

Archival Finding Aids in Linked Open Data between description and interpretation

Francesca Tomasi^(a)

a) Università di Bologna

Contact: Francesca Tomasi, francesca.tomasi@unibo.it

Received: 21 April 2023; **Accepted:** 20 June 2023; **First Published:** 15 September 2023

ABSTRACT

The Semantic Web in general, and the LOD in particular, suppose that the knowledge conveyed by documents must be adequately modeled and represented to produce reliable and trustworthy data. Following this statement, we understand that in the archival domain the tricky and subtle transition from the traditional methodologies for data description to LOD must be delegated to agents able to skilfully read the content of cultural objects. The *Digital Hermeneutics* model aims to propose a layered architecture that allows, beyond the descriptive specificities of each domain, to formalize the data transformation from the native system to LOD. The idea is to guarantee, through context information, that each moment of the transformation workflow is documented, finally strengthening the trust of the resultant dataset.

KEYWORDS

Semantic Web; Digital Hermeneutics; Interpretation; Trustworthiness; Digital Humanities.

Gli strumenti di corredo archivistici in LOD tra descrizione e interpretazione

ABSTRACT

Web semantico e Linked Data ci hanno insegnato che la conoscenza veicolata dai documenti deve essere adeguatamente modellata per produrre dati attendibili ed affidabili. In ambito archivistico, il delicato passaggio dal sistema nativo di descrizione dei complessi documentari alla pubblicazione in LOD deve essere allora delegato a chi sappia leggere sapientemente il dato archivistico.

Il modello *Digital Hermeneutics* vuole proporre un'architettura a livelli che permetta, al di là delle specificità descrittive di ogni dominio di conoscenza, di formalizzare il passaggio dal dato dal suo sistema nativo ai LOD, garantendo, attraverso la *provenance*, che ogni momento del workflow di trasformazione sia documentato e che quindi contribuisca a rafforzare la *trust* del dataset prodotto.

Alcune sperimentazioni, condotte già su tipologie documentarie diverse, mostrano che, con l'ausilio dei *named graphs* e delle *nanopublication*, questo modello è applicabile e garantisce qualità e scientificità dei dati, citabilità di ogni azione interpretativa, esportabilità dei risultati, in una prospettiva FAIR.

PAROLE CHIAVE

Web semantico; Ermeneutica digitale; Interpretazione; Affidabilità; Informatica umanistica.

Introduzione

L'organizzazione della conoscenza, assieme alla serie di azioni, metodi e procedure che comunemente le si attribuiscono, è l'attività che meglio coniuga l'operato di archivi, biblioteche e musei. Organizzare come strategia per l'accesso alla conoscenza è infatti l'obiettivo di ogni istituzione deputata a conservare, gestire e mettere a servizio della comunità le risorse, le fonti, i beni o, diremo anche, gli oggetti che detiene. La conoscenza, che si acquisisce dal significato espresso dal contenuto, è implicita in ogni tipologia documentaria – che si manifesti nella materialità del documento, del libro, dell'artefatto o in generale della risorsa di interesse storico – e trova nella descrizione, prima di ogni altra azione, la forma più alta di espressione (Tomasi 2022, 50-61).

Descrivere dunque è l'attività necessaria per far emergere il significato, nativamente latente, di un oggetto, naturalmente portatore di un contenuto. Ma la descrizione, per quanto veicolata e mediata dai numerosi standard che le istituzioni hanno nel corso degli anni elaborato, non è mai un atto univoco e nemmeno neutrale. Descrivere significa infatti prima di tutto effettuare una scelta: scelta delle categorie della descrizione, o diremo anche degli aspetti della risorsa di cui parlare; scelta degli elementi, che si traduce nell'individuazione delle etichette che qualificano i diversi punti di osservazione della risorsa; scelta dei valori associati a quelle categorie, o meglio degli attributi delle categorie, che è possibile esprimere tanto in linguaggio naturale, quanto riuscendo sistemi di controllo di autorità o schemi di classificazione o ancora ricorrendo, in generale, all'uso di un vocabolario controllato.

Che si parli di metadati, ovvero elementi, attributi e valori associati, di ontologie, ovvero classi e proprietà e di relative basi di conoscenza o *knowledge graphs*, o ancora di Linked (Open) Data (LOD) e RDF ovvero soggetti, predicati e oggetti, il punto è sempre lo stesso. Assumere un punto di vista sugli oggetti osservati e disambiguare per risolvere l'indeterminatezza e la potenziale polisemia del linguaggio naturale, attraverso la scelta di un vocabolario per descrittori e valori che sia condiviso, controllato e accessibile, anche nell'ottica della produzione di dati FAIR (*Findable, Accessible, Interoperable, Reusable*).

Il Semantic Web, da questo punto di vista, non ha cambiato l'approccio che da sempre le istituzioni adottano per far fronte al tema della descrizione, ma ha semmai enfatizzato e reso sempre più evidente il bisogno di adottare una semantica esplicita. Con il Semantic Web si è cioè richiesto alla comunità tutta dei produttori di conoscenza di provvedere ad una rappresentazione del contenuto delle risorse che esaudisca la capacità della macchina di capire il significato della descrizione e quindi saperlo manipolare. Ed è la semantica (ovvero dati espressi attraverso l'impiego di modelli concettuali computabili) la chiave per organizzare efficacemente la conoscenza.

Nel proporre i Linked Data come strategia di pubblicazione delle descrizioni delle "cose" del Web, o anche delle entità, Tim Berners Lee ha immaginato un modello che rispettasse il bisogno di una libera definizione delle categorie descrittive, ma con un forte vincolo sulla scelta e la formalizzazione dei nomi delle cose (ovvero gli URI) e delle relazioni (*typed links*), ovvero anche delle proprietà o dei predicati di uno o più vocabolari RDF, in un sistema in cui dati e metadati non si distinguono più in modo così netto, come avveniva invece nel Web 1.0.

E la realizzazione dell'idea del Semantic Web attraverso i LOD, come nell'esempio del *knowledge graph* di Google, ci mostra come i sistemi finalizzati all'estrazione di dati strutturati da *repositories* differenti (ed eterogenei) abbiano l'obiettivo di agevolare la creazione di nuove connessioni

fra contenuti nativamente prodotti in contesti diversi, e anche per scopi diversi. L'integrazione fra informazioni è lo strumento che trasforma la fruizione di dati in fruizione di conoscenza. E la descrizione passa da azione contraddistinta, nativamente, dall'autoreferenzialità, a processo di apertura verso nuove risorse, potenzialmente distanti dal dominio di partenza.

I dati archivistici non possono esimersi dall'adottare questo modello, che costringe però a ripensare alle tradizionali logiche della descrizione del complesso documentario, per abbracciare il tema della trasversalità dei dati che possono provenire da fonti di diversa origine, provenienza e quindi soggetto produttore, ma che vogliono esaudire lo stesso bisogno conoscitivo.

Diremo che il grafo, struttura dati assunta dal Semantic Web, ha senso solo se non è puro strumento per risolvere la riconciliazione dei dati con i record d'autorità, ma se diventa il mezzo per creare relazioni globali, che escono dai confini consolidati e sicuri del fondo conservato presso un'istituzione, per abbracciare la rete delle informazioni disponibili sul *cloud*. Determinando, archivisticamente parlando, nuovi contesti e quindi nuovi contenuti.

Ma affinché questo processo sia valido, affidabile e scientifico è necessario che, da un lato, si assuma una prospettiva di lettura critica delle descrizioni archivistiche in LOD e, dall'altro, si forniscano sistemi di validazione dell'attendibilità dell'informazione così trasmessa.

La descrizione dei dati archivistici nel LOD cloud

“Attraverso il progetto di sviluppo e pubblicazione di Linked Open Data del Sistema Archivistico Nazionale l'ICAR ha conseguito una serie di obiettivi volti alla realizzazione di servizi avanzati all'utenza e alla messa a punto di strumenti e metodologie innovative finalizzati alla interoperabilità tecnologica e semantica.” Così riporta il sito <http://dati.san.beniculturali.it>, che è la piattaforma di riferimento per l'integrazione di dati archivistici di diversa origine, provenienza e sistema informativo (ovvero i sistemi aderenti).

La lettura di una qualunque scheda del SAN (Figura 1), e quindi di ogni entità LOD (Figura 2) corrispondente (persona, ente, famiglia o complesso documentario nei suoi diversi livelli gerarchici di descrizione), fa emergere una serie di questioni di interesse per le nostre riflessioni:

1. È necessario prendere coscienza che è altamente improbabile che chi ha realizzato la versione in LOD di una descrizione archivistica, che sia agente umano o macchina, corrisponda a chi fa fatto la descrizione nel sistema nativo, realizzata secondo un modello concettuale e una struttura dati probabilmente differenti da quelli di destinazione (in questo caso un aggregatore);
2. L'ontologia, che viene espressa attraverso le proprietà visibili nella scheda e nei predicati dell'entità LOD, riflette il punto di vista di chi ha modellato il dominio osservato (ICAR nel nostro caso);
3. I valori associati ad ogni campo descrittivo (ovvero ad ogni proprietà o predicato) sono espressione della ricerca condotta sulle fonti al fine della costruzione dello strumento di corredo;
4. Non sempre disponiamo di strumenti che ci facciano dire che quanto leggiamo sia l'unico punto di vista esistente sullo stesso dominio osservato o se esistano posizioni alternative, contrastanti o divergenti.

Costa, Andrea (1851 - 1910) - Persona	
SCHEDA RISORSE COLLEGATE	
Tipologia	Persona
Forma autorizzata del nome	Costa, Andrea (1851 - 1910)
	Andrea Costa
Data di esistenza	29 novembre 1851 - 19 gennaio 1910
Luogo di nascita	Imola
Luogo di morte	Imola
Descrizione	Nasce a Imola il 29 novembre 1851 da Pietro e Rosa Tozzi in una famiglia cattolica praticante e di modeste condizioni. Il giorno successivo è battezzato nella cattedrale di S. Cassiano con i nomi di Andrea, Antonio e Baldassarre e suo padrino è Orso Orsini. Frequenta le scuole elementari gestite da un sacerdote e negli anni scolastici 1866-1867 e 1867-1868 frequenta la scuola tecnica comunale con Gaetano Darchini, Luigi Sassi e Angelo Negri. Negli anni scolastici 1868-1869 e 1869-1870 frequenta il liceo come uditore per le lezioni di letteratura italiana e latina. Il 15 dicembre 1870 si iscrive alla facoltà di filosofia e belle lettere dell'Università di Bologna come "studente libero" non avendo la possibilità di pagare le regolari tasse di ammissione e per mantenersi si impiega come scrivano in un'agenzia di assicurazioni triolese. L'11 maggio, Paolo Renzi, lo associa, o almeno lo avvicina, all'Internazionale. A Imola e a Bologna compie il suo noviziato, nell'atmosfera che presto si accenderà degli entusiasmi per la Comune, e nel contatto con Carducci, che lo predilige fra i suoi allievi. Frequenta i popolani garibaldini bolognesi e con i più audaci, senza passare attraverso l'esperienza mazziniana, partecipa al Fascio operaio, che da Bologna si diffonde in Romagna, nelle Marche, in Toscana, con contenuti socialisti. Incontra Bakunin e l'anarchia. Nel 1872 tocca a lui, poco più che ventenne, commemorare l'in [Descrizione completa consultabile in IBC Archivi.]
Sistema aderente	IBC Archivi - Sistema informativo partecipato degli Archivi storici in Emilia-Romagna
URL Scheda provenienza	Vali alla scheda del sistema di provenienza

Figura 1. La scheda di un soggetto produttore (Andrea Costa) dal SAN.

dc:date	29 novembre 1851 - 19 gennaio 1910
ha_qualifiedrelations_Cof	SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66749 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66752 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66737 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66738 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66745 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66851
ha_date_esistenza	29 novembre 1851 - 19 gennaio 1910
ha_statusProvenienza	scheda_pubblicata
abstract	Nasce a Imola il 29 novembre 1851 da Pietro e Rosa Tozzi in una famiglia cattolica praticante per le lezioni di letteratura italiana e latina. Il 15 dicembre 1870 si iscrive alla facoltà di Rilo Carlo Cafiero e da parte di Costa e di altri giovani si oppone un netto rifiuto al tentativo di Engels. Costa viene eletto segretario della commissione di corrispondenza della Federazione conseguente. Rilasciato il 20 luglio 1873 è presente alla fondazione della Federazione roma rispetto ai problemi teorici, e opponendosi vivacemente alle tendenze "operistiche" preser come epicentro Bologna e vedere la partecipazione di Bakunin in persona. Ma, non adeguati arresti e persecuzioni poliziesche, provoca la paralisi del movimento internazionale: Cost vengono infine assolti. L'assoluzione di Costa e dei suoi compagni crea le condizioni per un Andreino (1877-1937) e subito dopo Costa si rifugia in Svizzera per sfuggire all'arresto. Il j dell'evoluzione politica di Costa determinante è il contatto con l'ambiente del socialismo int distacco delle organizzazioni italiane dal bakuninismo. Questi in breve, gli antecedenti esse internazionale del socialismo» e l'«Avanti!» - per far passare i propri nuovi orientamenti ne caratterizzata da una intransigenza sul terreno dei principi, ma al tempo stesso da una spn esilio tra una elezione e l'altra) oltre a una presa di posizione in favore della gestione della campagna anticolonialista che, a partire da questo momento, Costa viene sviluppando nel j di questo tentativo è, nell'agosto 1881, il convegno di Rimini delle organizzazioni a lui vicini attee e le speranze di liberazione del rivoluzionario proprio alle soglie di un'attività pratica parte dei romagnoli, una tappa importante è rappresentata dal congresso di Mantova del 1: Mantova è anche il centro di un forte movimento contadino di ispirazione socialista che due parlano della sua partecipazione alle lotte, fra i braccianti di Ravenna, i risaioli del Bolognes presenza di Costa a Roma nelle lotte contro la disoccupazione e per lo sviluppo delle organ l'attività del Partito socialista rivoluzionario di Romagna. Il limitato suffragio, non solo politi pe ecc., con la rivendicazione di una nuova legge comunale e provinciale che avesse alla p politicamente dominanti. Nell'ottobre 1889 la coalizione democratica vince le elezioni ammi appunto l'obiettivo che Costa si era prefisso e che fino al 1886 aveva perseguito con parzia dell'Internazionale in Italia, il protagonista di una svolta decisiva nella storia del movimen maniera rilevante dalle posizioni del gruppo turatiano. Non molto tempo dopo, comunque, grande importanza, membro della direzione, al congresso nazionale del 1904 egli non espr quale per primo aveva portato, 25 anni prima, la parola dei lavoratori, e che tante volte av
ha_scheda_provenienza_href	http://archivi.ibc.repubblica.emilia-romagna.it/sac/cf/IT-ER-IBC-SP00001-0000264
ha_scheda_SAN	http://san.beniculturali.it/veb/san/dettaglio-soggetto-produttore?id=66756
ha_luogoProduttore	Imola
ha_luogo_morte	Imola
ha_luogo_nascita	Imola
sac:cof:hasPlace	Imola
è_prodotto da	Archivio Andrea Costa
forma autorizzata produttore	Andrea Costa
sac:cof:authorizedForm	Andrea Costa
record_provenienza_id	IT-ER-IBC-SP00001-0000264
sistema_provenienza	IT-ER-IBC
dc:coverage	Imola
is ha_prodotto di	Archivio Andrea Costa
is ha relazione con Soggetto Produttore of	SAN:qualifiedrelationsCof_san.cat.sogP.66650_san.cat.sogP.66726 SAN:qualifiedrelationsCof_san.cat.sogP.66650_san.cat.sogP.66728 SAN:qualifiedrelationsCof_san.cat.sogP.66756_san.cat.sogP.66756 SAN:qualifiedrelationsCof_san.cat.sogP.66737_san.cat.sogP.66756 SAN:qualifiedrelationsCof_san.cat.sogP.66745_san.cat.sogP.66726 SAN:qualifiedrelationsCof_san.cat.sogP.66749_san.cat.sogP.66756 SAN:qualifiedrelationsCof_san.cat.sogP.66752_san.cat.sogP.66756 SAN:qualifiedrelationsCof_san.cat.sogP.66851_san.cat.sogP.66756

Figura 2. Una porzione di scheda LOD della stessa entità 'Andrea Costa' di Figura 1. URI http://dati.san.beniculturali.it/SAN/produttore_IT-ER-IBC_san.cat.sogP.66756.

Aggiungeremo che, il campo denominato, nella scheda del SAN, "descrizione" (che corrisponde alla proprietà 'dc:description' e anche a quella 'abstract' in LOD) è un campo testuale che contiene

molte informazioni utili che potrebbero essere organizzate in modo strutturato, ovvero che, da potenziali stringhe di caratteri espresse come *literal*, potrebbero diventare nuovi soggetti di nuove triple (ci si riferisce in particolare ai riferimenti relativi a istituzioni, persone, eventi, coordinate spazio-temporali, concetti e oggetti).

Questa serie di riflessioni ci fanno capire che più che di ‘azione finalizzata alla descrizione’ abbiamo a che fare con una serie di atti interpretativi, tanto a livello di selezione di classi e proprietà/predicati, quanto a livello di valori associati. E non dobbiamo dimenticare che, essendo la creazione dei LOD affidata ad un operatore che tipicamente, come dicevamo, non corrisponde a chi ha prodotto lo strumento di descrizione originario, è necessario che anche la provenienza degli asserti sia documentata, ovvero che siano prodotti “provenance-aware RDF statements” (Sikos e Philp 2020). Se ogni azione è interpretazione, allora ogni azione deve essere dotata di informazioni contestuali che ne rafforzino la validità.

Prendiamo il caso di un inventario analitico, massima espressione del lavoro di ricerca condotto dall’archivista. Immaginandone una pubblicazione in LOD, dovremo affrontare una serie di operazioni logiche, al di là della dimensione strettamente tecnologia e operativa del processo:

- Dovremmo provvedere a trasformare le categorie della descrizione (diciamo tipicamente gli elementi di un modello ISAD) in classi e predicati. L’azione esprime il punto di vista dell’interprete: possiamo utilizzare l’ontologia OAD (Ontology for Archival Description elaborata da Regesta)¹ o usare RiC-O (la recentissima Record in Contexts Ontology dell’ICA)² oppure possiamo decidere di creare un set di descrittori formali ad hoc (cioè elaborare un nuovo modello concettuale, come accaduto nel caso del progetto Architettura della Conoscenza - ArCo³, da allineare poi con le ontologie esistenti) o ancora riusare modelli più trasversali (ad esempio il noto CIDOC Conceptual Reference Model), modelli *general-purpose* (ad esempio DC o il più recente DC Terms), modelli che garantiscano il maggiore interscambio a livello semantico o maggiore visibilità in rete (ad esempio Schema.org adottato da numerosi motori di ricerca);
- Ogni affermazione fatta in un inventario può potenzialmente diventare un asserto esprimibile come tripla RDF. Quindi i valori associati alle categorie di cui sopra possono essere un testo non ulteriormente manipolabile (a meno dell’impiego di tecniche di Name Entity Recognition, NER o, più in generale, di Natural Language Processing, NLP) oppure un URI. E in questo ultimo caso significa che l’oggetto diventa a sua volta un soggetto aperto ad ulteriori collegamenti;
- Ogni affermazione in un inventario è tipicamente documentata attraverso il linguaggio naturale. Ovvero, ogni contenuto porta potenzialmente con sé un’attribuzione di paternità, una fonte, una data o altri elementi contestuali necessari a garantirne la credibilità. Ovvio che queste informazioni contestuali vanno formalizzate;

¹ Per una descrizione del progetto (e per accedere all’ontologia), si veda: <https://labs.regesta.com/progettoReload/oad-ontology/>.

² Ultima release del 2021: https://www.ica.org/standards/RiC/RiC-O_v0-2.html.

³ “ArCo è il Knowledge Graph del patrimonio culturale italiano, e comprende: una rete di 7 vocabolari che descrivono il dominio dei beni culturali e i dati estratti dal Catalogo Generale dei Beni Culturali dell’ICCD-MiBAC e trasformati in RDF, in accordo con le ontologie”, <http://wit.istc.cnr.it/arco>.

- Molti elementi di un inventario hanno numerosi riferimenti ad oggetti, contenuti, eventi, persone, luoghi o date che sono esterni rispetto al complesso documentario descritto. Compito dell'interprete è individuare quali di questi riferimenti sono importanti per la creazione di connessioni ad entità già presenti sul *cloud*, utili all'arricchimento dei LOD che rappresentano l'inventario in senso stretto.

Ecco che la pubblicazione in LOD mostra l'esigenza di creare nuovi strumenti di corredo, che devono fare i conti con nuove problematiche relative alla rappresentazione di contenuti culturali in ambiente digitale, spostando l'attenzione da una dimensione strettamente descrittiva, più o meno consolidata, ad una interpretativa, che va formalizzata.

Le parole chiave del Semantic Web: affidabilità, attendibilità e fiducia

Il Semantic Web pone la *trust* in cima allo *stack* che qualifica un nuovo modo di intendere i dati sul Web. Canonicamente, come ormai noto, ci riferiamo a: un sistema di codifica dei caratteri (Unicode), un sistema di identificazione univoca delle entità (URI), un linguaggio formale di markup (XML), un framework per la descrizione (RDF), dei modelli concettuali (ontologie), il ragionamento inferenziale (*logic*) e, appunto, la fiducia (*trust*), ovvero la possibilità di accedere a dati sicuri, attendibili, validi, a cui affidarsi in qualità di fonte scientifica.

Una delle possibilità per garantire la potenziale affidabilità è la documentazione della provenienza (*provenance*) degli asserti: “storing provenance information along with content data is deemed essential to prevent inconsistencies when integrating sources, to emphasize content responsibility, and eventually to foster trust in data.” (Daquino et al. 2022).

Quando un'affermazione, espressa in uno strumento di corredo analogico o ancora più in un contesto digitale distribuito come il Web, sia corredata degli elementi che ne garantiscono gli opportuni contesti, essa non diviene certa, ma sicuramente è più affidabile di un'affermazione priva di contesto. Ecco che per rendere i dati LOD una risorsa utile per la ricerca scientifica sarà necessario: dichiarare il tipo di asserzione che si sta facendo (classificazione delle tipologie), specificare chi lo sta dicendo (attribuzione di paternità), esplicitare su quali fonti si basa tale affermazione (bibliografia di riferimento), stabilire con quale grado di certezza l'informazione è espressa (dimensione di validità assoluta o relativa dell'informazione), e identificare quando tale asserzione è stata fatta (componente temporale).

L'attendibilità dei dati del Web, soprattutto quando questi dati siano esito di un processo di trasformazione da un sistema all'altro o di passaggio da un modello ad uno differente, non può essere garantita dalla sola capacità di lettura critica dell'informazione da parte del lettore, sapiente interprete del contenuto del dato, ma deve essere accompagnata da una serie di strumenti che aiutino il lettore a fidarsi di quel dato. Anche perché ciò che il lettore finale fruisce, idealmente attraverso tecnologie diverse di accesso ai dati (ovvero anche con interfacce di tipo visuale), è la rappresentazione *user-oriented* di dati scritti in RDF, ed esito delle scelte concettuali di chi ha elaborato il modello di rappresentazione di quei dati.

Ma la *trust* deve essere gestita e trattata dalla macchina, per fornire all'utente dati altrettanto affidabili e attendibili, in prima battuta perché si tratta di contenuti che provengono da “utenti di fiducia” per il Web (Ceolin et al. 2016). È dunque necessario adottare modelli di rappresentazione dei dati che permettano alla macchina di riconoscere – a seconda del livello di granularità

del contesto – il valore di un dataset o anche di una affermazione e quindi stabilire, con criteri formali, l'affidabilità dei dati da consegnare all'utente finale. Solo un sistema di *knowledge graphs* documentati, ovvero dotati di contesto, può permettere alla macchina di fidarsi di quanto il dato trasmette e quindi di restituire informazione, e soprattutto conoscenza al fruitore, che può così identificare i processi e le procedure legati alla creazione del dato.

Dal punto di vista computazionale, infatti: “Computational provenance researchers seek to develop systematic, computationally-based processes and standards for capturing, and making available for use, information about who created an object, when it was created or modified and the process or procedure that modified the object.” (Sandusky 2016, 3266).

Un sistema di interpretazione a livelli

Quando parliamo di ‘atti interpretativi’ ci riferiamo, tradizionalmente, ad affermazioni su concetti, oggetti, entità che non sono necessariamente universalmente condivise (Barabucci, Tomasi, e Vitali 2022). O, meglio, siamo consapevoli che ogni affermazione può teoricamente riflettere il punto di vista di chi l’ha espressa. Come detto in premessa, ogni componente di uno strumento di corredo (analogico o digitale) è espressione di una scelta effettuata dall’agente responsabile, sulla base, certamente, di dati oggettivi, ma anche, in potenza, di personali interessi, e non può quindi prescindere dal background culturale di quell’agente. Ovvero anche, uno strumento di corredo (l’inventario in particolare) è il risultato dell’esigenza del raggiungimento di obiettivi e scopi descrittivi, ha l’obiettivo di documentare una ricerca condotta sulle fonti, riflette conoscenze e competenze di chi esprime quelle affermazioni. Ma un’affermazione, ancora più se non univoca e non sicuramente certa, va a sua volta descritta, ovvero documentata o anche contestualizzata.

Quando si vogliono trasformare in LOD strumenti di descrizione tradizionali vanno cioè documentate tanto le affermazioni prodotte dall’agente responsabile del contenuto espresso nello strumento, quanto dell’agente responsabile del processo di conversione e quindi di pubblicazione, che ha sua volta effettua scelte di modellazione che devono essere capaci di rappresentare in modo corretto l’informazione. Nuovi soggetti produttori, diremo, con ruoli diversi, ma tutti necessari.

Per realizzare questo processo, è possibile adottare un modello astratto trasversale, svincolato da uno specifico dominio, che può documentare: l’atto di descrizione, ogni asserto rilevante, l’attribuzione di contesto e quindi le informazioni relative alla pubblicazione, utili anche per rendere citabile ogni singolo momento della procedura.

Per questo scopo si è elaborato il modello *Digital Hermeneutics* (Daquino, Pasqual, e Tomasi 2020), che adotta il sistema dei *named graphs* (Caroll et al. 2005) per creare descrizioni più complesse e articolate rispetto al meccanismo della canonica tripla, e ragiona su differenti livelli della gestione dell’informazione relativa ai dati, trattandoli individualmente, ma in forma integrata.

In particolare il modello astratto di *Digital Hermeneutics* (Figura 3) riusa nella sua componente più alta le ontologie HiCO⁴ (Daquino e Tomasi 2015) e la canonica ontologia PROV-O⁵ (Moreau et al.

⁴ Daquino, Marilena, Silvio Peroni, e Francesca Tomasi. 2020. The Historical Context Ontology (HiCO), version 2. <http://purl.org/emmedi/hico>.

⁵ Lebo, Timothy, Satya Sahoo, e Deborah McGuinness, a c. di. 2013. PROV-O: The PROV Ontology. W3C Recommendation. <https://www.w3.org/TR/prov-o/>.

2015), assieme al Nanopublication data model (Groth, Gibson, e Velterop 2010) ovvero le *nanopublications* (prefisso :np) per far dialogare i diversi grafi. Il modello si basa su 4 livelli di descrizione:

“Layer 0. Factual data that is part of scholars’ background knowledge.

Layer 1. The scope of scholars’ questionable statements.

Layer 2. Context information for hypotheses assessment.

Layer 3. Provenance information of the mining processes.”

(Daquino, Pasqual, e Tomasi 2020)

In dettaglio:

Il livello 0 fa riferimento ai valori dei metadati oggettivi o incontrovertibili o comunque sui quali non si vuole esercitare alcun atto interpretativo per la natura canonicamente descrittiva di alcuni campi di ogni scheda. Potrebbero essere i dati già presenti in un repertorio, in un catalogo, dati prodotti da enti o istituzioni sui quali non c’è dibattito o non c’è interesse a documentarlo. Gli elementi scelti per la descrizione delle diverse caratteristiche dell’oggetto o della collezione dipendono dalle scelte dell’istituzione o dell’agente che ha generato la descrizione nel sistema nativo e possono essere eventualmente rappresentati, in un sistema LOD, con le ontologie più opportune rispetto al dominio (ma capaci di esaudire le esigenze descrittive attraverso la scelta del più adeguato modello concettuale stabilito dall’agente responsabile).

Il livello 1 è quello relativo ai dati potenzialmente contestabili, sui quali ci può essere pluralità di interpretazione, dibattito o divergenza di punti di vista. Possiamo anche dire che, in generale, si tratta dei dati che qualificano l’intervento critico dello studioso, o in generale dell’agente responsabile del contenuto, quelle affermazioni cioè che richiedono una serie di elementi di contesto per essere considerate valide e affidabili. Questo livello può documentare anche ulteriori informazioni relative all’oggetto descritto al livello 0, allo scopo di arricchire la descrizione stessa con dati che provengono da fonti esterne. E, anche in questo caso, i predicati per la gestione del collegamento fra la fonte descritta e le risorse esterne è delegata all’agente responsabile, sempre in un sistema di presa di conoscenza del rapporto modello di descrizione nativa/modello concettuale scelto.

Il livello 2 è la provenienza dell’asserto. Ovvero è la serie delle informazioni necessarie a trasformare l’asserto di livello 1 in una risorsa di cui, potenzialmente, fidarsi. Ecco che il contesto può essere determinato da: dichiarazione di paternità dell’affermazione (attribuzione di responsabilità intellettuale all’agente che ha prodotto l’asserto), menzione delle fonti di riferimento, tipologia di atto interpretativo (secondo una tassonomia da stabilire, e.g. filologico, storico, prosopografico, artistico, culturale, etc.) e relativo criterio adottato (metodo specifico per ogni tipologia tassonomica e quindi dipendente dalla disciplina di riferimento). In questo caso i predicati da utilizzare provengono dalle già citate ontologie PROV-O e HiCO.

Il livello 3 è quello che rende esplicito chi ha estratto e trasformato queste informazioni, quindi menziona il responsabile della procedura di generazione dello statement RDF e stabilisce le informazioni correlate al processo (ad esempio la data di pubblicazione dei dati e l’agente che ha effettuato l’operazione, descrivendoli sempre con PROV-O).

Le *nanopublications* infine fanno dialogare in un nuovo *named graph* i 3 livelli: l’asserto (“statement”), la provenienza delle informazioni di contesto (“context”) e i dati di pubblicazione (la “provenance” dell’operazione di creazione dello statement in RDF).

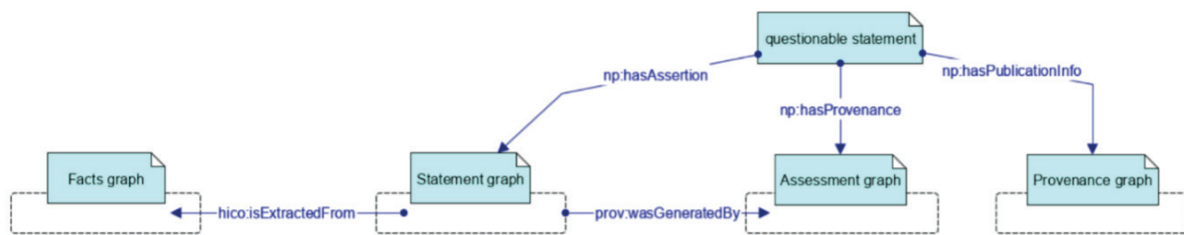


Figura 3. L'approccio a livelli di *Digital Hermeneutics* e il ruolo delle *nanopublications* (Daquino, Pasqual, e Tomasi 2020).

Abbiamo altrove già pubblicato esempi di applicazione di questo modello (si veda in particolare Daquino, Giovannetti, e Tomasi 2019; Daquino, Pasqual, e Tomasi 2020; Pasqual e Tomasi 2022). Come sempre, per testare infatti la bontà di un modello è il caso di applicarlo ad una base di conoscenza. Questa azione consente infatti di capire fino a che punto il modello è valido ed esportabile, potenzialmente anche in contesti differenti (tanto sul piano disciplinare, quanto dei supporti documentari o dei domini di riferimento).

Un progetto di ricerca che ha adottato il modello a scopo di sperimentazione è il progetto MythLOD⁶, un'iniziativa nata per gestire la conversione di un database relazionale (il progetto Mythologiae⁷) in un dataset LOD (Pasqual e Tomasi 2022). Nella creazione di questo nuovo strumento, l'atto interpretativo dello studioso ("assertion" di livello 1) è stato formalizzato rispetto a due elementi identificati nella scheda del progetto nativo: 1) l'attribuzione, da parte di un agente, di un soggetto che rappresenta il contenuto di un'immagine visuale (tipicamente un quadro, una statua o in genere un'opera d'arte) e che stabilisce la categoria concettuale di appartenenza dell'oggetto stesso; 2) il riconoscimento del passo letterario classico, moderno o contemporaneo, o anche di una riscrittura letteraria, capace di raccontare quanto riprodotto nell'opera d'arte. La fonte classica è richiamata attraverso un sistema di citazione canonico utilizzando il modello HuCit⁸ e collegato alla Perseus digital library grazie al sistema urn:cts⁹ per l'arricchimento e la fruizione diretta della fonte; il riferimento all'opera medievale, moderna e contemporanea riporta una tradizionale menzione bibliografica a livello di *work*, secondo il modello FRBR (autore, titolo), e la riconciliazione ai sistemi di controllo di autorità è esercitata con VIAF.

L'asserzione di livello 1 è arricchita di informazioni contestuali a livello 2 per garantire l'attendibilità dell'informazione stessa, con gli opportuni riferimenti alla provenienza dell'informazione riportata a livello 1: il tipo di interpretazione, il criterio adottato e l'agente, ovvero il responsabile dell'interpretazione.

Si legga il seguente esempio relativo all'oggetto identificato univocamente con ID item/312 nel catalogo¹⁰. In Figura 4 sono riportati i dati descrittivi tradizionali e reputati oggettivi. In Figura 5 è

⁶ Il progetto, comprensivo di data model, dataset, query, strumenti di visualizzazione e storytelling, a cura di Valentina Pasqual, si può consultare all'indirizzo <https://dharc-org.github.io/mythlod/static/mima.html>.

⁷ <https://mythologiae.unibo.it/>.

⁸ <https://github.com/mromanello/hucit>.

⁹ <https://sites.tufts.edu/perseusupdates/beta-features/perseus-cts-api/>.

¹⁰ L'intero catalogo, sempre a cura di Valentina Pasqual, si può navigare all'indirizzo: <https://dharc-org.github.io/mythlod/catalogue/index.html>.

attribuita la categoria di appartenenza per il contenuto espresso dall'oggetto ("Enea fugge da Troia in fiamme"), si riporta la citazione specifica da Virgilio, Eneide II, 721-724 e si menzionano le altre riscritture o citazioni dello stesso fenomeno nelle letterature moderne e contemporanee (Dante, Petrarca, Leopardi e Ungaretti). In più si forniscono gli strumenti per l'*assessment* degli asserti: il tipo di interpretazione, il criterio adottato e l'agente responsabile. In Figura 6 si dettaglia il riferimento alla citazione canonica, riportando il link al *work* su Perseus e i link ai sistemi di controllo d'autorità su VIAF (nella figura riportata si mostra il solo collegamento al *work* su VIAF per Dante).

Con le *nanopublications* (np:), si possono mettere in dialogo i tre grafi ("assertion", "provenance" e "pubInfo") espressi in MythLOD (myth:) relativi all'oggetto (si riporta qui di seguito il *named graph* in notazione Trig):

```
myth:head312 {
  myth:np-312 a np:Nanopublication ;
  np:hasAssertion myth:assertion312 ;
  np:hasProvenance myth:provenance312 ;
  np:hasPublicationInfo myth:pubInfo312 .
}
```



Factual Data

Title: Enea fugge da Troia in fiamme con il padre Anchise sulle spalle e con il figlioletto Ascanio (dettaglio)

Author: Chedel, Quentin-Pierre, 1705-1762

Keywords: fuga enea chedel anchise guerra-di-troia incendio ascanio eneide

Typology: Disegno

Collocation: Metropolitan Museum of Art, New York (United States of America)

Periodi: Arte moderna, XVIII secolo

Description Sorpreso nella città dall'attacco dei Greci, Enea fugge in mezzo alle fiamme con il vecchio padre An... Read More

See Also: [Responsible entity website](#)

Assertion

Categories: Enea fugge da Troia in fiamme

Canonical Citations: Eneide, II, 721-724

General References: Virgil. | Aeneis

Dante, Alighieri (1265-1321) Divina commedia

Leopardi, Giacomo, 1798-1837. | Canti

Petrarca, Francesco, 1304-1374. | Trionfi

Ungaretti, Giuseppe | Vita di un uomo

Provenance

Interpretation Type: Iconographical Approach

Interpretation Criterion: Associazione di Fonti

Interpretation Performer: Pitas, Angelica

Figura 4. Le informazioni fattuali nella scheda di catalogo per l'oggetto 312.

Figura 5. Le asserzioni e la provenienza degli asserti per lo stesso oggetto descritto in Figura 4.

Assertion

Categories: Enea fugge da Troia in fiamme

Canonical Citations:

[Go to the Perseus resource of this reference](#)

General References:

[Go to VIAF resource of this reference](#) ^{VI}_{AF}

Reference Type: Fonte Classica
Reference Author: Virgil. ^{VI}_{AF}

[Go to VIAF resource of this reference](#) ^{VI}_{AF}

Reference Type: Fonte Medievale o Moderna
Reference Author: Dante, Alighieri, 1265-1321 ^{VI}_{AF}

Figura 6. Dettaglio di Figura 5, con sistemi citazionali espliciti, arricchimento della descrizione e collegamenti ai sistemi di controllo d'autorità.

Conclusioni

Il Semantic Web in generale, e LOD in particolare, ci hanno ormai fatto capire che la conoscenza veicolata dai documenti deve essere adeguatamente modellata per produrre dati attendibili ed affidabili.

Questo significa che in campo archivistico il delicato passaggio dal sistema nativo di descrizione dei complessi documentari alla pubblicazione in LOD deve essere delegata a chi sappia leggere sapientemente il contenuto degli oggetti culturali.

A questo si aggiunge l'esigenza di una modellazione dei dati che non solo garantisca che le ontologie scelte siano sufficientemente capaci ad esprimere il potere informativo dei dati, ma che provveda a fornire tutti gli elementi di contesto necessari ad assicurare che i dati siano davvero utili a scopi di ricerca scientifica.

Il modello *Digital Hermeneutics* vuole proporre un'architettura a livelli che permetta, al di là delle specificità descrittive di ogni dominio di conoscenza, di formalizzare il passaggio dal dato dal suo sistema nativo ai LOD, garantendo, attraverso le informazioni di contesto, che ogni momento del workflow di trasformazione sia documentato e che quindi contribuisca a rafforzare la *trust* del dataset prodotto.

Alcune sperimentazioni, condotte già su tipologie documentarie diverse, mostrano che, con l'ausilio dei *named graphs* e delle *nanopublications*, questo modello è applicabile e garantisce qualità e scientificità dei dati, citabilità di ogni azione interpretativa, esportabilità dei risultati, in una prospettiva FAIR.

Se chi produce strumenti di corredo nel Semantic Web deciderà di passare da una dimensione strettamente descrittiva ad una interpretativa, cioè si impegnerà a fornire i necessari contesti, allora davvero si potrà garantire scientificità all'intero processo di produzione dei dati. Si potrà così immaginare la creazione di nuove forme di accesso alla conoscenza trasmessa dagli oggetti digitali, ovvero dai dati che ne raccontano il contenuto e quindi il significato. E alla creazione, potranno aggiungersi nuove forme di organizzazione della conoscenza così prodotta. E allora, nuovo sarà anche il sapere che potrà emergere dall'ibridazione fra risorse prodotte in ambiti diversi, realizzate per scopi diversi e magari disseminate con altrettanto diverse finalità.

Riferimenti bibliografici

- Barabucci, Gioele, Francesca Tomasi, e Fabio Vitali. 2022. "Modeling data complexity in public history and cultural heritage." In *Handbook of Digital Public History*, 459-74. Oldenbourg: De Gruyter.
- Carroll, Jeremy J., Christian Bizer, Pat Hayes, e Patrick Stickler. 2005. "Named graphs, provenance and trust." In *Proceedings of the 14th International Conference on World Wide Web*, 613-22. New York: ACM. <https://doi.org/10.1145/1060745.1060835>.
- Ceolin, Davide, Paul Groth, Valentina Maccatrozzo, Wan Fokkink, Willem Robert Van Hage, e Archana Nottamkandath. 2016. "Combining user reputation and provenance analysis for trust assessment." *Journal of Data and Information Quality (JDIQ)* 7 (1-2):1-28. <https://doi.org/10.1145/2818382>.
- Daquino, Marilena, e Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects." In *Metadata and Semantics Research*, a cura di Emmanouel Garoufallou, Richard J. Hartley, e Panorea Gaitanou, 544:424-36. Cham: Springer International Publishing.
- Daquino, Marilena, Francesca Giovannetti, e Francesca Tomasi. 2019. "Linked data ed edizioni scientifiche digitali. Esperimenti di trasformazione di un Quaderno di appunti." *Umanistica Digitale* 3(7). <https://doi.org/10.6092/issn.2532-8816/9091>.
- Daquino, Marilena, Valentina Pasqual, e Francesca Tomasi. 2020. "Knowledge representation of digital hermeneutics of archival and literary sources." *JLIS.it* 11 (3):59-76. <https://doi.org/10.4403/jlis.it-12642>.
- Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, e Francesca Tomasi. 2022. "CLEF. A Linked Open Data native system for Crowdsourcing." <https://arxiv.org/abs/2206.08259>.
- Groth, Paul, Andrew Gibson, e Jan Velterop. 2010. "The Anatomy of a Nanopublication." *Information Services & Use* 30 (1-2):51-56. <https://doi.org/10.3233/ISU-2010-0613>.
- Moreau, Luc, Paul Groth, James Cheney, Timothy Lebo, e Simon Miles. 2015. "The Rationale of PROV." *Web Semantics: Science, Services and Agents on the World Wide Web* 35:235-57. <https://doi.org/10.1016/j.websem.2015.04.001>.
- Pasqual, Valentina e Francesca Tomasi. 2022. "Linked Open Data per la valorizzazione di collezioni culturali: il dataset mythLOD." *AIB Studi* 62 (1):149-68. <https://doi.org/10.2426/aibstudi-13301>.
- Sandusky, Robert J. 2016. "Computational provenance: DataONE and implications for cultural heritage institutions." In *2016 IEEE International Conference on Big Data*, 3266-71. <https://doi.org/10.1109/BigData.2016.7840984>.
- Sikos, Leslie F., e Dean Philp. 2020. "Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs." *Data Science and Engineering* 5 (3):293-316. <https://doi.org/10.1007/s41019-020-00118-0>.
- Tomasi, Francesca. 2022. *Organizzare la conoscenza. Digital Humanities e Web Semantico*. Milano: Editrice Bibliografica.