will be appealing also for the general public because it will be possible to enhance the section dedicated to storytelling with new functions.

On another note RESTORE now has a complete set of tools for transformation of data from the original format to triples to visualization. The workflow can be replicated and applied to other resources in other contexts.

7. Acknowledgements

Special thanks are due to the project partners: Archivio di Stato di Prato, the Museo di Palazzo Pretorio and the Istituto Opera del Vocabolario Italiano. These institutions, in addition to the support provided for the interpretation and correct management of the data, were invited to test the platform for consulting the data provided within the RESTORE Project.

This evaluation process allowed the team to receive feedback from the data providers on the implemented platform and in particular on the accuracy of the search keys and the results obtained.

We would also like to thank, for their collaboration and for clarifications on the best solutions to proceed with the modeling, our colleagues Athina Kritsotaki and Eleni Tsoulouha from FORTH (Foundation for Research and Technology, Greece), active with part of the RESTORE Team within the SSHOC project. The Team works within the framework of Task 9.4 dealing with the Digital Humanities and Heritage Science Communities Data Pilot, in which the solution mentioned here was also developed.

8. References

[1] Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3, 160018.
https://doi.org/10.1038/sdata.2016.18

[2] Wisser, K. M. (2011). 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. https://doi.org/10.1080/19386389.2011.629960

[3] Pitti, D. V. (2013). Enhancing Access to Contextual Information on Individuals, Families, and Corporate Bodies for Archival Collections. 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 (2018). Normativa OA – Opere e oggetti d'arte, Versione 3.00. Struttura dei dati delle schede di catalogo ICCD. 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. http://www.tei-c.org/Guidelines/P5/ (ultimo accesso 20/09/2021)