
Mappatura e modellazione CIDOC-CRM per l’integrazione delle risorse del progetto RESTORE

¹Francesco Coradeschi, ²Emiliano Degl’Innocenti, ³Carmen Di Meo, ⁴Maurizio Sanesi, ⁵Alessia Spadi, ⁶Federica Spinelli

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

¹coradeschi@ovi.cnr.it

²emiliano.deglinnocenti@cnr.it

³dimeo@ovi.cnr.it

⁴maurizio.sanesi.eng@gmail.com

⁵spadi@ovi.cnr.it

⁶spinelli@ovi.cnr.it

Abstract

Il progetto RESTORE (smaRt accESs TO digital heRitage and mEmory) è iniziato nel giugno 2020 con una durata di 2 anni. Il consorzio del progetto, coordinato dall’Istituto Opera del Vocabolario Italiano del CNR, vede la partecipazione dell’Archivio di Stato e del Museo di Palazzo Pretorio di Prato, della Soprintendenza Archivistica e Bibliografica della Toscana e dell’azienda SPACE SpA, specializzata nella creazione di software. Il progetto - cofinanziato dalla Regione Toscana - ha come scopo principale il recupero, l’integrazione e l’accessibilità di dati e oggetti digitali che sono stati prodotti negli ultimi vent’anni dai suoi partner istituzionali, al fine di costituire una base di conoscenza riguardante la storia della città e delle sue istituzioni per poter ricostruire, ad esempio, lo sviluppo del suo tessuto economico e imprenditoriale e, non da ultimo, il ruolo centrale giocato dalle donne nella gestione e creazione di scambi d’affari, di network e nel favorire sviluppo per la città di Prato. A partire dalla figura del mercante Francesco di Marco Datini, della sua famiglia e del suo *entourage* e muovendo da una dimensione solo apparentemente locale, sarà possibile ricostruire una parte rilevante della storia delle città mercantili europee e del Mediterraneo del XIV secolo, compresi dunque aspetti di natura commerciale ed economica. Questo contributo si propone di offrire una panoramica della mappatura e della conversione dei formati realizzata per le risorse archivistiche e museali fornite dai partner (modellate sugli standard EAD, EAC-CPF, ICCD, TEI) sulla base del modello concettuale CIDOC-CRM, ovvero della ontologia scelta da RESTORE come “linguaggio comune” per l’integrazione semantica dei dati. Per quanto concerne il tema della sua sostenibilità, l’interattività e lo scambio collaborativo con gli attori centrali nel contesto delle più rilevanti infrastrutture di ricerca a livello europeo, come DARIAH-ERIC (ESFRI Landmark per le discipline umanistiche e le scienze sociali), E-RIHS (progetto ESFRI per la scienza del patrimonio) e con altri organismi presenti nel quadro della Cloud Europea della Ricerca (EOSC)

come il progetto SSHOC H2020, rappresentano sicuramente il contesto principale per la futura scalabilità ed estensione del progetto RESTORE.

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 - CNR (National Research Council of Italy), includes 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 - co-financed by the Regione Toscana - has as its main purpose the recovery, integration and accessibility of data and digital objects produced in the last twenty years by its partner, in order to build a knowledge base populated with information on the history of the city and of its institutions, to help in reconstructing (for example) the development of its economic and entrepreneurial system and the role of women in the development of a welfare state and network. Starting with the figure of the merchant Francesco di Marco Datini, his family and his entourage, and broadening the focus from the local dimension, it will be possible to also help 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 of the digital resources provided by archives and museums (encoded with different standards, such as: EAD, EAC-CPF, TEI, ICCD etc.) to the CIDOC - Conceptual Reference Model, the ontology chosen within the RESTORE project as the “common language” for semantic data integration. As for the sustainability of the project, 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, such as the SSHOC H2020 project, will represent the most relevant landscape for further extensions.

Introduzione

RESTORE (smaRt accESs TO digital heRitage and mEmory),¹ è un progetto coordinato dall’Istituto Opera del Vocabolario Italiano del Consiglio Nazionale delle Ricerche (CNR-OVI)² che ha sede a Firenze. Obiettivo di RESTORE è l’implementazione di una piattaforma software che consenta di realizzare, dalla raccolta alla pubblicazione dei dati, l’integrazione di dataset afferenti a diversi settori delle *Digital Humanities* e dell’*Heritage Science*.

La piattaforma di integrazione dei dati segue l’approccio FAIR³ - *Findability, Accessibility, Interoperability e Reuse* [1]: un insieme di principi, condivisi a livello internazionale, per la gestione e l’amministrazione di dati scientifici, codificati con diversi standard e prodotti con diversi flussi di lavoro. La piattaforma di RESTORE si basa su un’architettura modulare: il gruppo di ricerca sviluppa, infatti, componenti personalizzate o integra soluzioni già esistenti

1 Il sito web del progetto è disponibile all’indirizzo: <http://restore.ovi.cnr.it>

2 OVI-CNR: <http://ovi.cnr.it>

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

offrendo supporto agli standard di dominio, quali: TEI⁴ per i testi; EDM,⁵ oltre a MAG,⁶ MODS e METS⁷ per i contenuti prodotti dalle biblioteche; EAD⁸ e EAC⁹ per gli archivi; standard ICCD¹⁰ per la catalogazione dei beni custoditi in collezioni museali; ulteriori formati e standard in uso in diverse discipline della scienza del patrimonio culturale, come EDF,¹¹ HDF5.¹² Attraverso di essa, RESTORE intende elaborare un nuovo modello di gestione e integrazione dei dati che permetta di rappresentare - in maniera completa e interconnessa - le informazioni essenziali riguardanti gli aspetti tangibili e intangibili degli artefatti oggetto di studio delle discipline umanistiche e della scienza del patrimonio culturale.

Nella fase iniziale del progetto, sono state analizzate le risorse fornite dai partner e presi in considerazione i requisiti relativi alla loro integrazione, al fine di creare la base di conoscenza tenendo conto dei contesti di ricerca e di produzione (es. biblioteche, archivi, musei, ecc.). È stato, pertanto, necessario sviluppare strumenti specifici per la mappatura sintattica e semantica degli schemi dei metadati e dei sistemi di catalogazione rilevanti in ogni dominio di riferimento e procedere coerentemente all'allineamento e alla trasformazione dei dataset. Successivamente, è stato intrapreso il processo di conversione delle risorse fornite dai partner verso l'ontologia di riferimento del progetto RESTORE: il modello internazionale CIDOC - *Conceptual Reference Model* (CIDOC-CRM).¹³ Per raggiungere questo obiettivo, è stato sviluppato un flusso di lavoro che comprende diverse fasi - dalla raccolta e memorizzazione dei dati nel *datastore* CKAN (*Comprehensive Knowledge Archive Network*),¹⁴ fino alla loro effettiva mappatura, al caricamento dell'ontologia nel *triplestore* Virtuoso¹⁵ e al suo popolamento con informazioni provenienti dai contesti delle *Social Sciences and Humanities* (SSH) e dell'*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: <https://www.edfplus.info/>

12 Hierarchical Data Format: <https://www.hdfgroup.org/solutions/hdf5>

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

14 CKAN, acronimo di Comprehensive Knowledge Archive Network - è uno strumento open source che consente la creazione di siti web su base open data e la pubblicazione di dataset, rendendoli accessibili a molteplici utenti. Per saperne di più, si consulti il sito dedicato: <https://ckan.org/>

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

In questo contributo si presentano le modalità di integrazione attuate fino a questo momento, con particolare riguardo alle risorse fornite dall'Archivio di Stato di Prato¹⁶ e dal Museo civico di Palazzo Pretorio.¹⁷ Principali destinatari di questo progetto sono i professionisti e i ricercatori specializzati nei domini scientifici di riferimento già individuati e i cittadini attivi e partecipi ad attività collegate a quelle di natura scientifica (i.e.: citizen scientist), per i quali sarà necessario poter accedere alla consultazione dei dati mediante diversi tipi di interfacce di ricerca intuitive e *user-friendly*.

Il *workflow* del progetto è stato costantemente sviluppato tenendo presenti le linee guida per l'integrazione dei dati in una prospettiva interdisciplinare elaborate da organismi e progetti che, a livello europeo costituiscono punti di riferimento imprescindibili per la promozione della *Open Science* e dei suoi principi nei domini delle scienze umanistiche e del patrimonio culturale (SSH - CH). I principali interlocutori di riferimento per il team di sviluppo, sono stati, tra gli altri, nell'ambito del progetto SSHOC (Social Sciences and Humanities Open Cloud): CLARIN - ERIC, un'infrastruttura di ricerca che si occupa di dati provenienti dai settori di ricerca linguistico e socio-culturale; ortium of European Social Science Data Archives); DARIAH-EU-Digital Research Infrastructure for the Arts and Humanities (ESFRI Landmark); E-RIHS - European Research Infrastructure for Heritage Science (ESFRI Project); FORTH (Foundation for Research and Technology - Hellas).

Stato dell'arte

Contesto

L'Archivio di Stato di Prato è custode di un vasto patrimonio di fonti provenienti da enti e istituzioni civili, religiose e assistenziali che, integrate con quelle di altre istituzioni pratesi fra cui il Museo di Palazzo Pretorio, consentono di ricostruire la storia della comunità cittadina, del suo territorio e della sua popolazione, dal XIV secolo fino ai giorni nostri. Per i due fondi più antichi e ricchi di informazioni storiche sulla popolazione, ovvero i fondi "Datini"¹⁸ e "Ospedale Misericordia e Dolce",¹⁹ si disponeva di una descrizione archivistica informatizzata, delle immagini digitali dei documenti,²⁰ della trascrizione dei testi e di un corpus testuale lemmatizzato prodotto dall'OVI. In passato, sono stati realizzati strumenti digitali di ricerca (i.e.: il CD-ROM

16 Archivio di Stato di Prato: <http://archiviodistato.prato.it>

17 Museo di Palazzo Pretorio: <http://www.palazzopretorio.prato.it/>

18 Archivio di documenti amministrativi e corrispondenza del mercante Francesco di Marco Datini (1335-1410) che testimonia, attraverso la sua vasta attività in campo industriale, commerciale e bancario, uno spaccato dell'economia e della vita sociale dell'intero bacino del Mediterraneo.

19 Ente caritatevole che dal XIII secolo si occupa di viandanti, poveri e bambini abbandonati. Le risorse digitali relative a questo fondo possono essere consultate sul sito web dell'Archivio: <http://www.archiviodistato.prato.it/accedi-e-con-sulta/aspoSt005/tree>

20 Il progetto di digitalizzazione è iniziato nel 1999 ed è consultabile sul sito dell'Archivio: <http://datini.archiviodistato.prato.it/il-progetto/>

Per la tua Margherita,²¹ il sito web Datini online,²² ecc.), che risultano però ad oggi scarsamente interoperabili, poco accessibili e - in alcuni casi - del tutto inutilizzabili. Si tratta in larga misura di risorse risultanti da progetti e interventi distanti fra loro nel tempo e per scopi, privi del valore aggiunto dato dall'integrazione delle informazioni prodotte dai vari soggetti sul medesimo oggetto culturale.

L'obiettivo di RESTORE è stato innanzitutto quello di integrare le risorse fornite dall'Archivio con nuovi dati e oggetti digitali provenienti da altri istituti, tra cui l'ОВI e il Museo di Palazzo Pretorio. Quest'ultimo, conserva una parte rilevante del patrimonio artistico cittadino, il cui nucleo più consistente proviene dall'Ospedale Misericordia e Dolce e dalle opere che - nel corso del tempo, entrano a far parte di questa collezione - siano esse state realizzate espressamente per l'istituzione ospedaliera oppure acquisite da famiglie pratesi o enti religiosi e successivamente confluite in essa. In aggiunta alle risorse digitali qui descritte, il Museo di Palazzo Pretorio ha selezionato una serie di fonti, rilevanti per gli scopi del progetto, ancora da digitalizzare: la creazione di una base di conoscenza integrata per tali fonti consente di procedere alla creazione di un ambiente digitale personalizzato, che possa incentivare la conoscenza e la valorizzazione di una ricca porzione documentale poco nota ma in grado di gettare luce su temi fortemente radicati nella storia pratese e - in generale - toscana (il tessuto produttivo, la rete assistenziale, il ruolo delle donne in entrambi questi contesti, l'infanzia abbandonata). Le informazioni che diventano, così, accessibili, toccano da vicino tutti gli strati della cittadinanza e mirano a coinvolgere la popolazione in un percorso culturale di riappropriazione di spazi di conoscenza e di storie locali, strumenti significativi per accrescere e preservare la memoria collettiva.

Il team del progetto RESTORE ha innanzitutto analizzato le criticità individuate dagli istituti culturali coinvolti, nel contesto della raccolta, aggregazione, arricchimento semantico e messa a disposizione dei dati, al fine di costruire una base di conoscenza popolata con informazioni scientificamente affidabili, rese accessibili, interoperabili e riusabili. Al termine del progetto saranno elaborati modelli e soluzioni per la gestione, il trattamento e la fruizione del patrimonio culturale e documentario, replicabili in maniera sostenibile (sia tecnologicamente che in termini di risorse necessarie al loro mantenimento) presso altre istituzioni. La sostenibilità dell'infrastruttura tecnologica a supporto del flusso di lavoro elaborato è favorita dall'utilizzo di componenti *open source* ben documentate e già in uso presso la comunità scientifica che sono state integrate nella piattaforma di RESTORE con moduli creati espressamente per soddisfare i requisiti espressi dai partner. Il team di sviluppo sta producendo contestualmente anche i cosiddetti "training materials", ovvero manuali di utilizzo a disposizione degli utenti che vorranno replicare il *workflow* della piattaforma di RESTORE. La base di conoscenza e l'infrastruttura risultanti da queste attività saranno interoperabili nel contesto di altre iniziative a livello nazionale e internazionale, essendo basate sull'utilizzo e la mappatura di dati strutturati nel rispetto degli standard di dominio e riutilizzabile da altri soggetti con i medesimi requisiti.

21 http://datini.archiviodistato.prato.it/margherita/Root/Servizio/proj_fr.htm

22 Sito web dell'Archivio di Stato di Prato: <http://archiviodistato.prato.it>

Dati: Fonti archivistiche, collezioni museali, corpora testuali

Al momento, come detto in precedenza, il lavoro si è concentrato su risorse provenienti in larga misura dall'ambito degli archivi e delle biblioteche, con un certo numero di eccezioni costituite da collezioni museali, di cui sono state considerate principalmente le componenti semantiche legate alla dimensione intangibile. Va da sé che altre risorse - provenienti dall'ambito della scienza del patrimonio, più focalizzate su aspetti legati alla dimensione intangibile del patrimonio culturale - verranno trattate nella seconda fase del progetto.

Nel dettaglio, le risorse finora considerate sono:

- il fondo Datini, costituito da 150.000 lettere e circa 600 registri da cui si possono ricavare informazioni sulle persone coinvolte, sui costi e le tipologie delle merci, sui luoghi deputati agli scambi ecc., è ad oggi l'archivio mercantile per il Medioevo più grande al mondo;
- il fondo dell'Ospedale Misericordia e Dolce, con le sue 7000 unità archivistiche, presenta tutte le articolazioni delle funzioni di un ente di assistenza: dal sostegno al viandante, alla cura del povero e dell'ammalato, fino all'accoglienza dei gettatelli, bambini abbandonati e allevati, grazie allo stesso Spedale, dall'intera Comunità pratese;
- una selezione di opere tratte dalla collezione del Museo di Palazzo Pretorio, catalogate secondo lo standard ICCD (schede cd. "OA" - Opera d'Arte).

L'OVI ha inoltre messo a disposizione un corpus testuale che raccoglie una selezione di lettere edite, appartenenti all'ampio carteggio Datini, lemmatizzate selettivamente: nel corpus sono registrate le varietà grafiche e morfologiche di una serie di termini rilevanti, al fine di contribuire alla ricostruzione della vita economica e giuridica, nonché dei rapporti sociali dell'epoca. Sono stati lemmatizzati i nomi propri di persona, comprensivi di eventuali soprannomi e cariche specifiche (nel caso in cui l'indicazione rimandi ad un preciso personaggio storico individuato); di luogo, nomi di città, paesi, contrade, località, vie, piazze, porte, chiese, monasteri, palazzi, ospedali, enti e istituzioni, ecc.

Gli elementi lemmatizzati sono distribuiti entro 22 categorie (dette *iperlemmi*)²³ che comprendono, oltre ai nomi di persona e di luogo, termini e verbi afferenti al campo religioso o all'agricoltura, le parti del corpo, i nomi della settimana, i termini generici mese e anno, ecc. inclusi: abbigliamento e arredi, alimenti, animali, arti e mestieri, calendario, economia diritto e politica, edilizia e architettura, medicina, monete, navigazione, parentele, pelletteria e tessuti, ecc.

Il corpus lemmatizzato del carteggio Datini, realizzato dal CNR-OVI tra il 2003 e il 2005, è allestito con il software GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini),²⁴ lo

23 Cfr. la definizione di Iperlemma nella documentazione del programma GATTO - Gestione degli Archivi Testuali del Tesoro delle Origini:
[http://gattoweb.ovi.cnr.it/\(gn2uvv45jmttgo553m03r055\)/helpgattoweb/C02-P05-Iperlemmi.html](http://gattoweb.ovi.cnr.it/(gn2uvv45jmttgo553m03r055)/helpgattoweb/C02-P05-Iperlemmi.html)

24 GATTO: <http://www.ovi.cnr.it/Il-Software.html>

stesso programma che gestisce il corpus “Tesoro della Lingua Italia delle Origini” (TLIO),²⁵ in una versione appositamente dedicata e interrogabile via web.²⁶

Il corpus consta di:

- 2.511 testi
- 45259 forme
- 977.034 occorrenze di cui 126.663 lemmatizzate
- 6.510 lemmi e 22 iperlemmi.

Approccio Metodologico

Workflow

I partner del progetto RESTORE hanno fornito i dataset contenenti le informazioni sulle risorse da essi conservate. Lo scopo del progetto è, di nuovo, raggiungere l’integrazione di questi dati in modo da poterne estrarre la semantica (esplicita ed implicita) e i significati intrinseci.

Il problema è complesso, in quanto dati prodotti da istituti differenti fanno riferimento ad ambiti scientifico-disciplinari spesso distanti fra loro, che si riflettono nell’uso di standard e sistemi di catalogazione diversi per struttura e per finalità, molto spesso non automaticamente allineabili.

Si rende pertanto necessaria l’impostazione di un *workflow* che non solo permetta la semantizzazione dei dati all’interno di un singolo dominio di riferimento, ma che proponga un modello condiviso di trattamento dei dati - valido per più schemi di metadati e per più standard.

Alla fase tecnica si affianca quindi un momento di studio e approfondimento dei materiali che si vanno a trattare, svolta con il supporto degli esperti di dominio. Questa fase guida l’impiego e la realizzazione dei *tool* necessari a realizzare lo scopo prefissato. Una sintesi del *workflow* attualmente impiegato (ottobre 2021) è riportata, di seguito, nella Tabella 1.

Attività	Tool
Acquisizione dei dati forniti dai partner GLAMs (<i>Data Ingestion</i>)	CKAN Comprehensive Knowledge Archive Network

25 Tesoro della Lingua Italiana delle Origini - TLIO: <http://tlio.ovi.cnr.it/>

26 Corpus lemmatizzato del carteggio di Francesco Datini (1335-1410). Si compone di quasi 150.000 lettere ed è liberamente consultabile online attraverso il software GattoWeb: [http://aspweb.ovi.cnr.it/\(S\(qmmiy5m0sybb4lao4qqmexyo\)\)/CatForm01.aspx](http://aspweb.ovi.cnr.it/(S(qmmiy5m0sybb4lao4qqmexyo))/CatForm01.aspx)

Conversione dei dati, XML ²⁷ a CSV ²⁸	Python Algorithm (<i>custom</i>)
Mappatura e modellazione dei dati in entità e proprietà CIDOC-CRM	
Trasformazione dei dati, CSV a RDF ²⁹	Python Algorithm (<i>custom</i>) + 3M (Mapping Memory Manager) ³⁰
Importazione delle triple RDF e popolamento dell'ontologia	VIRTUOSO Endpoint SPARQL (+ <i>custom script</i>)
Documentazione del processo di trasformazione dei dati	Gogs ³¹ + JupyterLab ³²
Visualizzazione finale e <i>browsing</i> dei dati	VIRTUOSO Facets + LodLive View ³³ + EVT ³⁴

Tabella 1: Descrizione del workflow del progetto: attività e strumenti.

In primo luogo, i dati originali sono stati caricati in un *data store* per la conservazione e l'accesso: il progetto dispone di un'istanza del *tool* CKAN (cfr. nota 14) sulla quale i file forniti dai partner vengono memorizzati e descritti. CKAN permette infatti la descrizione dei dataset attraverso un sistema di metadattazione (essenzialmente, si tratta di una descrizione della collezione di dati) degli stessi, a cui sono stati associati i file originali inviati dai partner.

Il *workflow* del progetto prevede la trasformazione dei dati forniti, indipendentemente dal formato originale, in triple semantiche.³⁵ Per tripla si intende la trasformazione del dato in un

27 Extensible Markup Language - XML: <https://www.w3.org/TR/xml/>

28 Comma Separated Values - CSV: <https://www.w3.org/TR/tabular-data-primer/#tabular-data>

29 Resource Description Framework - RDF: <https://www.w3.org/RDF/>

30 Mapping Memory Manager - 3M: <https://www.ics.forth.gr/isl/x3ml-toolkit>

31 Gogs: <https://gogs.io/>

32 Jupyter: <https://jupyter.org/>

33 LodLive: <http://lodlive.it/>

34 Edition Visualization Technology - EVT: <http://evt.labcd.unipi.it/>

35 See: <https://www.w3.org/TR/rdf11-primer/#section-triple>

costruito logico-concettuale costituito da tre proprietà (entità), ovvero un soggetto, un oggetto, un predicato (*Subject, Object, Predicate*):

```
<http://datini.archiviostatato.prato.it/la-ricerca/scheda/ASPO00001427>, <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>, <http://www.cidoc-crm.org/cidoc-crm/E22_Man-Made_Object>
```

Il modello di riferimento per questo tipo di operazioni di semantizzazione è il *Resource Description Framework - RDF*. Una tripla, dunque, è il risultato di una disposizione dei dati stessi sotto forma di espressioni soggetto-predicato-oggetto, dove i diversi componenti dell'espressione sono associabili a degli URI (*Universal Resource Identifier*),³⁶ ovvero sequenze di caratteri che identificano le risorse che sono state mappate e rese in un contesto semantico. La conversione dei dati originali in triple è stata effettuata attraverso l'utilizzo di parser³⁷ sviluppati *ad hoc* in linguaggio Python.

Il primo parser previsto nel *workflow* prende in input i dati originali forniti dai partner e restituisce in output le tabelle in formato CSV (*Comma-separated Values*),³⁸ in rapporto 1:1 (per ogni file originale viene prodotto il file CSV corrispondente). Ciò vale a dire che il programma processa i dati originali ordinandoli in modo determinato e corrispondente alla visualizzazione degli stessi dati aggregati in forma tabellare. Il passo successivo nel trattamento dei dati prevede la trasformazione delle informazioni contenute nelle tabelle CSV in triple, nel formato turtle (TTL).³⁹ Anche questa fase si realizza attraverso l'impiego di un parser implementato *ad hoc* in Python. Per procedere alla triplicazione dei dati, ovvero alla conversione degli stessi dal formato tabellare (CSV) in triple (TTL), è necessario analizzarli ulteriormente, per comprendere la natura delle informazioni che essi veicolano: i) scegliendo le informazioni più rilevanti (con l'aiuto degli esperti del dominio scientifico di riferimento); ii) precisando il significato dell'informazione da rappresentare (a seconda del contesto d'uso ecc.); iii) provvedendo alla effettiva modellazione in CIDOC-CRM, scegliendo le entità e le proprietà adeguate.

Per mappatura, nel contesto delle attività descritte, si intende il processo di allineamento di informazioni dello stesso genere - espresse in dataset differenti mediante schemi e/o strutture semantiche diverse - attraverso l'abbinamento di elementi appartenenti a un dataset con quelli di un altro dataset nel caso che entrambi esprimano uno stesso concetto (ad. es.: segnatura, autore, luogo ecc.). Mappare i dati è il primo passo da compiere per facilitarne la migrazione verso un'ontologia di riferimento che, nel caso specifico è il modello CIDOC-CRM. Un'ontologia mette a disposizione degli sviluppatori un modello di gestione e rappresentazione

36 Universal Resource Identifier - URI: <https://www.w3.org/wiki/URI>

37 Per la definizione generica di parser si rimanda all'enciclopedia on line Treccani: <https://www.treccani.it/enciclopedia/parser/>. Nel caso specifico i parser sono programmi in linguaggio di programmazione Python atti alla trasformazione dei dati da un formato a un altro.

38 CSV: <https://datatracker.ietf.org/doc/html/rfc4180>

39 Turtle - TTL: <https://www.w3.org/TR/turtle/>

semantica di basi di conoscenza che consente di definire in maniera univoca⁴⁰ un insieme di concetti e descrivere le relazioni che intercorrono fra di essi. Obiettivo finale del processo è quello di formalizzare la conoscenza (i.e.: tradurla in un linguaggio formale),⁴¹ così da minimizzarne l'ambiguità⁴² (tipica del linguaggio naturale) e renderla interpretabile - ossia computabile - da un calcolatore (i.e.: attraverso un modulo denominato *reasoner*) per effettuare inferenze su di essa.

Attraverso la modellazione, vengono dunque strutturati i dati a seconda delle regole proprie dell'ontologia di riferimento, la quale costituirà la base per la trasformazione dei dataset in triple.

- Per procedere con la mappatura e la modellazione è stato necessario:
- lo studio approfondito degli standard EAD e EAC per la codifica dei dati d'archivio e delle informazioni relative ai soggetti produttori ([2];[3]);
- lo studio approfondito della normativa ICCD (schede OA, normativa 3.00, 2018) riguardante la compilazione delle schede OA - Opera d'arte ([4]);
- lo studio approfondito dello standard TEI per la rappresentazione dei testi in formato digitale ([5]);
- l'individuazione di un'ontologia di riferimento sulla cui base semantica costruire le triple.

Al fine di descrivere i diversi tipi di risorse, gli standard e normative a essi associate e procedere alla loro mappatura e modellazione si è resa inoltre necessaria l'analisi delle diverse ontologie già disponibili nel settore culturale: come modello di riferimento è stato individuato CIDOC-CRM (ISO 21127:2006) anche in considerazione della sua diffusione a livello internazionale nella comunità dei musei e del patrimonio culturale.⁴³ Pertanto, la fase di modellazione è consistita nell'allineamento delle entità descritte negli standard analizzati con le proprietà e le classi definite nell'ontologia CIDOC-CRM. La scelta degli elementi del modello concettuale CIDOC-CRM (i.e.: entità e proprietà) è stata validata con il contributo degli esperti di dominio (archivisti, conservatori museali, filologi); è stato riscontrato che l'utilizzo di CIDOC-CRM agevola il processo di "FAIRificazione" dei dati messi a disposizione dai partner del progetto, favorendo il superamento delle logiche intrinseche alle singole discipline e ai singoli sistemi di classificazione.

Attraverso l'utilizzo dell'ontologia comune i dati sono stati integrati e allineati sulle entità e proprietà CIDOC-CRM. Di particolare interesse sono le entità che al momento della mappatura sono state individuate come punti di possibile allineamento fra i diversi dataset, in quanto presenti in tutte queste risorse:

- Nomi di persona → antroponimi,
- Nomi di luogo → toponimi,

⁴⁰ Cfr. Extensional and intensional definitions

https://en.wikipedia.org/wiki/Extensional_and_intensional_definitions

⁴¹ 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>

⁴² Cfr. Bertrand Russell, *On Denoting*, *Mind* 14, no. 56 (1905): 479–93: <http://www.jstor.org/stable/2248381>

⁴³ Cfr. <https://cidoc.mini.icom.museum/>

- Estremi temporali → date/periodi.

Per ottenere una effettiva integrazione dei dati, diverse istanze di entità comuni a più dataset (i.e.: il toponimo Prato, o l'antroponimo Francesco di Marco Datini) devono fare riferimento ad un identificativo univoco (i.e.: un URI unico e possibilmente persistente), che costituisca il punto di contatto fra descrizioni concorrenti: informazioni differenti relative ai medesimi oggetti (i.e.: persone, luoghi, ecc.) provenienti da contesti informativi differenti (archivi, musei, biblioteche ecc.).

Vocabolari di riferimento

Per ottenere il risultato di cui sopra è stato necessario selezionare, oltre ad una ontologia (uno schema), una serie di vocabolari di riferimento, per ciascun dominio di interesse (autori, artisti, luoghi ecc.). I vocabolari del Getty *Art & Architecture Thesaurus* (AAT),⁴⁴ *Getty Thesaurus of Geographic Names* (TGN),⁴⁵ *Union List of Artist Names* (ULAN),⁴⁶ *Cultural Objects Name Authority* (CONA)⁴⁷ e *Iconography Authority* (IA)⁴⁸ sono liste strutturate di termini che possono essere utilizzate per migliorare l'accesso alle informazioni per quanto riguarda risorse afferenti ad arte, architettura e ai beni culturali in generale. I vocabolari del Getty Research Institute forniscono un ottimo canale per la creazione di conoscenza e per la definizione delle entità cui si riferiscono.

Il progetto RESTORE si è servito dei vocabolari AAT, ULAN e TGN per collegare le proprie entità alle URI generate dal Getty secondo le seguenti associazioni:

- ULAN come riferimento per gli antroponimi di persone e in particolare degli artisti;
- TGN come riferimento per i toponimi;
- AAT come riferimento per materie, tecniche e tipologie degli oggetti.

Nella seconda fase del progetto, una ulteriore attività sarà focalizzata all'allineamento delle informazioni onomastiche (i.e.: antroponimi) mediante il riferimento al Virtual International Authority File - VIAF.

I vocabolari CONA e IA, anch'essi sviluppati e mantenuti dal Getty, essendo ancora in fase di sviluppo e definizione, non sono stati impiegati. Attraverso i riferimenti a questi vocabolari è stato possibile identificare e indicizzare toponimi e antroponimi, due delle entità individuate come punti di contatto tra i vari dataset. Le entità comuni che ricorrono nei diversi dataset dovranno pertanto condividere il collegamento al vocabolario di riferimento. In questo modo si realizza l'integrazione dei dati sulle entità che si riferiscono allo stesso antroponimo o toponimo.

44 Art & Architecture Thesaurus - AAT: <https://www.getty.edu/research/tools/vocabularies/aat/>

45 Thesaurus of Geographic Names - TGN:
<https://www.getty.edu/research/tools/vocabularies/tgn/index.html>

46 Union List of Artist Names - ULAN:
<https://www.getty.edu/research/tools/vocabularies/ulan/index.html>

47 Cultural Objects Name Authority - CONA:
<https://www.getty.edu/research/tools/vocabularies/cona/index.html>

48 Iconography Authority - IA: <https://www.getty.edu/research/tools/vocabularies/cona/>

Infatti, se entità afferenti a diversi dataset fanno riferimento alla stessa URI, risulteranno automaticamente collegate anche attraverso i vocabolari stessi.

Vocabolario	Descrizione
Art & Architecture Thesaurus (AAT)	Vocabolario controllato utilizzato per descrivere beni di tipo storico artistico e demotnoantropologico (beni d'arte, architettura e cultura materiale). Fa riferimento a concetti di tipo ampio e generalizzato (es: dipinto, libro, cattedrale), ma non a specifici luoghi, persone, eventi.
Union List of Artist Names (ULAN)	Vocabolario controllato che contiene principalmente i riferimenti di tipo anagrafico relativi ad artisti. Include nomi propri, nomi d'arte e pseudonimi, varianti lessicali e ortografiche e temporali, traduzioni dei dati biografici in più lingue.
Getty Thesaurus of Geographic Names (TGN)	Vocabolario controllato che contiene nomi e informazioni associate ai luoghi. Include riferimenti a entità politiche-amministrative, caratteristiche fisiche degli stessi, luoghi attuali e storici.
Virtual International Authority File (VIAF)	Authority File internazionale che costituisce una base dati di voci di autorità controllate provenienti da diversi cataloghi nazionali
Iconclass system	Sistema di classificazione iconografica, strutturata per temi e motivi legati alla tradizione artistica occidentale. progettata per l'arte e l'iconografia. Contiene "classi" di concetti organizzate in modo gerarchico.

Tabella 2: Descrizione dei vocabolari di riferimento.

Modellazione

Di seguito, si descrivono alcune delle scelte operate in fase di modellazione dei dati - al fine di rappresentarne i contenuti in forma semantica, considerando la duplice dimensione - tangibile e intangibile - del patrimonio culturale.

Oggetto fisico e oggetto informativo

Per poter procedere con l'allineamento dei dati e la loro modellazione sulla base delle entità e proprietà dell'ontologia CIDOC-CRM, si è reso necessario considerare gli artefatti (opere d'arte, documenti d'archivio, ecc.) nella loro duplice veste di oggetti fisici e veicoli di contenuti (simbolici, intellettuali ecc.), considerando:

- La nascita dell'opera in quanto "idea" (le fasi della sua commissione e della sua progettazione da parte dell'autore/artista);
- La manifestazione fisica, tangibile dell'opera (quale elemento facente parte di una collezione fisicamente individuabile e, pertanto, collocato, musealizzato o archiviato, tutelato e anche custodito da Enti e "persone" a vario titolo nel corso della sua storia).

Il primo livello designa l'aspetto immateriale (intangibile) dell'opera, dal momento della sua creazione (anche solo concettuale), mentre il secondo livello ne descrive la produzione e trasposizione su di un supporto fisico (tangibile). La distinzione dei due aspetti è importante perché si tratta di punti di vista differenti che concorrono ugualmente alla realizzazione dell'oggetto - aristotelicamente considerato come *sinolo* di materia e forma - nel suo insieme e che, ai fini della modellazione dei dati, impongono una riflessione sulla consistenza dell'informazione da rappresentare. Al fine di rappresentare completamente lo spessore semantico dell'informazione contenuta nei dataset dei partner, si rendono necessari diversi passaggi interpretativi: colui che identificheremo come "padre dell'idea", ad esempio, non necessariamente corrisponderà a colui che ne crea la veste fisica o a colui che, infine, ne diventa il possessore o che ne detiene la custodia.

I due livelli appena distinti sono espressi, attraverso gli elementi dell'ontologia CIDOC-CRM, come segue:

- *E73 Information Object*: oggetto informativo (idea, concetto)
- *E22 Man-Made Object* : oggetto fisico (reificazione)

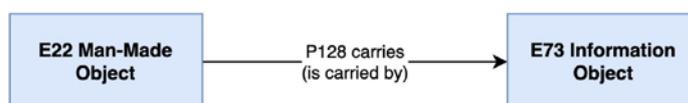


Figura 1: Diagramma delle classi e proprietà CIDOC-CRM definite per l'oggetto fisico e l'oggetto informativo.

All'oggetto fisico fanno riferimento tutte le caratteristiche fisiche e tangibili dell'opera, mentre all'oggetto informativo fanno riferimento tutte le informazioni di tipo concettuale.

Identificatori

Al fine di rendere possibile l'individuazione univoca degli oggetti, si è reso necessario associare ad essi identificatori (i.e.: signature, collocazioni, numeri d'inventario ecc.) mediante i quali essi sono descritti in cataloghi, inventari e archivi. Questi identificatori sono individuati, rispettivamente, nella signature per i materiali documentari conservati in archivi e biblioteche e nel Codice Univoco del Bene (NCT) per le opere d'arte. Va da sé che future estensioni del progetto includeranno set di identificatori in uso nelle diverse comunità di riferimento, attualmente non considerati.

Tradizionalmente, la segnatura dei documenti archivistici è composta dal nome del fondo e dall'identificativo dell'unità archivistica, eventualmente articolato in più sezioni (busta, inserto o fascicolo, codice dell'unità documentaria). Nel caso dell'Archivio di Stato di Prato, ogni fondo presenta una segnatura sua tipica che, nei file XML di origine, è stata suddivisa in più parti. Si è pertanto reso necessario, per ricostruirla, identificare gli elementi (i tag xml) che la componevano, esportarli separatamente nei campi dei file CSV e poi riunire le diverse parti, in modo coerente a seconda del documento descritto, mappando il risultato come un unico oggetto logico. Per quanto riguarda invece le opere d'arte, un bene culturale catalogato secondo la normativa ICCD (scheda OA)⁴⁹ presenta un codice univoco di identificazione (NCT): si tratta di un numero dato dalla sequenza dei valori assegnati ai sottocampi Codice Regione (NCT+R) e Numero catalogo generale (NCT+N).

Questi due valori sono da intendersi come identificatori delle risorse cui si riferiscono e sono pertanto mappati con la classe CIDOC-CRM *E42 Identifier*, che designa proprio gli identificatori, da associare all'entità che descrive l'oggetto fisico (*E22 Man-Made Object*), attraverso la proprietà *P1 is identified by*. Inoltre, per distinguere l'origine degli identificatori dei due enti, si è associata una tipologia all'entità identificatore, espressa dalla classe *E55 Type*, che si esprime rispettivamente in "segnatura" e "codice univoco del bene (NCT)". Anche l'oggetto informativo (*E73 Information Object*) è associato a un suo identificatore, che è stato individuato nel titolo assegnato alle risorse, espresso dalla classe CIDOC-CRM *E35 Title*. Il titolo costituisce infatti il riferimento più immediato al contenuto informativo dell'oggetto descritto.

Elemento EAD	Campo ICCD	Dominio CIDOC	Proprietà CIDOC	Entità/Rango CIDOC
<container type="codice">	NCT (NCTR + NCTN)	E22 Man Made Object	P1 is identified by	E42 Identifier
<unittitle>	SGTT	E73 Information Object	P1 is identified by	E35 Title

Tabella 3: Descrizione della modellazione CIDOC-CRM degli identificatori EAD e ICCD.

Personae

Le entità "persona" vengono rappresentate nella modellazione attraverso la classe *E21 Person*. Se le persone agiscono in gruppi o all'interno di organizzazioni complesse, si userà invece la classe *E74 Group*. La classe *E21 Person* si riferisce a tutte le persone esistenti - o realmente esistite - per cui si possano ricavare notizie storiche attendibili. Alla descrizione della persona sono legati gli antroponomi, i riferimenti a eventuali vocabolari specializzati e liste d'autorità - date e luoghi di nascita e morte, appartenenza a gruppi, eventuali identificatori, qualifiche e ruoli.

⁴⁹ OA - Opere/oggetti d'arte 3.00, http://www.iccd.beniculturali.it/it/ricercanormative/29/oa-opere-oggetti-d-arte-3_00

La nascita e la morte degli individui - sono descritte per mezzo delle classi *E67 Birth* e *E69 Death*, che esprimono, rispettivamente, l'evento di nascita e l'evento di morte, a cui è possibile collegare informazioni relative a date e luoghi. Gli elementi dei file XML EAD che riportano le denominazioni dei soggetti produttore sono *persname*, *famname* e *corporateBody*: a questi è associato un attributo, *authfilenumber*, il cui valore rappresenta l'identificatore che consente il collegamento ai file di autorità nel formato EAC-CPF.

Lo standard ICCD prevede a sua volta un file d'autorità a sé stante (AUT) dove sono riportate tutte le informazioni relative agli autori delle opere descritte nelle schede OA, a cui sono collegate attraverso il codice identificativo dell'autore.

Elemento EAD	Campo ICCD	Dominio CIDOC	Proprietà CIDOC	Entità/Rango CIDOC
<persname>	AUT	E7 Activity	P14 carried out by	E21 Person
		E21 Person	P98i was born	E67 Birth
		E21 Person	P100i died in	E69 Death
	AUTL	E67 Birth	P7 took place at	E53 Place
	AUTD	E67 Birth	P4 has time-span	E52 Time-Span
	AUTX	E69 Death	P7 took place at	E53 Place
	AUTT	E69 Death	P4 has time-span	E52 Time-Span
<famname> <corporateBody>	AUTU	E74 Group	P107 has current or former member	E21 Person
	AUTH	E21 Person	P1 is identified by	E42 Identifier

Tabella 4: Descrizione della mappatura CIDOC-CRM delle informazioni relative alle persone EAD e ICCD.

Ruoli

Per l'associazione degli antroponimi agli eventi, la principale criticità riscontrata ha riguardato la definizione dei ruoli: "l'individuo *x* ha partecipato all'evento *y* col ruolo *z*". Volendo integrare

informazioni relative agli individui presenti in più dataset, si è reso necessario considerare i differenti ruoli che questi possono ricoprire in essi: quando, ad es., un antroponimo ricorra in più contesti, la stessa persona può aver ricoperto funzioni e/o ruoli differenti in relazione ad eventi diversi (i.e.: Francesco di Marco Datini è un mercante e un banchiere quando tratta i suoi affari, un committente quando commissiona il suo ritratto a un artista, un mittente o un destinatario di una missiva, ecc.). Per fare questo è necessario estendere l'ontologia CIDOC-CRM con lo schema CRM-PC,⁵⁰ che introduce, fra le altre, la proprietà *P14.1 in the role of*. Infatti, usando solo la proprietà *P14 carried out by*, non sarebbe possibile definire il ruolo degli individui rispetto agli eventi: tutti i ruoli verrebbero mappati come qualifiche della persona - senza poterne caratterizzare ulteriormente il ruolo svolto.

Definendo la nuova entità *PC14 carried out by*, essa funge da collante e risulta possibile definire il ruolo di un individuo con riferimento a uno specifico evento. *PC14 carried out by* è definita insieme alle proprietà *P01 has domain* e *P02 has range*, che indicano rispettivamente il dominio (evento) e il *rango* (individuo/soggetto) dell'entità, e *P14.1 in the role of*, che definisce il ruolo dell'individuo nell'evento.

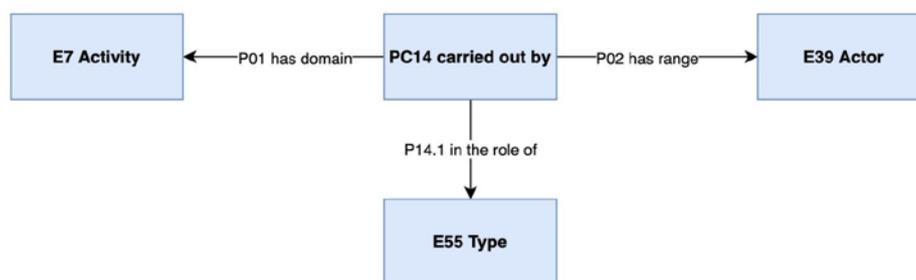


Figura 2: Diagramma delle classi e proprietà definite in CIDOC-CRM per i ruoli.

Localizzazione geo-amministrativa attuale

La definizione della collocazione fisica del bene - espressa attraverso un identificatore univoco assegnato da cataloghi o inventari e mappata mediante l'entità "Identifier" - include, oltre agli identificatori specifici per ogni risorsa, l'indicazione dell'attuale localizzazione geografica e amministrativa ad essa relativa. Secondo lo standard EAD, l'istituzione o l'agenzia responsabile di fornire l'accesso intellettuale ai materiali descritti è rappresentata mediante l'elemento *repository*. La scheda OA, include il campo PVC per documentare la posizione geografico-amministrativa attuale della risorsa, relativa al territorio italiano oppure ad organizzazioni amministrativo-territoriali di paesi esteri. In CIDOC-CRM tali posizioni sono descritte tramite

⁵⁰ Si ringraziano, per la collaborazione e per i chiarimenti sulle soluzioni migliori per procedere nell'ambito della modellazione, le colleghe Athina Kritsotaki ed Eleni Tsoulouha del FORTH (Foundation for Research and Technology, Grecia), attive con parte del Team RESTORE all'interno del progetto SSHOC (Social Sciences Open Cloud SSHOC - <https://sshopencloud.eu>). Il Team opera nell'ambito del Task 9.4 che si occupa del *Digital Humanities and Heritage Science Communities Data Pilot*, nell'ambito del quale è stata elaborata anche la soluzione qui menzionata.

la classe *E53 Place*; per definire la posizione attuale del bene è stata scelta la proprietà *P54 has current permanent location*.

Elemento EAD	Campo ICCD	Dominio CIDOC	Proprietà CIDOC	Entità/Rango CIDOC
<repository>	PVC	E22 Man Made Object location	P54 has current permanent location	E53 Place

Tabella 5: Descrizione della modellazione CIDOC-CRM della localizzazione geo-amministrativa attuale.

L'esigenza di specificare la posizione attuale delle risorse fisiche nasce dalla necessità di distinguere i differenti livelli di responsabilità degli istituti di conservazione che, pur fornendo l'accesso ed esercitando la custodia fisica dei materiali, possono non averne la proprietà. Per esempio, un archivio può assumersi la responsabilità dell'accesso intellettuale a lungo termine ai documenti elettronici, ma gli effettivi file o sistemi di dati elettronici possono continuare a risiedere nell'ufficio in cui sono stati creati e mantenuti o - ancora - possono essere tenuti in custodia a lungo termine da enti in grado di fornire le strutture tecniche appropriate per la conservazione.

Dati tecnici

Per quanto riguarda la modellazione dei dati tecnici, questi sono da considerarsi in relazione all'oggetto fisico - in quanto descrivono caratteristiche tangibili dell'oggetto stesso. Consideriamo dati tecnici associati all'oggetto fisico: i) l'informazione relativa alla composizione dell'oggetto; ii) la tecnica usata per produrlo e le sue dimensioni. Nello standard EAD, l'elemento *physdesc* (descrizione fisica) identifica la consistenza e tutte le informazioni sulle caratteristiche fisiche del materiale descritto. L'informazione è data come testo libero direttamente al suo interno o, in alternativa, è suddivisa negli elementi *dimension*, *extent* (consistenza del materiale descritto), *genreform* (caratteristiche del formato fisico), *physfacet* (informazioni sull'aspetto esteriore/fisico del materiale). Nello schema OA il campo MTC comprende sia la materia che la tecnica utilizzate durante la produzione dell'opera. Per la modellazione in CIDOC-CRM, materia e tecnica si configurano separatamente, in vista anche dell'allineamento con lo standard EAD. Indichiamo pertanto con MTC/M la materia di cui consiste l'oggetto fisico (supporto o altri materiali) e con MTC/T la tecnica utilizzata durante l'evento di produzione, mentre il campo MIS concerne le misure del supporto fisico dell'opera. Le misure degli oggetti sono espresse attraverso l'entità *E54 Dimension* di CIDOC-CRM. Attraverso l'entità *E54 Dimension* è possibile collegare l'unità di misura con *E58 Measurement Unit* e i valori numerici che vanno a definire la dimensione con *E60 Number*.

Elemento EAD	Campo ICCD	Dominio CIDOC	Proprietà CIDOC	Entità/Rango CIDOC
<physfacet type="supporto">	MTC/M	E22 Man Made Object	P45 consists of	E57 Material
	MTC/T	E12 Production	P32 used general technique	E55 Type
<extent>	MIS	E22 Man Made Object	P43 has dimension	E54 Dimension

Tabella 6: Descrizione della modellazione CIDOC-CRM dei dati tecnici EAD-ICCD.

Stato di conservazione

Lo stato di conservazione è un'informazione sulla condizione dell'oggetto in un certo momento; se non sono presenti riferimenti temporali si considera l'informazione come relativa al momento della catalogazione dell'opera.

Per la modellazione CIDOC-CRM, *E3 Condition State* descrive lo stato di un oggetto lungo un certo periodo di tempo. Si collega l'entità *E3 Condition State* all'oggetto fisico attraverso la proprietà *P44 has condition*. Lo standard EAD consente di codificare le informazioni che servono a descrivere lo stato di conservazione del documento: le informazioni specifiche sono fornite all'interno dell'elemento *phystech*. Nello schema OA, invece, il campo STCC indica se lo stato dell'oggetto è "buono", "discreto", "mediocre", "cattivo" o se il dato non è disponibile. Considerata la natura dell'informazione del campo STC e dell'elemento *phystech*, lo stato di conservazione si intende al momento della catalogazione, se non sono presenti altre specifiche. La natura della condizione può essere indicata con *P2 has type: E55 Type* associato a STCC.

Elemento EAD	Campo ICCD	Dominio CIDOC	Proprietà CIDOC	Entità/Rango CIDOC
<phystech>	STC	E22 Man Made Object	P44 has condition	E3 Condition State
	STCC	E3 Condition State	P2 has type	E55 Type

Tabella 7: Descrizione della modellazione CIDOC-CRM dello stato di conservazione EAD-ICCD.

Eventi

Per evento si intende l'insieme di processi e interazioni delimitati e coerenti, che portano dei cambiamenti in sistemi culturali, sociali o fisici, e che vanno a modificare lo stato dell'entità cui si riferiscono. In generale un evento si qualifica come un qualsiasi processo definito che abbia come risultato un cambiamento dello stato dell'entità cui fa riferimento. Gli effetti di un evento non sono necessariamente permanenti e incontrovertibili: l'evento stesso, tuttavia, deve poter essere documentato. Per documentare l'evento ne deve rimanere qualche traccia nella forma degli elementi che vi hanno partecipato. In generale un evento si intende come un'interazione fra:

- Uno o più partecipanti
- Luogo di svolgimento dell'evento
- Estremi temporali

Infatti, sebbene un evento possa apparire come se avesse un effetto “istantaneo”, in realtà qualsiasi processo o interazione materiale ha un'estensione spazio-temporale (per questo si fa riferimento agli eventi come ad entità temporalmente definite o “temporal entities”,⁵¹ mentre le altre entità, come persone ecc, sono considerate entità persistenti o “persistent entities”).⁵² Si descrivono di seguito gli eventi che hanno più ampio utilizzo all'interno della modellazione fin qui eseguita.

Produzione

La classe CIDOC-CRM *E12 Production* comprende le attività che hanno come scopo la creazione di uno o più nuovi oggetti (i.e.: caratterizzati da estensione spaziale)⁵³ a partire da altri materiali. Si può intendere la produzione come una specializzazione dell'attività di modifica, dove per “modifica” si intende una rielaborazione dello stesso oggetto, che non cambia nella sostanza, mentre “produrre” implica la creazione di qualcosa di “nuovo” a partire dai materiali iniziali, che saranno diversi dal risultato finale della produzione. Si intende come “nuovo” qualcosa che nella sostanza e nella forma non somiglia ai materiali coinvolti nella sua produzione oppure qualcosa che acquista un nuovo significato a livello di documentazione alla luce di una modifica, diverso da quello che aveva prima.

51 Cfr.: <http://www.cidoc-crm.org/Entity/E2-Temporal-Entity/Version-7.1.1>

52 Cfr.: <http://www.cidoc-crm.org/Entity/E77-Persistent-Item/version-7.1.1>

53 Cfr.: René Descartes *Principia Philosophiae* Amsterdam 1644, LXIII. *Quomodo cogitatio et extensio distincte cognosci possint, ut constituentes naturam mentis et corporis*: “Cogitatio et extensio, spectari possunt ut constituentes naturas substantiae intelligentis et corporeae”. Non si tratta di una distinzione astratta, quanto piuttosto della possibilità di modellare i dati in vista del supporto a scenari di utilizzo da parte di differenti comunità di ricercatori, interessati alle dimensioni tangibili e intangibili del patrimonio culturale: filologi, codicologi e paleografi, storici, filosofi, linguisti e lessicografi ecc.

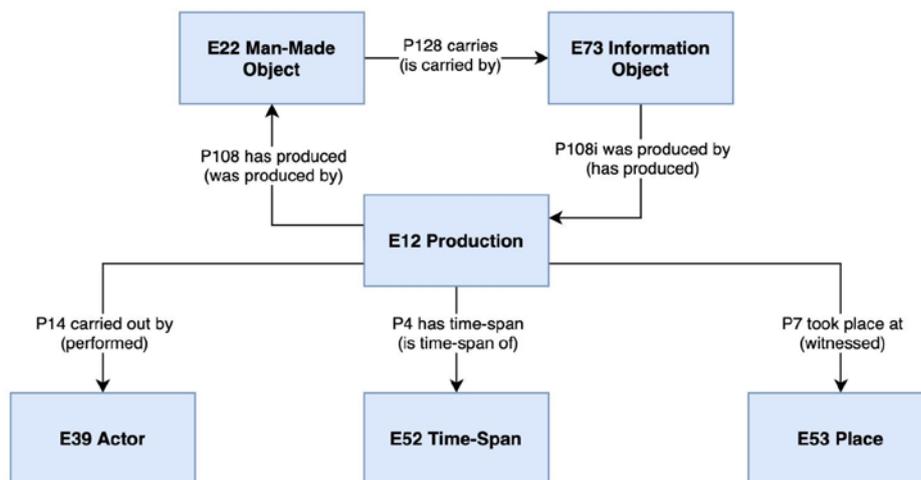


Figura 3: Diagramma delle classi e proprietà CIDOC-CRM definite per l'evento produzione.

Nel caso delle opere d'arte, la realizzazione di un'opera è la produzione di un nuovo oggetto fisico, con caratteristiche e portatore di significati diversi rispetto ai materiali che lo compongono. Alla produzione dell'opera concorrono uno o più partecipanti (i.e.: autore/i) in un determinato luogo e in un tempo definito. Per quanto riguarda invece i materiali archivistici, la produzione è da intendersi come riferita all'oggetto quale insieme delle sue caratteristiche fisiche, composto di carta, legno, inchiostro, ecc. Volendo invece documentare l'attività di stesura del manoscritto o l'ideazione della parte simbolica dei contenuti di un'opera d'arte, si farà riferimento alla classe *E65 Creation*, che descrive l'evento creazione.

Creazione

La classe *E65 Creation* descrive un evento che porta alla creazione di oggetti concettuali o prodotti immateriali,⁵⁴ come leggende, poesie, testi, musica, immagini, film, leggi ecc. Si è scelto pertanto di associare al contenuto informativo del materiale descritto l'evento di creazione, svolto da un autore (persona, famiglia o gruppo), in un determinato luogo e in un determinato tempo.

L'evento "creazione" è stato associato alla stesura dei documenti dell'Archivio di Stato di Prato (e potrà essere utilizzato - allo stesso modo - in contesti simili, legati al patrimonio manoscritto ecc.), in quanto si è ritenuto che il contenuto informativo espresso dal testo fosse da considerarsi separatamente rispetto al supporto da cui viene veicolato. In quanto oggetto concettuale il testo rientra pertanto negli scopi della classe *E65 Creation*.

⁵⁴ Cfr. la nota precedente.

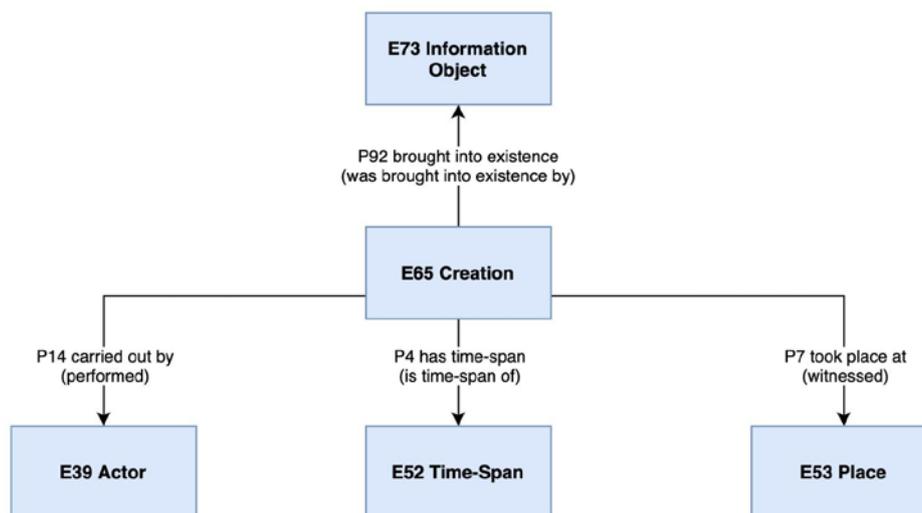


Figura 4: Diagramma delle classi e proprietà CIDOC-CRM definite per l'evento creazione.

Relativamente ai documenti dell'Archivio di Stato di Prato, si è scelto di mappare l'evento creazione solo per la tipologia di documento "registro", mentre si è scelto di associare alla tipologia "carteggio" l'evento *exchange letters*, appositamente creato per la modellazione di queste risorse.

Trasferimento di custodia

La classe *E10 Transfer of Custody* descrive il trasferimento della custodia fisica o della responsabilità legale di un oggetto fisico (quindi il riferimento è sempre all'entità *E22 Man-Made Object*). I casi di utilizzo dell'entità *E10 Transfer of Custody* sono:

- presa in carico della custodia (nel caso in cui non esista un responsabile precedente);
- fine della custodia (non c'è un responsabile successivo);
- trasferimento della custodia da un responsabile a un altro;
- presa in carico della custodia da un soggetto ignoto (il responsabile precedente è sconosciuto, non documentato);
- smarrimento dell'opera (l'attuale custode è sconosciuto).

Si noti che l'indicazione del donatore o del destinatario della custodia è opzionale, poiché è infatti possibile che essi non siano documentati. Inoltre, la responsabilità legale potrebbe ricadere su di un soggetto separato da colui o coloro che detengono la custodia fisica dell'oggetto, come specificato nel paragrafo 4.4: in questo caso, si può documentare il tipo di responsabilità in atto rispetto al trasferimento di custodia con la proprietà *P2 has type*. Quando si parla di custodia fisica, l'oggetto deve materialmente trovarsi in possesso del responsabile (interamente o in parte in caso di oggetti composti).

Nelle schede OA è presente il campo LA (altre localizzazioni geografico-amministrative) che, sostanzialmente, contiene i passaggi di mano dell'opera - documentati nel corso del tempo: l'inizio e la fine della custodia presso ciascun ente è documentato, in modo più o meno preciso, attraverso il campo PRD, che contiene l'informazione sugli estremi temporali in cui un'opera è documentata presso un determinato ente/istituto. La relazione tra questi luoghi è stata modellata con CIDOC-CRM attraverso istanze di *E10 Transfer of Custody*, mediante le quali si può tracciare la cronologia dei diversi passaggi e quindi ricostruire una parte della storia degli artefatti, relativa alla loro circolazione.⁵⁵ Se non fosse stato disponibile il dato temporale si sarebbe potuta utilizzare solo la proprietà *P49 has former or current keeper*, che associa all'oggetto fisico i suoi attuali e/o precedenti custodi e che è anche una scorciatoia rispetto al percorso più dettagliato, rappresentato dall'evento descritto dalla classe *E10 Transfer of Custody*.

Questa scelta si è rivelata efficace anche al fine di tenere traccia delle informazioni di carattere cronologico espresse nella scheda OA.

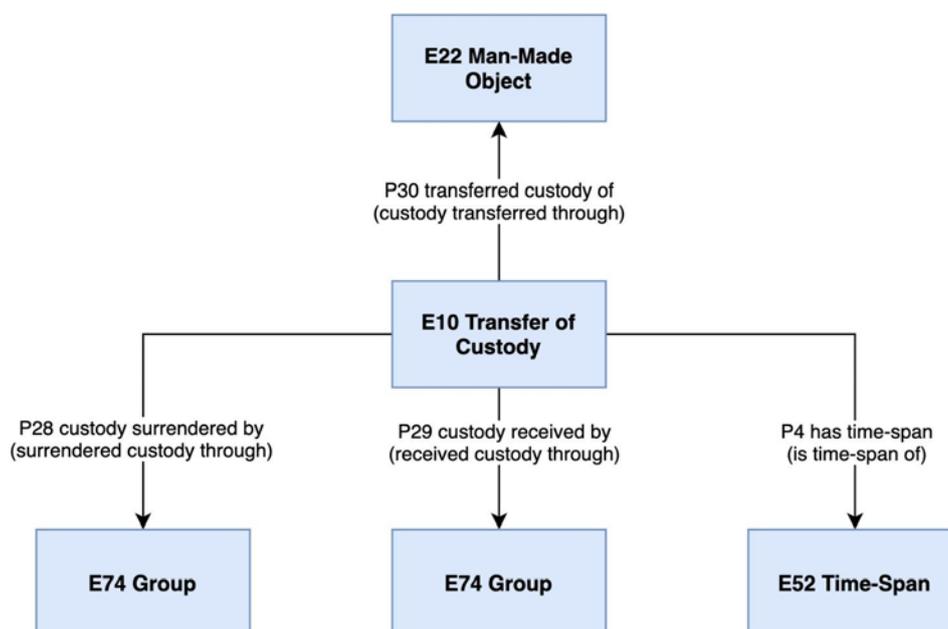


Figura 5: Diagramma delle classi e proprietà CIDOC-CRM definite per l'evento di trasferimento di custodia.

⁵⁵ Anche in questo caso, la modellazione tiene in considerazione la necessità degli esperti di dominio di ricostruire la storia (i.e.: origine e provenienza) dei documenti, come parte dell'indagine storico-filologica ecc.

L'ultimo responsabile dell'opera documentato, in realtà, è riportato nel paragrafo LC (localizzazione geografica-amministrativa attuale), che descrive l'attuale custode dell'opera. Quindi l'ultimo passaggio di custodia si articola tra l'ultima localizzazione geo-amministrativa riportata in LA (sulla base della cronologia) e la localizzazione riportata in LC. Tutti gli enti descritti sono o sono stati custodi (*keeper*) dell'opera. Per esprimere la cronologia (*E52 Time-span*) è stato usato il campo PRDI (data di inizio della custodia presso l'ente cui si riferisce), per indicare la data in cui si è svolto l'evento di trasferimento di custodia da un ente all'altro.

Scambio delle lettere

Relativamente alla tipologia documentaria del carteggio (cfr. sopra il paragrafo 4.7.2), si è

utilizzato l'evento *E9 Move* (i.e.: movimento) per descrivere le modifiche alla posizione fisica delle risorse considerate, ovvero il loro spostamento. Le proprietà *P27 moved from* e *P26 moved to* si collegano alle istanze della classe *E53 Place* che descrivono - rispettivamente - il punto di partenza e di arrivo dello spostamento considerato. Questa classe, focalizzandosi sullo spostamento dell'oggetto, si è rivelata non essere la soluzione migliore per la rappresentazione delle informazioni relative ai carteggi: una tipologia che prevede l'invio e la ricezione di un documento e che necessita quindi di esprimere anche informazioni essenziali sul destinatario e il mittente. Si è resa necessaria, pertanto, l'implementazione di una nuova classe, *EL1 Exchange of letters*, avente come superclasse *E7 Activity*, per la definizione dello scambio lettere. Le due sottoclassi *EL2 Send letter* e *EL3 Receive letter*, indicano in maniera più specifica l'attività di invio e ricezione di una lettera. Associando a queste classi la proprietà *P14 in the role of*, è possibile specificare gli attori coinvolti e i rispettivi ruoli di partecipazione (mittente e destinatario).

Prospettive future

In questo contributo è stato presentato lo stato di avanzamento dei lavori del progetto RESTORE, con un particolare focus sulla mappatura dei dati relativi alle risorse museali e archivistiche e alla loro modellazione semantica.⁵⁶ Nel secondo anno di progetto, il lavoro si concentrerà, sullo sviluppo di strumenti di ricerca e visualizzazione dei dati che, per motivi di spazio, sono stati solo accennati in questa sede. La mappatura dei dati originali, forniti dai partner in diversi formati (i.e.: XML-EAD, XML-EAC, XML-TEI e ICCD/OA), rappresenta la base logica su cui poggia la modellazione delle informazioni rilevanti in essi contenute, mediante l'ontologia CIDOC-CRM e il loro allineamento semantico. Questo lavoro si presta come base per gli sviluppi successivi del progetto (elaborazione di ragionatori automatici, sistemi specializzati per la visualizzazione integrata di dati in formati differenti, estensione delle procedure di mapping dei dati con il supporto per altri standard di dominio, ecc.). Al termine del progetto, la piattaforma tecnologica di RESTORE sarà aperta a ulteriori partner ed estesa in termini di funzionalità offerte e tipologie di risorse digitali gestite. Per la sostenibilità di lungo

⁵⁶ In questo contesto ci si è concentrati maggiormente sull'aspetto di integrazione delle risorse, ma molto altro ci sarebbe da dire su criticità e sviluppi del lavoro incontrati nell'elaborazione della mappatura messa in atto rispetto ai singoli standard.

periodo (organizzativa e tecnologica), il potenziamento della collaborazione interdisciplinare e la proiezione dei risultati a livello europeo, i punti di riferimento saranno costituiti dalle infrastrutture ESFRI,⁵⁷ attive nel campo delle scienze umane (DARIAH-ERIC)⁵⁸ e del patrimonio (E-RIHS),⁵⁹ il cui coinvolgimento garantirà il pieno rispetto delle linee guida europee in materia di accessibilità, interoperabilità, riuso e creazione di valore aggiunto per la società a partire dai risultati ottenuti dalla ricerca.

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)
- [6] Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. 2021. *Definition of the CIDOC Conceptual Reference Model v7.1.1 (Version v7.1.1). The CIDOC Conceptual Reference Model Special Interest Group*. <https://doi.org/10.26225/FDZH-X261>
- [7] Toccafondi D., Cascone G. 2002. *Per la tua Margherita: lettere di una donna del '300 al marito mercante: Margherita Datini a Francesco di Marco, 1384-1401*. Comune di Prato, Archivio di Stato di Prato, Provincia di Prato.

57 ESFRI: <https://www.esfri.eu/esfri-roadmap-2021>

58 DARIAH-ERIC: <https://www.dariah.eu/>

59 E-RIHS: <http://www.e-rihs.it>