JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare

Francesca Tomasi, Marilena Daquino

Introduzione

Interessanti sono i traguardi raggiunti nella modellazione concettuale per il dominio archivistico. L'uso di tecnologie legate al Semantic Web, in particolare, ha contribuito a ridefinire modi e sistemi di valorizzazione del potere informativo espresso dalle descrizioni archivistiche. Progetti come le ontologie del progetto ReLoad[1] e recentemente la pubblicazione di SAN Ontology[2], ci portano ad aprire altre - se non nuove, certamente riproposte in nuovi termini - considerazioni sulla natura stessa dei dati archivistici e sul loro riuso nell'ottica di un dialogo con altre fonti di dati, al fine di estendere le capacità semantiche ed espressive del patrimonio informativo che gli archivi custodiscono.

[1] Archivio Centrale dello Stato, Istituto dei Beni Culturali Regione Emilia Romagna, Regesta.exe, *ReLoad, repository for linked open archival data*, http://labs.regesta.com/progettoReload/.

[2] ICAR, Centro MAAS, *SAN ontology*, http://www.maas.ccr.it/SAN-LOD/lode/.

JLIS.it Vol. 6, n. 3 (September 2015)

DOI: 10.4403/jlis.it-11133

Argomentare nuovamente sulla natura della descrizione archivistica riporta alle tanto discusse definizioni e affiliazioni delle discipline archivistiche, biblioteconomiche, museologiche nell'alveo dei 'sistemi di organizzazione della conoscenza' (Gnoli, Marino e Rosati 2006), dove le finalità e gli obiettivi informativi connessi alla natura documentale si uniscono alle questioni di rappresentazione formale delle informazioni, le quali ad oggi paiono confluire sempre più, o quantomeno in modo congiunto, nel dibattito sul 'web di dati', sollecitato dalla prospettiva Linked Data.

La crescita di interesse verso la formalizzazione di thesauri, vocabolari e ontologie, per domini di conoscenza chiaramente delineabili attorno ad una comunità scientifica o disciplinare, è sintomo dell'auspicabile e continuo desiderio di arricchimento del valore potenziale del patrimonio culturale; l'ausilio di queste tecnologie permette infatti di sperimentare nuovi percorsi per la condivisione e la valorizzazione delle collezioni di dati, costringendo gli attori del processo a riformulare le scelte di organizzazione dei dati stessi alla luce di rinnovate esigenze comunicative.

Per la comunità archivistica, questo non è che il proseguimento della strada verso "l'uscita dei sistemi archivistici da loro stessi", per citare un'affermazione di Stefano Vitali (Vitali 2003), che necessariamente richiama all'ordine del giorno la definizione del ruolo che gli archivi e le descrizioni archivistiche rivestono nei confronti della crescita e dell'evoluzione di altre discipline, le quali dipendono, o possono appoggiarsi, al contributo che l'archivista, e l'archivistica, forniscono nei processi di organizzazione della conoscenza.

Assumendo come indiscusso il valore intrinseco delle descrizioni archivistiche quali oggetti aventi dignità di fonte storica, resta aperta la discussione su come queste descrizioni debbano essere riutilizzate in altri ambiti e contesti e su come possano essere arricchite con informazioni provenienti da altri domini. La descrizione archivistica è capace di fornire un apporto informativo non solo finalizzato all'interpretazione da parte di soggetti terzi, ma già autoesplicativo e

contestualizzato. Andranno allora studiate le opportune strategie finalizzate ad integrare campi di conoscenza eterogenei e offrire un ancor più alto livello qualitativo dell'informazione, fornendo al contempo i necessari mezzi per sfruttare queste potenzialità nella fruizione.

Se la concettualizzazione in campo archivistico rappresenta un modello di eccellenza cui le altre discipline guardano, approcci alla conoscenza assunti in altri contesti teorici favoriscono una ridefinizione di scopi e ragioni dei 'sistemi di descrizione archivistica'.

Le fasi del processo

Acquisita questa prospettiva, il processo che conduce alla sistematizzazione delle esigenze espressive di ambiti disciplinari dialoganti passa inevitabilmente per alcune fasi:

- la comprensione dei diversi approcci all'organizzazione dell'informazione, attività necessaria a valutare adeguate soluzioni di integrazione. Nel caso specifico, che si vuole in questa sede discutere, vengono considerate quali fonti di dati semanticamente eterogenee le descrizioni archivistiche e le informazioni estratte dal *full text* di una fonte, che nativamente vengono strutturate - vale a dire descritte e organizzate - sulla base di diverse esigenze di rappresentazione, conservazione e interrogazione;
- la scelta di un dominio di conoscenza, quello filologico, avente le necessarie affinità con quello archivistico, con il fine di un'integrazione dei dati che sia orientata all'arricchimento vicendevole dei contenuti ed i cui benefici, in termini di fruizione, siano evidenti anche ad altri attori e comunità, quali ad esempio gli storici;
- la modellazione mediante un approccio *top-down*, con componenti ibride, delle entità che fungono da *trade union* tra i

domini e che possono considerarsi di interesse in fase di acquisizione di conoscenza dai dati. In particolare, oggetto di questa analisi interdisciplinare è la formalizzazione ontologica dei ruoli delle persone, o meglio degli agenti (in senso estensivo, oltre alle persone, anche enti e famiglie). Tre sono i livelli del processo interpretativo che, in una dimensione evento-centrica, muovono dal concetto di relazione:

- la relazione tra agente e ruolo/funzione ricoperto, così come sarebbe previsto dallo standard ISDF dell'ICA (CPBS 2007). In questo contesto, ruoli e funzioni attribuibili ad un agente (sia esso soggetto produttore o meno) vengono considerati, pur nella consapevolezza della diversità concettuale, sullo stesso piano semantico, essendo – da un punto di vista formale – entrambe possibili relazioni tra entità di diversa natura (e.g. il ruolo esercitato da una persona nell'atto della stesura di una lettera non è dissimile - sul piano formale - dalla funzione amministrativa espletata dal soggetto nella produzione di un documento);
- tra agente e fonte, analizzando le diverse tipologie di relazioni dell'entità nei confronti del documento - compresa la paternità - e ampliando di conseguenza le possibilità relazionali offerte dallo standard EAC-CPF3. In questa prospettiva, si vuole anche estendere l'applicazione dello standard oltre il solo soggetto produttore, considerando entità diversamente relazionate ad una fonte;
- tra agente e informazioni acquisibili dalle fonti, in un'ottica che tenga conto della provenienza delle asserzioni, come ad esempio la formalizzazione dell'atto interpretativo esercitato da un editore su un documento. Questo procedimento deve, di conseguenza, tenere conto anche della possibile

[3] StaatsBibliothek zu Berlin, Encoded Archival Context – Corporate Bodies, persons, families (EAC-CPF), http://eac.staatsbibliothek-berlin.de.

contraddittorietà degli asserti, poiché le informazioni estratte dai documenti rappresentano un atto di lettura soggettivo.

L'importanza di disporre del testo pieno delle fonti, l'analisi di approcci diversi ma potenzialmente integrati, la riflessione sul concetto di ruolo e relazione - che travalica il concetto di soggetto produttore in senso stretto e che sposa l'idea di persona, ente o famiglia dotata di un ruolo o di una funzione - ed infine l'importanza della paternità delle asserzioni, rappresentano l'intervento critico che si vuole qui descrivere. Scopo di questo approccio è di favorire un effettivo processo di *cultural heritage enhancement* come esito dell'integrazione disciplinare, che significa anche integrazione di sistemi di metadatazione o concettualizzazione eterogenei (Peroni, Tomasi e Vitali 2013). L'ontologia PRoles (cfr. sezione relativa) rappresenta il tentativo di formalizzare il dominio, acquisita questa prospettiva teorica. Prima di arrivare a PRoles sarà necessario però introdurre alcuni concetti necessari a chiarire l'approccio.

L'organizzazione dell'informazione. Metodi e strategie

Discipline diverse optano per altrettanto differenti modalità di organizzazione dei dati in relazione sia ai criteri di conservazione (e preservazione) del patrimonio, sia agli obiettivi informativi e comunicativi che si intendono raggiungere. In particolare diremo che i processi di organizzazione della conoscenza insegnano l'approccio semantico all'informazione come strategia di esaustività rappresentazionale[4].

[4] Innegabile che i sistemi di organizzazione della conoscenza prospettino la cooperazione attiva di àmbiti disciplinari eterogeni. Per una visione d'insieme cfr. le attività di ISKO (*International Society for Knowledege Organization*) Italia, http://www.iskoi.org e le pagine di ISKO internazionale, http://www.isko.org.

L'assunto che emerge dal confronto tra metodologie prospetta risvolti ambivalenti: le comunità che hanno adottato un approccio documento-centrico per la strutturazione informazioni - come per esempio la filologia nel caso della realizzazione di edizioni digitali - hanno la costante necessità di saper coniugare questa scelta con l'approccio di altre discipline, che hanno invece previsto un'organizzazione prettamente data-centrica, come i dati bibliografici e archivistici; allo stesso modo, le comunità che hanno scelto quest'ultimo approccio nell'organizzazione dei dati, nel tentativo di concettualizzare un modello coerente e complessivo per l'intero dominio, risultano comunque lontane dall'uscita dall'autoreferenzialità dei propri dati.

La descrizione archivistica, inserendosi in questo secondo filone, ha posto solide basi e prassi per l'organizzazione dell'informazione, trovandosi spesso ad essere il primo attore a lavorare sull'informazione stessa (i *raw data*), al fine di esplicitarne la struttura logica. L'elaborazione degli standard per le descrizioni archivistiche[5] hanno proseguito il percorso verso l'ambìto obiettivo di integrare i propri dati semanticamente eterogenei all'interno di un unico modello concettuale. Questo *step*, che può dirsi parzialmente raggiunto con l'elaborazione di SAN Ontology, pone all'ordine del giorno nuove prospettive per l'effettiva possibilità di creare punti di contatto e intersezione tra discipline e le rispettive fonti di dati.

L'approccio data-centrico, che tipicamente riflette il modello logico del database, fa proprio il principio del documento inteso come set di metadati descrittivi, laddove lo scopo è esprimere aspetti estrinseci tipici della descrizione formale ad un livello alto di astrazione. L'approccio documento-centrico individua tipicamente nel *markup* a livello di *full text*, e quindi sulla costruzione di dati

[5] Per un riferimento degli standard in uso per la descrizione archivistica cfr. International Council on Archives (ICA), http://www.ica.org/10206/standards/standards-list.html.

semi-strutturati, il modello da seguire in una prospettiva che parte invece dal contenuto veicolato dal documento piuttosto che dai suoi aspetti estrinseci, o diremo anche, paratestuali.

Ora, solo un approccio combinato può essere capace di un'esaustività interpretativa che soddisfi appieno i bisogni dell'utente finale. L'utente, consultando dati aggregati da ambiti disciplinari affini benché differenti per natura e metodologie, potrà ottenere cioè una migliore esperienza informativa; in particolare l'utente sarà in grado di utilizzare un numero significativo di *access points* ai dati – siano essi a livello di sovrastruttura, come i dati archivistici, o di natura contenutistica e interpretativa, come i dati storico-filologici. Considerando, nel solco della tradizione archivistica, le entità (soggetti produttori e non) quali *access point*s principali nella ricerca e interrogarsi sulle ramificazioni di questo concetto – quindi sui ruoli che un'entità può avere e le relazioni che questa intrattiene con altre entità, nel merito di quel ruolo – è un punto di partenza necessario e un utile strumento finale per la ricerca lato utente. Crediamo che in quest'ottica di dialogo e scambio tra comunità, l'archivista, o meglio l'archivista informatico, sia l'attore culturale in grado di porsi in prima persona a concettualizzare strategie per l'integrazione di dati provenienti da domini diversi. Questa figura dovrebbe perciò porsi all'interno del dibattito interdisciplinare portando in dote la sua tradizione e la sua capacità di elaborazione, spronando il dibattito verso nuovi scenari collaborativi e di sperimentazione, proseguendo nel suo ruolo di organizzatore e valorizzatore della conoscenza, rispondendo ai bisogni delle comunità scientifiche che sono interessate ai documenti d'archivio quali fonti storiche.

Dal punto di vista formale, i modelli ontologici offrono un notevole potenziale espressivo per definire i termini e l'organizzazione di un dominio di conoscenza, essendo obiettivo della concettualizzazione, storicamente dichiarato, la capacità di riorganizzare e condividere i dati di ambiti differenti (Chandrasekaran, Josephson e Benjamins

1999). Perciò le ontologie possono rivelarsi, nel contesto più ampio del *cultural heritage management*, una metodologia strategica.

Approcci all'ontology-based data integration

Nell'ottica di anticipare alcune considerazioni preliminari al dibattito sulle ontologie (Biagetti 2010) per l'archivistica (e 'oltre l'archivistica'), è necessaria una breve disamina dei principali approcci per l'integrazione di dati semanticamente eterogenei mediante modelli ontologici (Cruz e Xiao 2005). Le scelte possibili ricadono su tre soluzioni di modellizzazione, dipendenti dal contesto di applicazione:

- un'unica ontologia per descrivere il dominio di conoscenza. Questo approccio consiste nella creazione di un unico vocabolario che consente di mantenere un alto livello di precisione terminologica (identificando univocamente tutti gli elementi che contribuiscono a definire il dominio), con il limite evidente di non poter essere immediatamente intellegibile da altri domini di conoscenza per il riuso immediato dei dati e delle definizioni;
- molteplici ontologie. Questo approccio prevede l'utilizzo flessibile di un'ontologia per ogni singola fonte di dati e/o ambito del dominio di conoscenza e la successiva parziale integrazione di queste attraverso processi di *mapping*, *matching* o *merging* (Kalfogloui e Schorlemmer 2003);
- un approccio ibrido. Questa scelta comporta la convivenza di più ontologie per la descrizione di sotto-domini congiuntamente ad un vocabolario *top-level* per la definizione dei termini di riferimento del dominio.

Senza poter esprimere un netto giudizio sull'adozione di una scelta di modellazione piuttosto che un'altra, possiamo dire che quest'ultimo approccio può considerarsi il giusto medium per l'intento interdisciplinare, ovvero il riuso delle informazioni formalizzate in un dominio, nel contesto di altri domini di conoscenza; ogni dominio può adottare, parzialmente o integralmente, il vocabolario di altri domini per raccordare le definizioni e/o le singole ontologie allo scopo di estrarre e combinare le informazioni.

Similmente, il ruolo di un'ontologia per il dominio archivistico dovrebbe essere quello di fornire uno strumento che sia al contempo esaustivo e riusabile dalle altre discipline, che guardano ai modelli dell'archivistica col fine di arricchire i dati delle loro collezioni con quelli provenienti da una fonte autorevole, in modo da completare il quadro descrittivo ed il contesto di riferimento.

In questo senso, poter disporre di un modello che consenta di relazionare *authorities*, entità archivistiche e unità documentali mediante l'attestazione di un ruolo svolto e.g., da una persona su un documento - e/o che questo sia asserito nel testo pieno di un documento - creerebbe un ponte tra fonti di dati eterogenee.

Dove finisce l'archivistica

Nell'analisi degli elementi che intervengono nella descrizione di un complesso archivistico, gli aspetti formali per l'integrazione tra *data sources* si uniscono ad alcune lacune del dibattito teorico. Come già teorizzato altrove (Daquino, Peroni, Tomasi e Vitali 2014), la mancata formalizzazione in Italia dello standard per la descrizione delle funzioni dei soggetti produttori, ISDF, crea un gap tra l'effettiva necessità di rappresentare questo aspetto fondamentale della descrizione archivistica e la possibile correlazione tra domini differenti.

Guardando alla mole di informazioni veicolate dalla descrizione archivistica si scorgono numerosi possibili punti di raccordo con le altre discipline. Ma se le descrizioni archivistiche sono autoesplicative, autonome a livello informativo e identificabili univocamente, rasentano, proprio per questo, l'ottica dei 'silos di dati', ovvero danno vita ad un dominio di conoscenza caratterizzato da un'autosufficienza informativa che non richiede esplicitamente integrazioni provenienti da altri domini, come, per esempio, *la LAM* (*Library, Archives and Museums*) *activity*[6] invece prescriverebbe (Zorich, Waibel e Erway 2008).

Se aggiungessimo il fattore trascurato, ovvero la descrizione delle funzioni o, nel nostro caso, dei ruoli svolti dai soggetti (produttori e non) nei confronti della documentazione che hanno prodotto o a cui sono genericamente relazionati, l'autoreferenzialità dei complessi informativi si aprirebbe all'intervento e al dialogo con altre forme di concettualizzazione e di relazionalità tra gli elementi del dominio. La già elaborata EAC-CPF ontology[7] ha aperto la discussione verso la formalizzazione del concetto di relazione in campo strettamente archivistico. Ma queste relazioni, che non devono necessariamente fermarsi all'interno del dominio archivistico, potrebbero aprirsi alle informazioni provenienti da altre fonti di dati e confrontarsi con altri approcci all'organizzazione dell'informazione. È questo un campo di studi e sperimentazione che può dare interessanti frutti, sia a livello strutturale, proseguendo l'interazione e lo scambio tra attori culturali, sia a livello contenutistico, quindi modificando le finalità con cui un progetto di ricerca si pone nei confronti del proprio oggetto di studio.

---

[6] Si veda, per esempio: "An investigation into the incentives and strategies for deep and transformative collaboration among libraries, archives and museums (or LAMs)" fra le attività di OCLC, http://oclc.org/research/activities/lamsurvey.html. E soprattutto, sempre fra le attività di OCLC: http://hangingtogether.org/?cat=5. Nel panorama italiano si auspica che analoghe iniziative possano trovare spazio nelle attività del MAB (Musei, Archivi, Biblioteche), http://www.mab-italia.org/.

[7] *EAC-CPF Ontology*, http://labs.regesta.com/progettoReload/lontologia-eac-cpf.

L'approccio quindi descrittivo dell'archivistica si potrebbe aprire verso una prospettiva di analisi delle fonti che trova in alcune discipline, come appunto la filologia, presupposti teorici, metodologici e tecnici consolidati. L'auspicio è di formalizzare un *environment* che si possa qualificare come un: "complete and flexible system of archival description that would interrelate record description, creator description and the description of functions and activities" (Pitti in Gartner 2014). Potremmo anche dire che un approccio data-centrico come quello dell'archivista potrebbe trovare nel documento, vale a dire nella fonte storica come veicolo di contenuti, una risorsa utile ad arricchire la descrizione degli oggetti di interesse del dominio.

Dove iniziano le altre discipline

Tipicamente le discipline che lavorano sul documento, allo scopo di formalizzarne il contenuto, mirano a enucleare concetti attraverso un processo di annotazione sul *full text* che ha lo scopo di descrivere elementi pertinenti. Tali elementi possono prendere la forma, per esempio, delle persone, dei luoghi, delle date. Come discusso in letteratura la marcatura è carente in espressività, non solo per l'assenza di una semantica dei linguaggi di *markup* (Renear, Dubin e Sperberg-McQueen 2002), ma soprattutto per la mancanza di relazioni fra gli elementi dell'annotazione. Il concetto di persona nei testi letterari è fondamentalmente correlato alla funzione che la persona svolge in quello specifico contesto. I diversi ruoli che diverse persone ricoprono legano il concetto stesso di individuo al documento che attesta l'entità (dalla persona citata in un documento alla persona identificabile come il curatore dell'edizione digitale). In un dato tempo e in un dato luogo una funzione viene svolta da un individuo, dando vita ad un evento. Il vocabolario della TEI[8], i cui

[8] *Text Encoding Initiative* (TEI), http://www.tei-c.org.

studi sul modello ontologico sono ancora in fase sperimentale[9], potrebbe trarre allora giovamento dalla concettualizzazione già avviata in campo archivistico. Allo stesso modo alcune componenti che identificano l'orientamento di TEI all'approccio documento-centrico arricchirebbero le capacità espressive delle descrizioni archivistiche.

Nell'ottica allora di ragionare e sperimentare queste forme di integrazione tra fonti di dati in ambiti affini, alcune considerazioni vanno fatte proprio sulla natura e sui contenuti che possono essere utili ai fini di un arricchimento vicendevole delle collezioni di dati.

È assodato il contributo archivistico alla definizione di norme per l'*authority control*, come è principio consolidato che questo non si limita alla mera disambiguazione di una stringa che identifica il nome del soggetto produttore. Quel che si vuole far emergere è la necessità di investigare, con il fine di formalizzare, le tipologie di relazione tra quelli che possono essere considerati senza dubbio gli *access point* principali nella ricerca - ovvero le persone, le famiglie e le istituzioni - e altre informazioni che sono utili allo stesso fine, ovvero l'identificazione univoca di un soggetto produttore tramite l'esplicitazione del contesto di produzione del complesso archivistico.

Per fare ciò, crediamo non possano non intervenire fattori che fino ad ora sono stati trascurati o considerati secondari rispetto alle finalità descrittive degli *authority file* (o *reference file*), quali appunto la descrizione dei ruoli e delle funzioni che questi soggetti svolgono nei confronti della documentazione e, viceversa, che la documentazione asserisce o attesta su essi, attraverso l'estrazione di informazioni provenienti dal testo pieno delle fonti.

Esplicitare questi collegamenti, che sono al confine tra il dominio archivistico e quello prettamente filologico (ma con risvolti negli

[9] Special Interest Group (SIG) on *Ontologies*, http://www.tei-c.org/SIG/Ontologies.

studi storici, sociali, politici, filosofici), può diventare allora un terreno proficuo di modellizzazione concettuale, che sappia riusare il contributo delle ontologie per l'archivistica nell'ottica di formulare nuove relazioni e creare nuova informazione.

I modelli per l'integrazione di domini di conoscenza

Molti e diversi domini di conoscenza si sono posti l'obiettivo di concettualizzare i punti di raccordo tra le discipline. Scopo dell'integrazione semantica è quello di agevolare il dialogo e l'interscambio in un'ottica di condivisione di descrizioni che arricchisca il potere informativo del (meta)dato. Tra i modelli fino ad oggi elaborati, potremmo ad esempio menzionare modelli generali come EDM[10] e la Prov-o ontology[11] o ancora il CIDOC-CRM[12] e la sua versione FRBRoo;[13] più specificamente nell'ambito storico potremmo ricordare la Factoid ontology (Pasin e Bradley 2013), gli SNAP:DRGN[14] e, in ambito bibliografico, le SPAR ontologies.[15]

Questi (ma anche altri) modelli hanno in comune, nelle reciproche differenze di approccio e soggetto di modellazione, l'intento di fornire una concettualizzazione non solo o non espressamente esaustiva delle componenti di un dominio di conoscenza, ma la creazione di una rete di legami tra entità a livello più alto (Doerr 2003), che possono travalicare i confini degli àmbiti disciplinari e

[10] *Europeana Data Model (EDM) Documentation,* http://pro.europeana.eu/edm-documentation.

[11] *PROV-O: The PROV Ontology*. 30 April 2013, W3C Recommendation, http://www.w3.org/TR/2013/REC-prov-o-20130430.

[12] CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/index.html

[13] *FRBRoo Introduction*, http://www.cidoc-crm.org/frbr_inro.html.

[14] SNAP:DRGN, Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names, http://snapdrgn.net.

[15] The *Semantic Publishing And Referencing* ontologies (SPAR), http://purl.org/spar.

permettere il dialogo e lo scambio di informazioni nel contesto del *cultural heritage*.

Sulla base delle indicazioni e degli spunti offerti da queste tipologie di modelli, si possono sperimentare innumerevoli possibilità di modellazione e integrazione con l'obiettivo implicito, come già accennato, di arricchire i contenuti di una collezione di dati, ad esempio congiungendo dati della descrizione archivistica a quelli provenienti dalla marcatura di un testo in XML/TEI, integrando così un livello critico (trascrizione, organizzazione, stratificazione dell'informazione estrapolata da una fonte) e un livello gerarchico e relazionale dell'informazione. Allo stesso modo EAC-CPF, come formalizzazione dello standard ISAAR-CPF, pur nella problematicità dello schema (Michetti 2008), ha avviato una prima formalizzazione della descrizione separata del soggetto produttore - e del concetto di contesto sotteso - che consente un dibattito critico stimolante e agevola proposte di concettualizzazione espressive.

Un modello formale: l'ontologia PRoles

I tentativi di coniugare i desiderata della marcatura filologica in un'edizione digitale con quelli della descrizione archivistica possono dunque partire proprio dalla concettualizzazione di un modello che, con un approccio *top-down*, indichi le relazioni e gli agenti coinvolti in entrambi i domini.

In questa direzione, un esperimento è stato fatto con PRoles (Daquino, Peroni, Tomasi e Vitali 2014),[16] un'ontologia per la definizione dei ruoli che agenti possono svolgere nei confronti di un testo e che un testo (pieno) asserisce sugli agenti stessi, dando una precipua importanza all'autore (e autorevolezza) di chi formalizza le asserzioni estratte dal testo. Particolare attenzione è stata data ai ruoli e alle relazioni politiche, per iniziare un ragionamento che uscisse dalla sola formalizzazione delle relazioni di paternità agente-

[16] *Political Roles Ontology (PRoles)*, http://www.essepuntato.it/2013/10/politicalroles.

documento e intraprendesse l'analisi delle varie sfaccettature del contenuto informativo di un documento, approfondendo la multilivellarità delle relazioni tra i soggetti dello studio.

L'esperimento condotto attraverso la realizzazione dell'ontologia PRoles vuole proporre una concettualizzazione di dominio esattamente nell'ottica dell'integrazione disciplinare.

Questo modello nasce come un primo tentativo di inserire in EAC-CPF Ontology (Mazzini e Ricci 2011) la descrizione dei ruoli (non tanto, attualmente, le funzioni amministrative) che le entità archivistiche svolgono nei confronti dei documenti correlati e nei confronti di altre entità, estendendo poi le possibilità relazionali, ovvero creando uno scenario per lo sfruttamento di informazioni estrapolare dal testo pieno delle fonti.

In un'ottica evento-centrica, diversi scenari vengono creati per descrivere la partecipazione di più entità eterogenee (agenti, documenti, eventi in senso stretto, luoghi, archi temporali, ruoli) in un àmbito in cui vengono esercitate e operate 'azioni' e 'asserzioni'.

Per esemplificare: è possibile esplicitare con predicati ontologici il ruolo che un agente (persona, famiglia o ente) riveste nei confronti di un documento, ricalcando i presupposti di ISDF e ampliando la descrizione offerta dalle relazioni di EAC-CPF, al fine di definire il rapporto di paternità culturale dell'agente sull'oggetto o una qualsiasi altra relazione fra essi. È possibile esplicitare altresì un ruolo, un'azione o un evento in cui è coinvolto un agente, asserendo la fonte da cui questa informazione è stata estrapolata e l'autore dell'asserzione stessa. È necessario poi consentire la convivenza di asserzioni contraddittorie, effettuate da più autori (dei metadati) e/o su più fonti che riportano informazioni contraddittorie.

Questi ultimi fattori, l'attestazione della provenienza di un'asserzione e l'asserzione stessa estrapolata dal *full text* di un documento, entrambe mediate dalla formalizzazione di un ruolo detenuto da un agente, diventano i *matching points* tra i due domini

di conoscenza, aprendo ad alcune interessanti possibilità informative in fase di interrogazione dei dati: infatti, è possibile ricercare non solo le informazioni su un agente e le fonti che esplicitano informazioni su un agente, ma anche le fonti manipolate (marcate) da un editore, quindi la sua interpretazione sui fatti che descrive, e confrontare le interpretazioni di più editori.

Per definire le tipologie dei ruoli e inquadrarli nello spazio, nel tempo e in un contesto d'azione, sono state istanziate due apposite classi, *proles:PoliticalRoleInTime* e *proles:ParticipationWithPoliticalRole*, che riusano ed estendono due modelli preesistenti, PRO Ontology (Peroni, Shotton e Vitali 2012)[17], parte del set delle SPAR ontologies, per la definizione dei ruoli ed il pattern Nary Participation[18] per gli scenari più complessi.

La scelta terminologica dei ruoli ricade nell'adozione di una tassonomia aperta (ovvero ampliabile aggiungendo individui alla classe *pro:Role*), che può essere estesa a seconda della tipologia di scenario che si intende descrivere: attualmente il modello prevede due tipologie di ruoli, legati al ciclo di vita del documento (dalla sua scrittura alla pubblicazione) e legati alle relazioni politiche descritte nel testo. A questo livello di formalizzazione, un utente finale può:

- risalire ai documenti in cui un'entità è citata;
- identificare il ruolo con il quale l'entità è stata citata;
- riconoscere quale ruolo l'entità ricopre nei confronti del documento;
- stabilire l'editore che attesta l'esistenza di un dato ruolo o relazione come propria interpretazione sul testo.

Questa metodologia rende il modello facilmente estensibile in base al caso d'uso, poiché permette di definire nuovi ruoli e scenari senza dover modificare la parte terminologica dell'ontologia (TBox).

[17] PRO, the Publishing Roles Ontology, http://purl.org/spar/pro.
[18] Nary Participation,
http://ontologydesignpatterns.org/wiki/Submissions:Nary_Participation.

Infine, per le attestazioni di *provenance* sulle asserzioni fatte in merito al contenuto di un testo, vengono riusate alcune proprietà della già citata PROV-o, che consentono di assicurare, per ogni informazione estratta dal *full text*, il legame con la fonte testuale e l'autore dell'affermazione.

Gli elementi marcati di un documento possono diventare istanze dell'ontologia e consentire ai documenti di arrivare ad un grado di espressività che il solo *markup* non è in grado di garantire. Solo attraverso questo approccio i documenti possono trasformarsi in vere e proprie basi di conoscenza.

Un caso di studio

Per testare le potenzialità del modello, una prima sperimentazione è stata fatta sull'edizione delle lettere di Vespasiano da Bisticci (Tomasi, *Vespasiano*, 2013). Si tratta della trascrizione in formato XML/TEI delle missive ricevute ed inviate dal/al copista fiorentino Vespasiano da Bisticci, vissuto nell'arco del XV secolo. Tali lettere sono tràdite da fonti eterogenee (documenti d'archivio, codici miscellanei e di dedica, esemplari a stampa moderni) e sono state variamente edite nel corso degli anni (dall'800 fino ai giorni nostri). L'edizione digitale sulla quale è stata fatta la sperimentazione dell'ontologia, riporta: la trascrizione delle lettere in una *facies* stabilita dall'editore, la segnatura dell'esemplare (e quando possibile l'accesso alla versione digitale del codice o del documento), l'estrazione dei metadati basilari per l'identificazione dell'oggetto lettera (mittente, destinatario, data cronica e data topica), l'indicazione dei precedenti editori, ed infine un commento storico, prosopografico e lessicografico.

Tutti i nomi di persona menzionati nelle lettere sono stati indicizzati, normalizzati rispetto alle specifiche di VIAF e mappati sui principali repertori esistenti (LCCN, SBN, BNF, etc.) per stabilire le forme controllate dei nomi e quindi i punti di accesso e connessione. Sono

stati anche stabiliti i primi ruoli (e.g. copista, committente, miniatore, regnante, combattente, etc.) utili a classificare le persone ed identificarle rispetto al contesto in cui sono menzionate (il documento inteso come supporto di trasmissione) e all'evento a cui sono associate (la copia di un codice, il coinvolgimento in un episodio storico, la richiesta di un esemplare manoscritto, la partecipazione ad un evento politico, etc.). Sono stati acquisiti i link a Dbpedia (cfr. una visualizzazione dei collegamenti della voce 'Vespasiano da Bisticci', Fig. 1), in particolare per destinatari e mittenti dotati di una voce su Wikipedia. Tutti i riferimenti alle persone, ai codici prodotti dalla scuola di Vespasiano e al lessico tecnico della copia e della circolazione di codici manoscritti nel corso del XV secolo sono stati marcati, identificati in modo univoco attraverso URI ed espressi come asserti RDF (Tomasi, *L'edizione*, 2013).

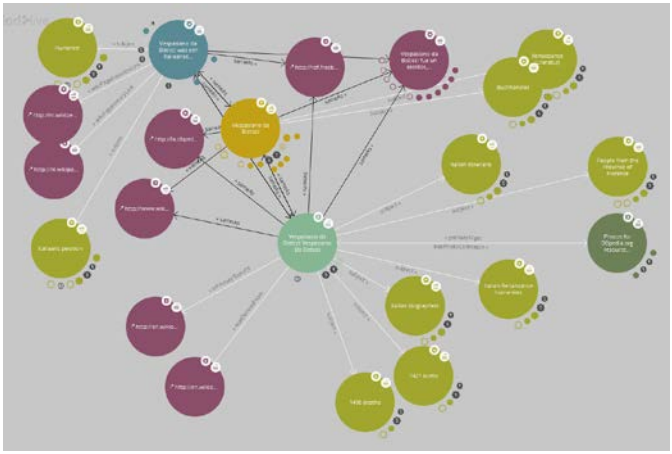

Figura 1. Vespasiano da Bisticci visualizzato attraverso il sistema di browsing grafico del 'web of data', Lod Live[19].

[19] http://lodlive.it

Non è questa la sede per descrivere integralmente le specifiche dell'edizione, ma sarà utile un caso per esemplificare il processo. Consideriamo la lettera diciassette della collezione[20] (cfr. Fig 2.). Si tratta della carta ASF, Mediceo avanti il Principato, filza XVII, n. 165, inviata da Vespasiano da Bisticci a Piero de' Medici, il 19 aprile 1458 da Firenze. Scopo della missiva è informare uno fra i più noti committenti del copista sullo stato di avanzamento di alcuni codici che il Medici stesso ha commissionato a Vespasiano, affidando al contempo alla scuola il reperimento dell'antigrafo, dei materiali scrittorei, la realizzazione della copia e delle miniature.

In prima battuta andrà notato come Vespasiano assuma in questa lettera, come in ogni lettera della collezione, un ruolo ben preciso, a seconda dell'evento in cui viene inquadrato. Il suo essere soggetto delle carte di cui detiene autorialità si somma alla varianza del ruolo che in ogni lettera viene a ricoprire. A seconda delle relazioni che instaura con il suo interlocutore, il profilo dell'entità, secondo le specifiche EAC-CPF, arricchisce la descrizione con nuove connessioni semantiche: fra persone (Vespasiano quale interlocutore e produttore di codici per Piero de' Medici), fra persona e carta (Vespasiano e la carta 165, filza XVII, Mediceo avanti il Principato, ASF in cui Vespasiano è classificabile come mittente), fra persona e altre risorse correlate a quella persona (il codice Plinio, *Storia naturale*, attuale manoscritto Laur. Plut. 82,3 prodotto dalla scuola di Vespasiano).

Ma si può entrare ancor più nel dettaglio. All'interno della missiva di parla di un tal Benedetto, riconosciuto variamente dagli editori e probabilmente identificabile in Benedetto Strozzi. Sappiamo dalle fonti che Benedetto lavorò come copista per conto di Vespasiano da Bisticci. Nello specifico Benedetto è copista del Laur. Plut. 82, 3 sopra menzionato. Questa entità, in qualità di agente che riveste un ruolo all'interno di un dato contesto – la lettera diciassette della collezione

[20] http://vespasianodabisticciletters.unibo.it/lettere/lettera17.html

– ed in qualità di produttore di manoscritti copiati per conto di Vespasiano, diventa un punto di accesso, ma anche di raccordo fra i domini. Un utente che voglia quindi interrogare la collezione delle lettere ed estrapolare in quale unità documentaria il copista viene citato, in particolare con questo ruolo - ruolo peraltro non esplicitamente dichiarato nella lettera -, potrebbe voler sapere anche su quali fonti si è basato l'editore per dedurre questa asserzione o conoscere come altri editori hanno descritto la stessa persona, in quali altri contesti questa persona compare e quali altri ruoli questa persona altrove assume.

Queste sono alcune delle riflessioni che hanno costituito le basi per il proseguimento del lavoro che andiamo a descrivere nella sezione conclusiva che segue.

Figura 2. Vespasiano da Bisticci a Piero de' Medici, Firenze, 19 aprile 1458[21]

Prospettive di lavoro future

L'elaborazione di un modello che si ponga al di sopra, e insieme sia ponte, tra i domini di conoscenza deve saper coniugare le esigenze di obiettività e riusabilità dei predicati, tipiche delle ontologie fondazionali, con le esigenze specifiche situazionali.

Nel caso di uno studio sulle relazioni tra il dominio archivistico e quello filologico, i problemi che si pongono coinvolgono allora la definizione del 'contesto' dell'unità documentaria che si vuole descrivere: possiamo avere interesse nella formalizzazione di

[21] http://vespasianodabisticciletters.unibo.it/lettere/lettera17.html.

predicati che descrivano il contesto di produzione del documento, per il quale il riuso delle informazioni di dominio archivistico e bibliografico è il più indicato, o il contesto storico in cui si inserisce il documento, con tutte le specificità circostanziali (geografiche, lessicali, sociali, politiche) che ne derivano.

Su queste basi, il lavoro svolto nel concepire PRoles è stato ridefinito e rielaborato in un modello più articolato, e in fase di ulteriore definizione, denominato HiCO[22], che sarà di nuovo testato sulla già menzionata edizione digitale delle lettere di Vespasiano da Bisticci (Tomasi, *Vespasiano*, 2013). Questa ontologia (Daquino e Tomasi 2015), seguendo le già citate buone pratiche di riuso di modelli provenienti da domini di conoscenza diversi benché contigui, ambisce a definire un workflow per l'attribuzione e la descrizione di interpretazioni in senso lato storiche - ovvero che tendono a storicizzare il documento - di cui un editore fornisce una descrizione data-centrica dei contenuti e delle relazioni intrattenute dai soggetti-agenti, che compaiono nel testo pieno del documento. Entità archivistiche, testi ed interpretazioni, congiuntamente, tendono in questa prospettiva a porsi come punti di partenza per potenziali nuove relazioni, e non esclusivamente oggetti il cui fine ultimo è la descrizione in un dominio di conoscenza.

Saper fornire una valida formalizzazione a queste necessità conoscitive è un obiettivo che si pone all'ordine del giorno per le discipline umanistiche, o meglio, per le *digital humanities*, che vagliano costantemente le potenzialità offerte dalle tecnologie del Semantic Web per poter arricchire gli strumenti per l'apprendimento e la ricerca.

L'interdisciplinarità così declinata si propone quale mezzo per creare nuovi strumenti di lavoro, definire nuove finalità di ricerca e, in definitiva, creare vere e proprie nuove fonti di sapere. È in questa prospettiva che si vuole continuare a ragionare e produrre

[22] *Historical Context Ontology* (*HiCO),* http://hico.sourceforge.net/index.html.

documentazione, linee guida e modelli concettuali, auspicando una sempre maggiore complementarità negli indirizzi di ricerca all'interno degli ambiti umanistici, facendo fronte alle difficoltà di formalizzazione di domini così vasti utilizzando la tradizione e le competenze di chi di mestiere si occupa di organizzare la conoscenza.

Diremo quindi che l'unità documentaria come fonte d'informazione testuale consistente, l'integrazione di domini apparentemente difformi e certamente eterogenei e infine le persone, intese non solo come soggetti produttori di documentazione archivistica, ma come agenti che svolgono ruoli nel contesto del concetto di evento, rappresentano insieme un quadro teorico che modella un approccio al *cultural heritage* come nuovo dominio di conoscenza integrata.

Promuovendo l'adozione degli standard archivistici e riusando modelli e teorizzazioni provenienti da altri domini si potranno così davvero sfruttare le possibilità offerte dai Linked Data per l'integrazione e la creazione di nuovi percorsi informativi tra i dati, allo scopo di ampliare il potenziale comunicativo dei documenti d'archivio e garantirne gli adeguati 'strumenti di corredo' (Linked Archival Metadata 2014).

References

Biagetti, Maria Teresa (a cura di). "Le ontologie". *AIDAinformazioni*: 28 (2010). Accessed April 15, 2014, http://www.aidainformazioni.it/2010/122010.html .

Chandrasekaran, B., Josephson John. R., and Benjamins V. Richard. "What are ontologies? And why do we need them?". *IEEE Intelligent Systems* 01-14 (1999): 20-26.

CPBS Sub-Committee on Descriptive Standards, *International Standard for Describing Functions* (ISDF), 2007, http://www.ica.org/10208/standards/isdf-international-standard-for-describing-functions.html. Trad. it. by Salvatore Vassallo, 2009, http://media.regesta.com/dm_0/ANAI/anaiCMS//ANAI/000/0111/ANAI.000.0111.0005.pdf .

Cruz, Isabelle, and Xiao Huiyong. "The Role of Ontologies in Data Integration". *Journal of Engineering Intelligent Systems* 13-4 (2005): 1-18.

Daquino, Marilena, Peroni Silvio, Tomasi Francesca, and Vitali Fabio. "Political Roles Ontology (PRoles): enhancing archival authority records through Semantic Web technologies". *Procedia Computer Science* 38 (2014): 60-67.

Daquino, Marilena, and Tomasi, Francesca. "Ontological approaches to information description and extraction in the cultural heritage domain". In *Humanities and Their Methods in the Digital Ecosystem*. Proceedings of the Third AIUCD Annual Conference (AIUCD2014). Selected papers, edited by Francesca Tomasi, Roberto Rosselli Del Turco, and Anna Maria Tammaro, article 8. New York: ACM, 2015.

Doerr, Martin. "The CIDOC CRM - An ontological approach to semantic interoperability of metadata". *AI Magazine* 24 (2003). doi:10.1609/aimag.v24i3.1720.

Gartner, Richard. "An XML schema for enhancing the semantic interoperability of archival description". *Archival Science* 15-3 (2014). doi:10.1007/s10502-014-9225-1.

Gnoli, Claudio, Marino Vittorio, and Rosati Luca. Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il Web. Milano: Tecniche nuove, 2006.

Kalfogloui, Yannis, and Schorlemmer Marco. "Ontology mapping: the state of the art". *The Knowledge Engineering Review* 18-01 (2003): 1-31. doi:10.1017/S0269888903000651.

Linked Archival Metadata: A Guidebook. Eric Lease Morgan and LiAM, Version 0.99, April 23, 2014, http://sites.tufts.edu/liam/.

Mazzini, Silvia, and Ricci, Francesca. "EAC-CPF Ontology and Linked Archival Data". In *Proceedings of the 1st International Workshop on Semantic Digital Archives*, Berlin 29/9/2011 (SDA 2011), 72-81, http://ceur-ws.org/Vol-801/paper6.pdf.

Michetti, Giovanni. "EAC: Elementi per un Approccio Critico". *Archivi & Computer* 18 (2008): 40-55.

Pasin, M. and Bradley J. "Factoid-based prosopography and computer ontologies: towards an integrated approach". *Literary and Linguistic Computing* (2013). doi: 10.1093/llc/fqt037.

Peroni, S., Shotton D., and Vitali F. "Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents". In *Proceedings of the 8th International Conference on Semantic Systems*, edited by Harald Sack, and Tassilo Pellegrini. New York: ACM, 2012, http://speroni.web.cs.unibo.it/ publications/peroni-2012-scholarly-publishing-linked.pdf.

Peroni, S., Tomasi F., Vitali F. "The aggregation of heterogeneous metadata in Web-based cultural heritage collections. A case study". *International Journal of Web Engineering and Technology* 8 (2013): 412-432.

Renear, Allen, Dubin, David, and Sperberg-McQueen, C. Michael. "Towards a semantics for XML markup". In *Proceedings of the*

2002 ACM symposium on Document engineering (DocEng '02), 119-126. New York: ACM, 2002.

Tomasi, Francesca. *Vespasiano da Bisticci, Lettere*. Bologna: AlmaDL - CRRMM - Università di Bologna, 2013, http://vespasianodabisticciletters.unibo.it. doi:10.6092/unibo/vespasianodabisticciletters.

Tomasi, Francesca. "L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere". *ECDOTICA* 2013, 9 (2012): 264-286.

Vitali, Stefano. "La seconda edizione di ISAAR (CPF) e il controllo d'autorità nei sistemi di descrizione archivistica". In *Authority Control: definizione ed esperienze internazionali*, a cura di Mauro Guerrini, and Barbara B. Tillet, 6. Firenze: University Press, 2003.

Zorich, Diane M., Günter Waibel, and Ricky Erway. *Beyond the silos of the LAMs: Collaboration among libraries, archives and museums*. Report produced by OCLC Research (2008), http://www.oclc.org/content/dam/research/publications/library/2008/2008-05.pdf.

FRANCESCA TOMASI. Alma Mater Studiorum, Università di Bologna. francesca.tomasi@unibo.it.

MARILENA DAQUINO. Alma Mater Studiorum, Università di Bologna. marilena.daquino2@unibo.it .

Tomasi F., Marilena Daquino. "Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare". *JLIS.it*. Vol. 6, n. 3 (September 2015): Art: 11133. DOI: 10.4403/jlis.it-11133.

ABSTRACT: This paper presents a reflection on some topics related to the semantic modeling of cultural heritage description. In particular, the authors move from some ontologies as developed in - and for - the archival domain. The purpose of this approach is double: to understand, from the one hand, the role of the archival conceptual metodology for the cultural heritage enhancement; from the other, to propose a model to let heterogeneous disciplines able to dialogue in a shared semantic perspective. We adopt a triple level vision: 1. the importance of the documentary unit as a primary full text source, 2. the possibility to integrate models from potentially different research environments and domain, 3. the relevance of agents' roles and functions as an exploratory approach to the meaning of documents. In particular we reflect on the concept of 'creator' - the agent - as a key to manage multiple relationships (between people and between people and resources) in a provenance-oriented perspective. We finally discuss about an ontology that formally describes our vision: PRoles (Political Roles Ontology).

KEYWORDS: Data; Documents; EAC-CPF; Proles; Roles; TEI.

Submitted: 2014-04-15

Accepted: 2014-05-18

Published: 2015–09-15