Umanistica Digitale - ISSN:2532-8816 - n.7, 2019
M. Daquino, F. Giovannetti, F. Tomasi – Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini
DOI: http://doi.org/10.6092/issn.2532-8816/9091

Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini

Marilena Daquino,[1] Francesca Giovannetti,[2] Francesca Tomasi[3]

Digital Humanities Advanced Research Centre (/DH.arc), University of Bologna
[1]marilena.daquino2@unibo.it
[2]francesc.giovannett6@unibo.it
[3]francesca.tomasi@unibo.it

Abstract

Digital Scholarly Editions (DSE) are a powerful tool for disseminating cultural heritage. As long as Semantic Web technologies become a de facto standard for disseminating cultural heritage data, DSE are a missing bit that must be integrated in the LOD Cloud. Despite a number of standards are in place for exchanging data about encoded texts (XML/TEI, noSQL database, RESTful API), a number of key elements are missing, namely: (1) a comprehensive workflow for publishing DSE as knowledge graphs, (2) data models for identifying concepts and relationships characterising DSE, and (3) a cost-benefit analysis of the usage of such technologies. In this paper we present the proof-of-concept DSE of Paolo Bufalini's notebook so as to address, discuss, and evaluate the aforementioned issues.

Le edizioni scientifiche digitali (ESD) sono uno strumento necessario alla divulgazione del patrimonio culturale. Con l'affermarsi delle tecnologie del *Semantic Web* per la disseminazione dei dati dei beni culturali, le ESD diventano un tassello fondamentale da integrare nella Linked Open Data cloud. Linee guida per la codifica del testo e standard tecnologici vengono correntemente utilizzati nella comunità accademica per condividere i informazioni tra ricercatori e sviluppatori (p.e. XML/TEI, noSQL database, RESTful API). Nondimeno, esistono diverse lacune nella creazione di edizioni scientifiche semantiche: (1) l'assenza di un workflow per la pubblicazione di ESD come *knowledge graph*, (2) l'identificazione di modelli concettuali per la rappresentazione di aspetti caratterizzanti le ESD, (3) un'analisi dei costi, dei vantaggi e dei benefici derivanti dall'utilizzo di *cutting-edge technology* in giustapposizione alla "tradizionale" codifica del testo. In questo articolo presentiamo il *proof-of-concept* dell'edizione semantica digitale del quaderno di appunti di Paolo Bufalini che tenta di rispondere a tali lacune.

Keywords: TEI, Linked open data, Ontologies, Digital scholarly editions, Semantic publishing

Introduzione[1]

Il tema delle edizioni scientifiche digitali (ESD) è oggi al centro di un dibattito critico tanto vivace quanto prolifico (si vedano almeno [14], [18], [27], [31]). Tale dibattito ha animato riflessioni e contributi critici su forme, modelli, processi e metodi di quella speciale forma di edizione scientifica di testi che trova nel digitale non solo un supporto di trasmissione, ma una rinnovata metodologia di espressione dell'atto interpretativo. In tempi recenti sono state le tecnologie del Web semantico ad aver ricoperto un ruolo determinante nel ripensare la testualità digitale. L'affermazione di nuovi modelli di descrizione delle 'cose' del Web non è stata acquisita dalla comunità scientifica come un mero espediente tecnologico, ma ha disvelato la necessità di ripensare alla rappresentazione del sapere attraverso la valorizzazione del sistema di relazioni implicite ed esplicite che il testo veicola. Ecco che il principio del grafo della conoscenza, o anche "knowledge graph", qualifica un nuovo bisogno: passare dall'informazione alla conoscenza attraverso un sistema di interconnessioni che usano il grafo come struttura di dati. Ma il passaggio da un modello tipico del Web 1.0 a quello richiesto dal Web 3.0 non è operazione indolore. E tale processo merita un'attenzione che va oltre gli aspetti strettamente tecnologici, per abbracciare l'analisi critica dell'intero percorso che dall'edizione 1.0 porta all'edizione 3.0.

Molte delle ESD ad oggi disponibili online presentano una struttura documento-centrica. Tale struttura prevede che i testi siano fruiti, in primo luogo, in quanto documenti, ovvero come una collezione di unità informative che si articolano in uno spazio bidimensionale. Si tratta del paradigma tipico della stampa tipografica, trasportato dall'ambiente analogico a quello digitale. L'espansione del Web of data sollecita la ridefinizione del concetto di ESD da parte della comunità accademica. L'ESD, oggi strutturata essenzialmente come rete di documenti, deve essere ripensata come insieme di entità, ovvero anche risorse identificate univocamente con URI e interconnesse attraverso l'uso di link tipizzati (le proprietà RDF), secondo un paradigma dato-centrico.

Tale trasformazione pone quesiti metodologici fondamentali. Si tratta infatti di un passaggio di paradigma, non solo di una migrazione tecnologica, che impone di tradurre il documento da entità semi-strutturata in collezione di dati altamente strutturati. Tale operazione implica un passaggio di modellazione che su scala molto più ampia avviene nel Web. Il documento (o la collezione di documenti) che costituisce l'edizione, diviene soggetto di un processo di decostruzione, che mira ad identificare atomi informativi univocamente identificati. Ma tale processo costringe anche tali elementi atomici ad essere estratti, arricchiti, organizzati, manipolati, interrogati al di fuori della 'camicia di forza' del documento all'interno del quale occupano una precisa posizione e che concorrono a costituire ([15]).

Nell'ottica dato-centrica, anche la collezione di documenti e il singolo documento possono essere considerati atomi informativi. Anche ad essi può essere associata un'etichetta univoca e

1 Le autrici hanno disegnato e sviluppato il progetto, nonchè partecipato alla stesura dell'articolo, che è stato così suddiviso: le Sezioni 1 e 2 sono responsabilità di Francesca Tomasi; le Sezioni 3 e 7 sono responsabilità di Marilena Daquino; le Sezioni 5 e 6 sono responsabilità di Francesca Giovannetti. La responsabilità è condivisa per la Sezione 4 e le conclusioni.

possono dunque essere considerati tanto come entità unitarie quanto come collezioni di entità. A partire dalle unità informative così definite si può generare una rete di relazioni fra i dati facenti parte dell'edizione, presenti nel Web o provenienti dal mondo reale e a cui non per forza è associata una presenza online (per esempio persone, luoghi, opere, eventi, periodi storici ecc.). Ed è in questo modo che il knowledge graph prende forma.

L'insieme di dati strutturati estratti dalla canonica edizione XML/TEI determina un livello di informazione che convive, aumenta ed è in comunicazione con lo strato di annotazioni incorporate nel documento. Si viene così a creare una combinazione di stand-off ed embedded markup che concorrono alla descrizione dello stesso testo secondo due modelli differenti (documento-centrico e dato-centrico).

L'arricchimento dell'edizione scientifica digitale attraverso l'estrazione di un dataset a partire dal documento XML/TEI ha il potenziale di ridefinire il paradigma delle edizioni accademiche: basi di dati interconnesse e non più silos isolati. L'ESD deve essere considerata un bene culturale a pieno titolo ed un tassello necessario al completamento del network delle risorse del patrimonio culturale, la cui presenza online sotto forma di linked open data è in rapidissima espansione.

Proviamo allora a sintetizzare come il principio di applicazione del modello LOD alle canoniche edizioni XML/TEI possa essere concretamente immaginato, poiché il workflow richiederà poi un'attenta valutazione critica.

Il processo di transizione dell'edizione digitale da XML/TEI (modello ad albero) a LOD (modello a grafo) impone un'attenta analisi del testo e dell'informazione in esso contenuta iniziando dal modello di markup XML/TEI messo a punto dall'editore. Si tratta di un'operazione concettuale ancor prima che tecnica, che implica una visione del testo 'dall'alto', non come sequenza di dati ma come rete di relazioni.

Potenzialmente, ogni elemento del markup funge da soggetto, oggetto o proprietà (ossia un typed link) collegato al documento di origine (anch'esso entità parte del dataset).

La definizione della rete di relazioni che attraverso gli asserti (soggetto-predicato-oggetto) è possibile costruire richiede un'analisi altrettanto dettagliata del documento TEI di partenza. La scelta dei predicati ontologici adeguati a dare reale consistenza semantica al markup deve essere consapevolmente intrapresa con l'obiettivo di aumentare l'espressività della base di conoscenza contenuta nel testo, evitando di disperdere il suo potenziale informativo in nome della, pur necessaria, semplificazione. La scelta dei più adatti modelli ontologici costituisce l'aspetto centrale di questa fase.

Si prosegue poi con la costruzione del grafo della conoscenza, un insieme di triple RDF estratte dal documento TEI. I dati in essa contenuti sono *linked* quando sono in grado di dialogare con il sapere disseminato nel web, in grado di arricchire il testo digitale e di espandere l'esperienza informativa dell'utente finale che può muoversi oltre i confini del testo. Si arriva dunque ad avere la collezione di documenti sufficiente per la pubblicazione dell'edizione online.

Ci sono dunque una serie di interrogativi da porsi durante il passaggio dal modello XML/TEI

al modello LOD. La progettazione si baserà su una serie di riflessioni metodologiche fondamentali:

- *Passaggio dal markup embedded al dataset*: come trasformare XML/TEI in RDF? Non si tratta di un semplice esercizio meccanico, ma di una presa di coscienza della funzione dei sistemi di marcatura inline. Sarà allora necessario distinguere le entità reali dalle loro occorrenze all'interno del testo (ossia le stringhe di caratteri) e preservare i fenomeni registrati nel testo per la sua visualizzazione e fruizione (ad es. aspetti grafici, impaginazione).
- *Passaggio dagli elementi TEI all'ontologia*: quali modelli ontologici riutilizzare? E' necessario creare nuove classi e proprietà? Si deve invece 'ontologizzare' lo Schema? Difatti, i modelli concettuali scelti per il mapping su TEI sono determinanti per stabilire le capacità comunicative della base di conoscenza e svelare concetti latenti.
- *Creazione dei collegamenti con i sistemi di controllo delle autorità*: quali collegamenti istituire? Come dare consistenza alla rete del sapere? La necessità di aprire i dati al dialogo con altre risorse in grado di arricchire la base di conoscenza, così da soddisfare i bisogni informativi di varie tipologie di utenti, è alla base della scelta dei dataset relazionati e dei modelli.

Questi interrogativi invitano a elaborare una strategia di valorizzazione delle edizioni scientifiche digitali come patrimonio culturale. Il presente studio si propone di esplorare la relazione fra testo, codifica TEI e rappresentazione ontologica, attraverso il caso studio del taccuino inedito di Paolo Bufalini. Il lavoro condotto sul quaderno ci aiuterà ad affrontare la dimensione epistemica del processo, tentando di illustrare un processo largamente scalabile, che porti alla pubblicazione di un'edizione digitale semantica. In particolare nella sezione 2 si intende illustrare la storia del quaderno di Bufalini, introducendo alcune delle problematiche sottese; la sezione 3 vuole fornire gli strumenti necessari a illustrare il workflow per la costruzione di un'edizione digitale semantica; la sezione 4 stabilisce i modelli concettuali necessari al processo di conversione; la sezione 5 illustra le modalità di arricchimento dell'edizione XML/TEI in vista della sua trasformazione; la sezione 6 spiega il processo di conversione a HTML e RDF; la sezione 7 descrive i principi che stanno a fondamento della realizzazione dell'edizione digitale accessibile attraverso un'applicazione Web.

Il quaderno di Paolo Bufalini, 'Appunti 1981–1991'

Paolo Bufalini (Roma, 1915–2001) è stato fra i protagonisti dello scenario politico del secondo dopoguerra italiano: antifascista, dirigente del Partito Comunista Italiano (PCI), senatore della Repubblica dal 1963 al 1992. Laureato in giurisprudenza, ricopre un ruolo centrale in politica interna e in politica estera, soprattutto per ciò che riguarda i rapporti con il Vaticano nonché le leggi a favore del divorzio e dell'aborto ([21]). Parallelamente all'impegno in politica, Bufalini coltiva la propria passione per la letteratura italiana, europea e soprattutto latina, confermandosi apprezzato traduttore delle opere poetiche di Quinto Orazio Flacco ([23]).

Nel periodo che va dal 1981 al 1991, Bufalini tiene un quaderno privato di appunti manoscritti che intitola in modo generico 'Appunti, 1981–1991'. Il quaderno è composto da 145 pagine rilegate e due carte sciolte. Si tratta di una miscellanea di citazioni letterarie accompagnate da commenti e traduzioni dal latino all'italiano che corrispondono a momenti significativi della vita intellettuale e sociale dell'autore ([5]). Alcuni fra gli autori citati tra le pagine del quaderno di Paolo Bufalini sono Catullo, Lucrezio, Marziale, Orazio, Seneca, Tacito, Virgilio, Cicerone, Ennio, Giovenale, Giustiniano, Nevio, Plinio il Giovane, Quintiliano, Svetonio per i classici della letteratura latina; Dante, Manzoni, Petrarca, Alfieri, Belli, Campana, Campanella, Carducci, Croce, Passavanti, Pontano, Ripamonti, Tasso per la letteratura italiana; Flaubert, Hegel, Mann, Shakespeare, Tolstoj, Toqueville, Yourcenar per la letteratura europea.

Dopo la morte di Bufalini i figli hanno donato il quaderno al Centro studi 'La permanenza del Classico' del Dipartimento di Filologia Classica e Italianistica dell'Università di Bologna in vista di una pubblicazione del quaderno. Tale documento offre alla comunità di studiosi l'opportunità di approfondire alcuni aspetti della personalità del Bufalini uomo politico e di aprire una finestra sul laboratorio intellettuale del Bufalini classicista e traduttore. Il Centro studi si è occupato di una prima trascrizione del documento e dello scioglimento dei riferimenti alle fonti di oltre 250 citazioni, spesso impliciti a causa della natura privata del quaderno. Inoltre, il Centro è responsabile di una prima di marcatura del documento secondo lo standard XML/TEI P4.

Oltre a stabilire da quali fonti letterarie sono stati tratti i testi che Bufalini cita nei suoi 'Appunti', gli studiosi del Centro si sono occupati di ristabilire le relazioni implicite che legano i frammenti testuali gli uni agli altri. I testi del quaderno formano una complessa rete di citazioni, commenti, traduzioni, citazioni di traduzioni, note e narrazioni personali. L'intento di Bufalini è ragionare sulla storia della letteratura, portare alla luce le influenze fra autori, creando connessioni fra testi e idee. Ad esempio, brani di letterati e filosofi giustapposti indicano un rapporto di influenza tra gli autori, così come l'accostamento di testi della letteratura italiana e latina che presentano termini in comune, proponendo poi diverse traduzioni in italiano dal testo latino.

Il quaderno è dunque contraddistinto da complesse dinamiche testuali, le quali hanno richiesto interventi di ricostruzione al fine di determinare, ad esempio, i frammenti testuali presenti, le traduzioni utilizzate (di Bufalini o preesistenti), le singole edizioni consultate, i rapporti fra i commenti e i testi citati, la paternità dei passi non attribuiti.

Sarà allora utile fare qualche esempio, che possa descrivere la complessità del dettato, svelando il rapporto fra il testo del quaderno e la rete dei diversi riferimenti espliciti e sottesi.

Le pagine 43–44 del quaderno mostrano un esempio di relazioni intratestuali e intertestuali (Illustrazione 1). La sezione si apre con la citazione in latino del libro VII, vv. 5-9, dell'Eneide di Virgilio. Enea, dopo aver dato sepoltura alla propria nutrice, riprende la navigazione. Il verso 9 descrive la visione del mare che tremola colpito dalla scintillante luce lunare, immagine che Bufalini ritrova nei vv. 115-118 del canto I del Purgatorio dantesco, citati a seguire. Sono poi

riportate tre differenti traduzioni di Giuseppe Albini, la traduzione Carlo Saggio, la traduzione di Rosa Calzecchi Onesti e infine l'ipotesi di traduzione dello stesso Bufalini. Gli stessi passi dell'Eneide e della Commedia si ritrovano a pagina 12 del quaderno, giustapposti alle scene notturne evocate dai versi di Ennio e Lucrezio (Illustrazione 2).

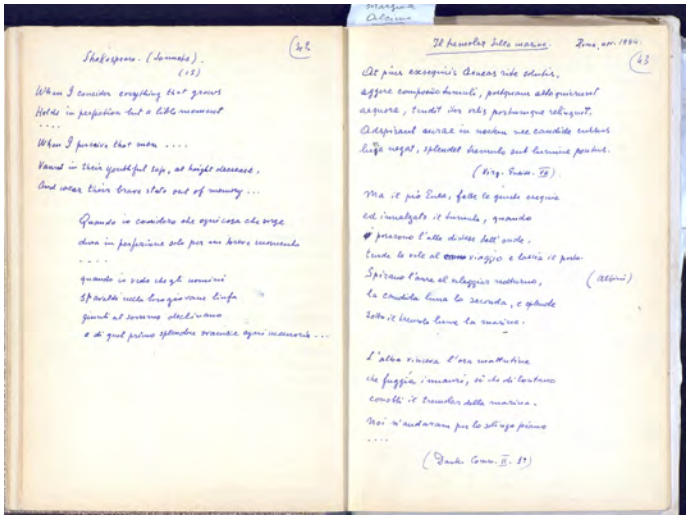

Illustrazione 1: Facsimile delle pp. 43-44 del quaderno di Appunti di Paolo Bufalini.

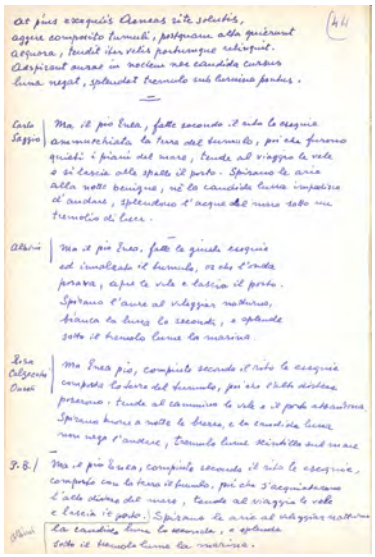

Illustrazione 2: Facsimile della p. 12 del quaderno di Appunti di Paolo Bufalini.

Altro esempio interessante, estratto dalle pagine 34–35 (Illustrazione 3), mostra una serie di riflessioni su Friedrich Nietzsche. Nella prima parte, Bufalini sostiene che il nietzschiano "culto

estetico del genio e dell'eroe", "insieme con il fermo convincimento che la felicità è impossibile […]", affondi le proprie radici nel pensiero di Schopenhauer. Nella seconda parte, Bufalini riporta un breve commento di Thomas Mann su Nietzsche, arricchendo poi il commento a Nietzsche di Mann con considerazioni personali.

È proprio l'esigenza di esprimere tale complessità di relazioni fra testi e autori citati nel quaderno che ha fin dagli inizi convinto gli studiosi del Centro dei vantaggi legati alla produzione di un'edizione digitale del quaderno.

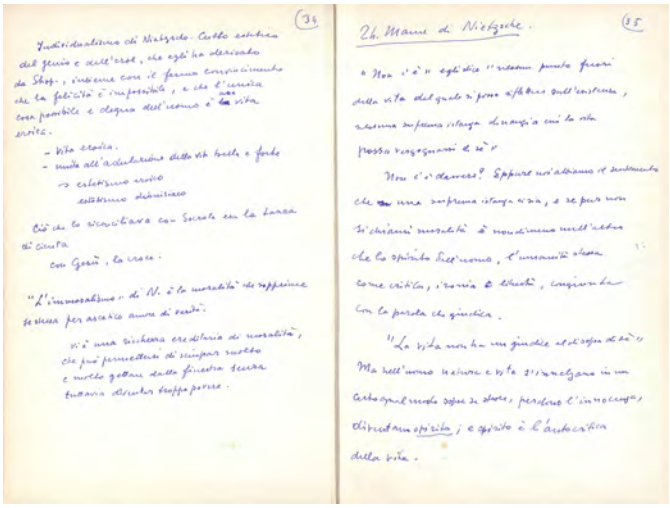

Illustrazione 3: Facsimile delle pp. 34-35 del quaderno di Appunti di Paolo Bufalini.

3. Il workflow di pubblicazione di una edizione semantica digitale

Il caso Bufalini è dunque esemplificativo per riflettere sull'utilizzo di tecnologie del Semantic Web per la pubblicazione di ESD. Si tratta di un fenomeno recentemente studiato dalla letteratura ([36]), che vede ad oggi alcuni tentativi di sperimentazione di un metodo di lavoro condivisibile ([10]; [12]; [20]). Non di meno, una attenta analisi dei vantaggi e degli svantaggi, degli obiettivi e dei costi e benefici che un nuovo workflow di pubblicazione di edizioni semantiche digitali comporterebbe nelle pratiche consolidate della comunità di filologia digitale è del tutto assente.

3.1 Problemi, vantaggi e controindicazioni

Tra le pratiche in uso nella filologia digitale, la codifica XML/TEI del testo è largamente adottata - anche se non unanimemente accettata (per cui cfr. [34] e [35]) - come passaggio fondamentale per la redazione di edizioni digitali. Tra gli indubbi vantaggi di questa scelta tecnologica c'è la possibilità di codificare aspetti legati al testo e al paratesto contestualmente, preservare informazioni relative al contesto dell'edizione in un unico file sorgente, trasformare agilmente documenti da XML a HTML e consentire la visualizzazione del testo su browser. A questi si aggiunge l'indubbio vantaggio dell'interoperabilità sintattica, dell'interscambio a livello di formato e la possibilità della condivisione di uno schema, ovvero di un vocabolario controllato e formale.

Tra gli svantaggi della codifica XML c'è innanzitutto la perdita di alcune informazioni significative registrate nella codifica TEI (p.e. il riferimento di stringhe a persone, eventi, o altri concetti) nel passaggio ad HTML. Benché diverse soluzioni - come l'evoluzione di HTML5 ([28]) o il progetto Polymer[2] - tentino di arricchire semanticamente il markup HTML, la trasformazione da XML/TEI a HTML non è immune a perdite di informazione o a scelte arbitrarie nella definizione del markup finale. Viceversa, il modello RDF permette di rappresentare univocamente concetti e relazioni espresse nei dati utilizzando vocabolari di riferimento, ma non è pensato per la visualizzazione dei dati stessi. Inoltre, RDF consente di rappresentare fenomeni tipici del testo, come la sequenza di elementi testuali tramite la creazione liste ordinate di entità (W3C 2014), ma risulta inefficace e complesso in fase di interrogazione e quindi di restituzione all'utente.

In secondo luogo, il vocabolario TEI consente ai filologi di descrivere gli stessi fenomeni testuali utilizzando diversi (o diverse combinazioni di) descrittori. A tale flessibilità corrisponde una minore interoperabilità tra edizioni digitali e l'interrogazione di più edizioni contemporaneamente risulta inevitabilmente più complessa. L'utilizzo di ontologie, schemi RDF (W3C 2014) e i thesauri SKOS ([22]) è una soluzione ottimale per superare tali difficoltà e garantire l'interoperabilità semantica tra fonti di dati.

In terzo luogo, all'interno di un processo di pubblicazione digitale convenzionale, la fase di codifica del testo è quella che richiede il maggior dispendio in termini di risorse, tempo e *know-how*. Inoltre il modello dell'edizione digitale basato su codifica XML rimane un sistema chiuso, che difficilmente dialoga con risorse esterne esistenti (come accennato sopra), e pertanto deve includere tutte le informazioni necessarie a supportare la corretta comprensione del testo (p.e. bibliografia, riferimenti biografici, commenti editoriali, metadati descrittivi e gestionali del testo). Il reperimento di tali informazioni richiede tempo, può risultare parziale e può generare informazioni divergenti con altre fonti autorevoli. L'integrazione di informazioni parziali o mancanti in una base di conoscenza è invece uno dei benefici maggiori derivanti dall'utilizzo dei Linked Open Data, favorito dall'utilizzo di ontologie condivise dalla comunità di riferimento per semplificare il dialogo.

In ultimo, la preservazione a lungo termine del prodotto digitale dipende dalle scelte e le

2 https://www.polymer-project.org/

possibilità dell'autore. Un identificativo persistente (p.e. un DOI) viene generalmente attribuito all'artefatto finale. In casi di eccellenza come la banca dati Perseus, identificativi persistenti possono essere attribuiti anche ai singoli versi del testo, i.e., `cts:urn`.[3] Ad oggi in nessuna edizione digitale vengono invece identificati i contenuti dell'edizione (p.e. letture di varianti, scelte stilistiche, note esplicative). Questi elementi possono invece variare da una versione all'altra dell'edizione digitale, senza essere registrati formalmente (i.e., assicurando un *versioning* dell'edizione stessa) e garantendone la citabilità. Le tecnologie del Semantic Web offrono alcune soluzioni in merito. Innanzitutto, l'utilizzo di URI persistenti consente di identificare versioni delle asserzioni estratte dai documenti. In secondo luogo, il modello RDF consente di distinguere diversi livelli di annotazione. I *named graphs,* ovvero grafi identificati univocamente da URI persistenti, possono essere a loro volta annotati con informazioni sulla *provenance* delle asserzioni incluse nel grafo, p.e. sull'autorialità del testo codificato o autori citati dall'autore del testo, o sulla provenance del grafo stesso, p.e sull'autore dell'edizione del testo. Il modello delle *nanopublication* ([17]) è pensato per rappresentare informazioni estratte da articoli scientifici mediante l'organizzazione in named graphs. Inoltre, PROV-O Ontology (Lebo et al. 2013) è lo standard identificato dal W3C per rappresentare informazioni inerenti le attività di creazione e manipolazione dei dati. La combinazione di queste tecnologie consente di organizzare ad alto livello e rappresentare formalmente gli aspetti ermeneutici relativi all'edizione del testo, identificare ogni singolo asserto frutto dell'interpretazione dell'autore dell'edizione e assicurarne la citabilità.

3.2 Obiettivi e contributi scientifici

Alla luce di questa breve analisi, la congiunzione di tecnologie tradizionali per la codifica del testo (ai fini della visualizzazione) e la rappresentazione semantica (per l'estrazione e l'identificazione di informazioni) si configura come uno scenario di partenza ideale su cui declinare un workflow di pubblicazione di edizioni semantiche digitali. La giustapposizione di markup e *knowledge extraction* per la creazione di un nuovo workflow di pubblicazione ci consente di (1) tenere effettivamente conto dello stato dell'arte nelle pratiche di filologia digitale e (2) metterlo a valore sperimentando altre tecnologie dedicate e consolidate. Gli obiettivi del nostro approccio possono riassumersi in quattro macro aree, descritte come segue:

- *Separazione degli aspetti di organizzazione della conoscenza dalla visualizzazione del testo*. Gli aspetti di semantica vengono formalizzati e rappresentati utilizzando tecnologie di Semantic Web, mentre gli aspetti relativi alla formattazione del testo vengono gestiti tramite markup e non vengono rappresentati semanticamente. Il contributo qui proposto è la selezione di alcune ontologie per la rappresentazione dei fenomeni caratteristici delle ESD indipendentemente dalla natura del testo in analisi (vedi Sezione 4). Nel contesto del quaderno di P. Bufalini, la scelta delle ontologie consente anche di rappresentare le specificità del testo e fornire viste significative sui dati che vadano oltre la lettura sequenziale, p.e. viste che consentano di passare dai

3 http://sites.tufts.edu/perseusupdates/beta-features/perseus-cts-api/

testi alle persone e dalle persone alle idee.

- *Migliorare gli aspetti di Knowledge Discovery indipendentemente dalle scelte implementative per la disseminazione dell'edizione.* Inserendo riferimenti ad authority, fonti di dati esistenti e vocabolari nella codifica XML/TEI del testo, l'obiettivo è registrare individui, classi e proprietà che verranno poi estratti automaticamente in fase di trasformazione in RDF. Così facendo si riducono i margini di errore derivanti da una riconciliazione effettuata con metodi automatici *a posteriori* (p.e. OpenRefine), garantendo l'accuratezza delle informazioni registrate, mantenendo l'indipendenza degli aspetti di semantica dalle scelte implementative per lo sviluppo dell'edizione digitale e assicurando la preservazione di tali aspetti curatoriali nel file XML sorgente (dove si registra la scelta degli authority effettuata consapevolmente dell'autore dell'edizione). Il contributo è la proposta di una codifica XML/TEI di authority file per rappresentare alcuni elementi di interesse emersi dall'edizione del quaderno di Bufalini, ovvero: persone, influenze, citazioni, riferimenti bibliografici (vedi Sezione 5). L'obiettivo specifico nel contesto del caso d'uso è evidenziare rapporti inter e intratestuali in una lettura non sequenziale del testo.
- *Riduzione dei costi nella fase di lavorazione del testo e reperimento fonti.* Introducendo riferimenti a fonti di dati esterne nella codifica XML/TEI abbiamo ridotto sensibilmente i tempi di lavorazione del testo, delegando le fasi di ricerca bibliografica e biografica ma senza cedere alla qualità dei contenuti. Il processo di trasformazione e riconciliazione effettuato a partire dalla codifica XML/TEI del quaderno di P. Bufalini è illustrato in Sezione 6.
- *Garanzia di citabilità delle asserzioni e versioning della edizione digitale.* La rappresentazione in named graphs secondo le specifiche delle nanopublication garantisce la citabilità persistente dei singoli fenomeni ermeneutici, contestualizzati alla versione corrente dell'ESD. Il contributo è discusso in Sezione 6 e i benefici sono discussi in Sezione 7.

3.3 Le fasi del processo di pubblicazione

Nella seguente tabella vengono riassunte le fasi del workflow di pubblicazione, le risorse (ore uomo) impiegate, le tecnologie utilizzate e gli output nel contesto del caso d'uso del quaderno di P. Bufalini. La discussione delle singole fasi è illustrata nelle sezioni seguenti.

Fase	Descrizione	Ore uomo	Tecnologie e tool	Output
Scelta dei modelli concettuali	Definizione delle *competency questions*	8	LOV	Descrizione testuale degli scenari

	Survey delle ontologie esistenti			Lista di ontologie
Revisione codifica XML/TEI	Inserimento authority record per persone, relazioni (citazioni, influenze) e riferimenti bibliografici. Revisione della codifica del testo per uniformare le scelte editoriali.	40	XML	Nuova versione del documento XML/TEI dell'edizione Documento di mapping da elementi TEI a statement RDF
Trasformazione in HTML/RDF	Trasformazione del documento XML in HTML Estrazione dati dal documento XML e creazione file RDF	8	XSLT Python	Foglio di stile XSL, file HTML Script Python, file RDF (.nq)
Sviluppo dell'applicazione web	Caricamento dati su triplestore Visualizzazione del testo Interrogazione triplestore e fonti esterne e creazione di viste sui dati	24	Python (Webpy), Blazegraph Triplestore, HTML, JS,CSS,	Applicazione Python
	Tot.	80		

Tabella 1: Fasi del workflow di pubblicazione dell'edizione semantica digitale del quaderno di P. Bufalini, risorse e tecnologie impiegate e output.

Il workflow presentato si basa su alcune assunzioni, alcune già illustrate, ovvero:

- Le ontologie esistenti sono in grado di rappresentare la maggior parte degli aspetti caratteristici delle ESD (p.e. rappresentazione del testo e delle sue parti, descrizione dei contenuti e delle relazioni tra entità estratte dal *full-text*).
- Il workflow proposto tiene conto dello scenario in cui una codifica XML/TEI del testo esiste già e deve essere rivisitata alla luce delle scelte di modellazione concettuale per uniformare la codifica e facilitare la trasformazione in RDF.
- la codifica del testo in XML e la seguente visualizzazione in HTML non vengono sostituiti dalla rappresentazione dei dati in RDF ma convivono all'interno dell'edizione finale. Quest'ultima si intende quindi composta da tre costituenti: (1) uno o più documenti XML, (2) uno o più documenti HTML e (3) una base di conoscenza inclusiva di grafi RDF e le ontologie scelte per la rappresentazione.

Il workflow illustrato tiene conto delle specificità del caso d'uso del quaderno di P. Bufalini, in particolare del livello di complessità della codifica XML/TEI del testo. Pertanto esiste un grado di variabilità nella definizione delle ore uomo che al momento non è generalizzabile.

Le tecnologie scelte in questa sede, benché non prescrittive, sono motivate da aspetti pragmatici:

1) XSL (Clark 1999) è un linguaggio standard per la manipolazione e la trasformazione di documenti XML in altri documenti XML. Pertanto la trasformazione da XML a HTML è adeguatamente supportata dal linguaggio.
2) Python (Van Rossum 1995) è un linguaggio di programmazione agile, adeguato sia ad assolvere operazioni legate alla computazione su diverse tipologie di dati, sia allo sviluppo web. In particolare, fornisce un ampio numero di librerie e metodi per manipolare dati RDF, interrogare triplestore e sviluppare applicazioni Model-View-Controller (MVC), p.e. Webpy.
3) Blazegraph[4] è stato scelto per la sua scalabilità, efficienza e per la semplicità di implementazione e utilizzo anche per non esperti.

4. Scelta dei modelli concettuali per la rappresentazione del testo

La fase di scelta delle ontologie è preceduta dall'analisi e la definizione di *competency questions*, ovvero degli scenari descrittivi a cui le ontologie devono essere in grado di rispondere. In particolare, le ontologie scelte per il workflow di pubblicazione delle ESD devono fornire termini per descrivere due macro-scenari inerenti il testo e i suoi componenti:

1) la descrizione degli elementi del testo (p.e. paragrafi, versi) e del peritesto (p.e. frontespizio, titoli, note, citazioni) così come registrati dall'autore;

4 https://www.blazegraph.com/

2) la descrizione degli elementi di epitesto (p.e. biografie, riferimenti bibliografici, analisi del testo) inseriti dall'autore dell'edizione.

Per rappresentare tali scenari si è optato per il riuso di modelli esistenti, quale buona pratica nella comunità del Semantic Web. Per reperire le ontologie esistenti e verificarne la diffusione e l'adeguatezza abbiamo scelto di utilizzare il catalogo online di ontologie Linked Open Vocabularies (LOV).[5]

In particolare, l'articolazione dei livelli che compongono l'oggetto testuale, gli elementi di peritesto ed eventualmente i testi citati nel corpo del testo (il contenuto del testo citato, l'edizione specifica citata, l'esemplare posseduto dall'autore - quando noto) sono rappresentati attraverso FaBiO, la formalizzazione OWL di FRBR inclusa nelle SPAR Ontologies ([25]). La scelta è motivata dalla necessità di rappresentare i quattro livelli dell'opera letteraria (opera, contenuto, manifestazione e individuo) identificati dal modello FRBR e ampiamente rappresentati dalle ontologie selezionate. Nel seguente esempio è riportata la rappresentazione (in sintassi Turtle) dei versi 5-9 del settimo volume dell'Eneide:

```
@prefix cito: <http://purl.org/spar/cito/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix frbr: <http://purl.org/vocab/frbr/core#> .
@prefic fabio: <http://purl.org/spar/fabio/>.
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix prism1: <http://prismstandard.org/namespaces/1.2/basic/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix seq:
<http://www.ontologydesignpatterns.org/cp/owl/sequence.owl#> .
@prefix tei: <http://www.tei-c.org/ns/1.0> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<https://w3id.org/bufalinis-notebook/excerpt/bibl016> a fabio:Excerpt ;
    rdfs:label "Eneide, volume 7, verse 5-9" ;
    frbr:partOf <https://w3id.org/bufalinis-notebook/text/pvm-
eneide/volume-7> .
<https://w3id.org/bufalinis-notebook/text/pvm-eneide> a fabio:Book ;
    rdfs:label "Eneide" ;
    frbr:realizationOf <https://w3id.org/bufalinis-notebook/work/pvm-
eneide> .

<https://w3id.org/bufalinis-notebook/text/pvm-eneide/volume-7> a
fabio:Expression ;
    rdfs:label "Eneide, volume 7" ;
    frbr:partOf <https://w3id.org/bufalinis-notebook/text/pvm-eneide> .

<https://w3id.org/bufalinis-notebook/work/pvm-eneide> a fabio:Poem ;
    rdfs:label "Eneide" ;
```

5 http://lov.linkeddata.es

```
    dcterms:creator <https://w3id.org/bufalinis-notebook/person/pvm> ;
    = <http://viaf.org/viaf/313680866>,
        <http://worldcat.org/entity/work/id/3144508483> .
```

Le informazioni estratte dal testo pieno, p.e. i ruoli di persone citate, le relazioni tra opere citate da Bufalini, i.e. le citazioni tra opere terze registrate da Bufalini e le citazioni tra autori (accordo, disaccordo, disputa, citazione generica), vengono descritte mediante l'uso di (1) Open Annotation Model ([32]), che consente di esplicitare le annotazioni di un testo, e (2) CiTO ([26]), che consente di rappresentare relazioni intra e intertestuali.

Ad esempio, la traduzione dei sopra citati versi dell'Eneide può essere rappresentata come segue (in sintassi Turtle):

```
<https://w3id.org/bufalinis-notebook/quote/quot-077> a oa:Annotation ;
    oa:hasBody <https://w3id.org/bufalinis-notebook/excerpt/bibl016> ;
    oa:hasTarget
       <https://w3id.org/bufalinis-notebook/citation/virgilio-eneide-7-
5-9c> ;
    oa:motivatedBy <https://w3id.org/bufalinis-notebook/quotation> .

<https://w3id.org/bufalinis-notebook/citation/virgilio-eneide-7-5-9c> a
cito:Citation ;
    cito:hasCitedEntity <https://w3id.org/bufalinis-
notebook/excerpt/bibl016> ;
    cito:hasCitingEntity <https://w3id.org/bufalinis-notebook/page/44> .

<https://w3id.org/bufalinis-notebook/quotation> rdfs:label "quotation" .

<https://w3id.org/bufalinis-notebook/excerpt/bibl016>
prism1:startingPage <https://w3id.org/bufalinis-notebook/page/44> ;
    rdf:value """
                <section xmlns="http://www.tei-c.org/ns/1.0">
                   <lg>
                      <l>At pius exequiis Aeneas rite solutis,</l>
                      <l>aggere composito tumuli, postquam alta
quierunt</l>
                      <l>aequora, tendit iter velis portumque
relinquit.</l>
                      <l>Adspirant aurae in noctem nec candida cursus</l>
                      <l>luna negat, splendet tremulo sub lumine
pontus.</l>
                   </lg>
                </section>
             """^^rdf:XMLLiteral.
```

Le affermazioni sui testi citati da Bufalini, le influenze tra autori così come riportate nelle note di P. Bufalini e le assunzioni fatte dagli autori dell'edizione vengono rappresentate mediante HiCO ([9]), un'estensione di PROV-O, che permette di descrivere relazioni tra persone, attività e artefatti e rappresentare gli aspetti ermeneutici. Ad esempio, Bufalini afferma che il pensiero di Nietzsche affonda le sue radici in Schopenhauer e rafforza la propria tesi citando un commento di Mann sul rapporto tra Nietzsche e Schopenhauer. Questa situazione può essere

descritta (in sintassi Turtle) come segue:

```
pb:infl-fn-agreesWith-as-quot-059int-act
       a hico:InterpretationAct ;
       hico:hasInterpretationType pbtype:relation-between-people ;
       hico:hasInterpretationCriterion criterion:author-quotation ;
       cito:obtainsBackgroundFrom pb:infl-fn-agreesWith-as-quot-059int-
act-tm ;
       hico:isExtractedFrom pb-quot:quot-059 ;
       cito:citesAsAuthority person:tm ;
       prov:wasAssociatedWith person:pb .
```

Pertanto abbiamo una speculazione fatta da Bufalini (`pb:infl-fn-agreesWith-as-quot-059int-act`), supportata dalla citazione (`criterion:author-quotation`) di un frammento del quaderno (`pb-quot:quot-059`) in cui si riporta una simile speculazione fatta questa volta da Thomas Mann (`pb:infl-fn-agreesWith-as-quot-059int-act-tm`).

5. Revisione del modello di codifica XML/TEI

Il passaggio da TEI a RDF richiede una fase di revisione della marcatura TEI alla luce della necessità di individuare entità e relazioni che andranno a popolare l'ontologia. Stabilite le domande di ricerca e i modelli ontologici per esprimere i temi rilevanti per il quaderno di Bufalini, il modello di codifica XML/TEI è stato rivisto per facilitare la trasformazione, garantendo inoltre una maggiore coerenza e consistenza delle scelte di codifica. Inoltre, il vocabolario stesso è stato aggiornato alla più recente versione TEI P5. Il nuovo modello di codifica ha come obiettivo principale l'esplicitazione delle relazioni fra le diverse componenti testuali del quaderno (citazioni, commenti, traduzioni, ecc.) e l'identificazione dei collegamenti tra authority file locale e dataset esterni. Le principali tipologie di relazione formalizzate sono illustrate di seguito, accompagnate da esempi.

1. Relazione fra un testo citato e la sua traduzione (ad opera di autori vari oppure di Bufalini stesso). Bufalini riporta i versi 5-9 dell'*Eneide*, volume VII, e diversi esempi di traduzione. Il testo dell'*Eneide* è racchiuso in un elemento `quote`, la traduzione di Bufalini è racchiusa in un elemento `note` di tipo 'traduzione' e le traduzioni di altri autori sono racchiuse in un elemento quote di tipo 'traduzione'. Alle traduzioni è attribuito un `@corresp`, contenente il riferimento all'identificativo univoco del frammento tradotto.

```
<cit xml:id="virgilio-eneide-7-5-9c" sameAs="#virgilio-eneide-7-5-9">
    <quote xml:id="quot-077" xml:lang="lat" source="#bibl016">
        <lg><l>At pius exequiis Aeneas rite solutis,</l>[...]</lg>
    </quote>
```

```
</cit>

<cit xml:id="g-albini-translation-eneide-b">
    <bibl><editor role="translator" ref="#GA">Albini</editor></bibl>
    <quote type="translation" xml:id="tra-015" corresp="#quot-077"
resp="#GA" source="#bibl070b">
        <lg><l>Ma il pio Enea, fatte le giuste esequie</l>[...]</lg>
    </quote>
</cit>
```

2. Relazione fra le diverse occorrenze di uno stesso testo. Bufalini riporta i versi sopracitati dell'Eneide alle pp. 8, 12, 43 e 44. Ogni occorrenza di questi esatti versi è racchiusa in un elemento `quote`, a sua volta racchiuso in un elemento `cit`, e ogni atto citazionale è identificato univocamente (`@xml:id`). Il frammento citato è ricondotto al record bibliografico dell'opera contenuto nell'elemento `teiHeader` per mezzo dell'attributo `@source`, mentre l'equivalenza tra frammenti citati è esplicitata dall'attributo `@sameAs`.

```
<cit xml:id="virgilio-eneide-7-5-9">
    <quote xml:id="quot-017" xml:lang="lat" source="#bibl016">
        <lg><l>At pius exequiis Aeneas rite solutis,</l>[...]</lg>
     </quote>
    [...]
</cit>

<cit xml:id="virgilio-eneide-7-5-9c" sameAs="#virgilio-eneide-7-5-9">
    <quote xml:id="quot-077" xml:lang="lat" source="#bibl016">
        <lg><l>At pius exequiis Aeneas rite solutis,</l>[...]</lg>
    </quote>
</cit>
```

3. Relazione fra testi accostati da Bufalini per affinità di forma e/o contenuti. Bufalini riporta *Eneide*, VII, 5-9 e *Purgatorio*, I, 115-117. L'accostamento dei versi dell'*Eneide* e del *Purgatorio* sottintende un rapporto di influenza fra Virgilio e Dante. Tale rapporto, è descritto sinteticamente nell'elemento `teiHeader` in un elemento `relation`; il testo precedente da un punto di vista temporale costituisce l'entità influente (attributo `@active`), mentre il secondo rappresenta l'entità influenzata (attributo `@passive`).

```
<teiHeader>
    [...]
    <relation type="influence" active="#quot-077" passive="#quot-076" >
        <desc>Verg. Aen. 7, 8-9 in D. Alighieri, D. C. 2, 1, 115-
117</desc>
    </relation>
   […]
</teiHeader>
```

4. Relazione fra intellettuali, dei quali Bufalini riporta e commenta il pensiero. Bufalini sostiene che il pensiero di Nietzsche affondi le sue radici in Schopenhauer e rafforza la propria tesi citando un commento di Mann sul rapporto tra Nietzsche e Schopenhauer. Il rapporto di

relazione fra il pensiero di Nietzsche e quello di Schopenhauer è descritto nel `teiHeader` nell'elemento `relation`. Il tipo di relazione, in questo caso di accordo, è esplicitato dal valore `agreesWith` dell'attributo `@name`. La relazione riporta il collegamento al frammento testuale da cui la relazione è estratta attraverso l'attributo `@corresp`. Infine, l'attributo `@source` indica l'autore dell'affermazione, in questo caso Thomas Mann, seguita da una breve descrizione testuale fatta dagli editori.

```
<teiHeader>
    [...]
       <relation name="agreesWith" active="#FN"
                     passive="#AS" corresp="#quot-059" source="#TM">
        <desc>T. Mann sottolinea come F. Nietzsche abbia
              derivato il culto dell'eroe da A. Schopenhauer</desc>
     </relation>
    [...]
</teiHeader>
```

5. Collegamento ad authority file e/o basi di conoscenza esterne, quali VIAF, DBpedia e Worldcat. L'inserimento di riferimenti ad authority file esterni per identificare persone e opere citate nel quaderno è fra le più significative modifiche apportate alla codifica TEI originale. Ogni occorrenza di nome di persona è disambiguata tramite l'utilizzo di un attributo `@ref` che rimanda all'identificativo univoco della persona descritta nel `teiHeader` e collegata ai record VIAF e DBpedia corrispondenti attraverso l'attributo `@sameAs`.

```
<teiHeader>
  <listPerson>[…]
       <person xml:id="BC">
              <persName sameAs="http://viaf.org/viaf/61544292
                     http://it.dbpedia.org/resource/Benedetto_Croce">Cro
                     ce, Benedetto (1866-1952)</persName>
        </person>[…]
  </listPerson>
</teiHeader>

<text>[…]
    <persName ref="#BC">B. Croce</persName>[…]
</text>
```

La fonte di ogni testo citato è disambiguata tramite l'utilizzo di un attributo `@source` che contiene un rimando all'identificativo univoco dell'opera, a sua volta descritta nel `teiHeader` e collegata ai record VIAF, DBpedia e Worldcat corrispondenti attraverso l'attributo `@sameAs`.

```
<teiHeader>[…]
    <bibl xml:id="eneide-7" resp="#CSPC">
       <author ref="#PVM"/>
       <title level="m" type="Poem" subtype="Book"
              sameAs="http://viaf.org/viaf/313680866
                     http://worldcat.org/entity/work/id/3144508483">Enei
```

```
                de</title>
        <biblScope unit="volume"
                sameAs="http://viaf.org/viaf/176577261">7</biblScope>
        <citedRange xml:id="bibl016" unit="verse" from="5" to="9">5-
9</citedRange>
    </bibl>[…]
</teiHeader>

<quote xml:id="quot-077" xml:lang="lat" source="#bibl016">
[…]
</quote>
```

6. La trasformazione in HTML e RDF e le risorse realizzate

La fase successiva del workflow comprende (1) il passaggio dalla codifica XML/TEI ad un documento HTML contenente la trascrizione del testo e (2) l'estrazione dei dati e la trasformazione in RDF secondo le ontologie scelte.

Assumiamo che la trascrizione del testo definitivo per la pubblicazione non subisca modifiche continue e che ad ogni modifica corrisponda una nuova versione dell'edizione da identificare univocamente. Pertanto il documento HTML finale può essere ottenuto mediante una trasformazione XSL effettuata *una tantum*. Tale scelta è motivata anche da considerazioni tecniche relative alla disseminazione dell'edizione digitale. Le applicazioni web statiche hanno indubbi vantaggi rispetto ad applicazioni dinamiche in termini di manutenzione e aggiornamento. File HTML statici possono essere visualizzati senza richiedere una trasformazione XSL *server-side* e dover dipendere dall'utilizzo di tecnologie instabili (come l'accesso al triplestore in cui sono conservati i dati RDF).

Nel passaggio ad HTML viene fatta una selezione delle informazioni da visualizzare, ottimizzando il processo di trasformazione e focalizzandolo solo sugli aspetti di impaginazione. Ad esempio gli authority file non vengono inclusi nel documento finale, mentre verranno utilizzati per estrarre dati e creare la base di conoscenza.

Similmente, l'estrazione dei dati dal file XML per la creazione della base di conoscenza avviene su un numero selezionato di elementi TEI, identificati nel documento di mapping menzionato in Sezione 5 così da rispondere alla descrizione degli scenari illustrati in Sezione 4. In particolare vengono estratte le seguenti entità:

1. le componenti del testo (note, commenti, traduzioni, citazioni) vengono estratte dall'elemento `body` del documento XML;
2. le persone menzionate dall'autore vengono estratte dall'authority inserito nel `:teiHeader`;
3. le entità bibliografiche citate dall'autore vengono estratte dall'authority inserito nel `teiHeader`.
4. i luoghi in cui l'autore si è trovato a scrivere i suoi appunti vengono estratti dall'elemento `body` del documento XML;

Date le specificità del testo di Bufalini, si è scelto di estrarre le seguenti relazioni tra entità:

5. le relazioni intratestuali (p.e. citazione di frammenti di testo di altri autori), estratte dall'elemento `body` del documento XML;
6. le relazioni intertestuali tra frammenti del testo così come dedotte dagli editori (p.e. tra citazione e commento, tra citazione e traduzione del testo citato), estratte dall'elemento `body` del documento XML;
7. la menzione di persone in frammenti del testo del quaderno e riferimento al punto in cui vengono menzionate, estratte dall'elemento `body` del documento XML;
8. le influenze tra autori, così come esplicitato in nota dell'autore o come dedotto dagli editori, interpretando il pensiero - non esplicito - di Bufalini, estratte da un authority dedicato inserito nel `teiHeader`;
9. le influenze tra autori e testi, come evidenziato dalla giustapposizione di frammenti testuali in cui sono evidenziate graficamente relazioni intra-testuali, estratte da un authority dedicato inserito nel `teiHeader`.

In particolare, i punti 2, 4 e 5 hanno richiesto un ulteriore livello di annotazione per registrare il livello di incertezza dell'asserzione e la provenienza delle asserzioni stesse, ovvero quando un fenomeno è esplicitamente registrato da Bufalini o è stato dedotto dagli autori dell'edizione. Per tali situazioni le asserzioni, la provenance delle asserzioni e le informazioni relative all'edizione sono inserite in named graphs dedicati e collegati tra loro tramite un grafo *header*. Il modello per rappresentare tali collegamenti è quello delle nanopublication, come esemplificato nella seguente figura.

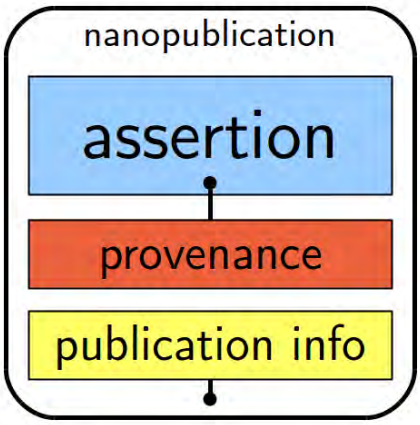

Illustrazione 4: Rappresentazione grafica dei grafi di una nanopublication. Khun, T., Nanopublications. Provenance-Aware Linked Data Publishing, 2015.

Riprendendo l'esempio sopra citato dell'influenza tra Nietzsche e Schopenauer estratta da un passaggio di Thomas Mann, questo scenario può essere descritto come segue (in notazione Trig):

```
# the head of the nanopublication
<http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-059-
head/> {
   pb:infl-fn-agreesWith-as-quot-059-np a np:Nanopublication ;
               np:hasProvenance
         <http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-
         059-provenance/> ;
               np:hasAssertion
         <http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-
059-assertion/> ;
               np:hasPublicationInfo
         <http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-
059-pubinfo/> .
}

# the assertions about the influence between two authors
<http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-059-
assertion/> {
         person:fn cito:agreesWith person:as .
         person:as prov:qualifiedInfluence pb:infl-fn-agreesWith-as-quot-
059 .
         pb:infl-fn-agreesWith-as-quot-059
               rdfs:label "T. Mann sottolinea come F. Nietzsche abbia
derivato
                        il culto dell'eroe da A.
               Schopenhauer"^^xsd:string .
               prov:agent <http://w3id.org/bufalinis-
notebook/person/fn> .
}

# the interpretation of such an influence by Bufalini and Mann's source
of information
<http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-059-
provenance/> {
         pb:infl-fn-agreesWith-as-quot-059int-act a hico:InterpretationAct
;
               cito:obtainsBackgroundFrom
               <http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-
         quot-059int-act-tm> ;
               hico:isExtractedFrom pbquot:quot-059> ;
               hico:hasInterpretationType pbtype:relation-between-
people ;
               hico:hasInterpretationCriterion criterion:author-quotation
;
               cito:citesAsAuthority person:tm ;
```

```
              prov:wasAssociatedWith person:pb .
       pb:infl-fn-agreesWith-as-quot-059int-act-tm  a
hico:InterpretationAct .

       <http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-
059-assertion/>
              prov:wasGeneratedBy pb:infl-fn-agreesWith-as-quot-059int-
act .
}

# info about editors
<http://w3id.org/bufalinis-notebook/infl-fn-agreesWith-as-quot-059-
pubinfo/> {
       pb:infl-fn-agreesWith-as-quot-059-np
              prov:wasAttributedTo <http://w3id.org/bufalinis-
notebook/organization/cspc> ;
              prov:generatedAtTime "2018-01-
01T00:00:01+00:00"^^xsd:dateTime .
       pb:cspc a foaf:Organization ;
              rdfs:label "Centro Studi Permanenza del Classico,
              Università di Bologna"^^xsd:string .
}
```

In ultimo, la scelta degli authority file e dei dataset da collegare è stata fondamentale per completare la base di conoscenza e fornire una migliore esperienza informativa all'utente. Come già esemplificato, link a VIAF e DBpedia sono stati creati per identificare ed estendere le informazioni biografiche delle personalità citate. Similmente, quando è stato possibile identificare la specifica edizione di un testo citato è stato creato un link a Worldcat e VIAF; nei casi in cui è nota l'opera ma non l'edizione citata, è stato creato il solo link a VIAF, in cui sono rappresentate unicamente le opere ma non le edizioni. In totale, sono stati creati circa 200 link a entità esterne.

Il risultato della trasformazione è una serie di grafi RDF distinti per autorialità: (1) un grafo dell'edizione, che include tutti gli elementi testuali (circa 400, di cui 240 citazioni) così come riportati da Bufalini e che conta circa 6000 triple; (2) 164 grafi contenenti circa 900 triple inerenti le asserzioni, la provenance e le note editoriali relative ai passaggi del testo che sono stati interpretati dagli autori dell'edizione (relazioni tra persone, autori e testi).

7. L'edizione online: una valutazione preliminare di costi e benefici

Le risorse che compongono l'ESD finale sono: un documento XML/TEI, un documento HTML, la base di conoscenza comprensiva di grafi RDF e ontologie e l'applicazione web che

ospita l'edizione. L'applicazione[6] e il codice per realizzarla[7] sono disponibili online. Da questa è possibile (1) fruire del testo dell'edizione in due modalità, navigando i facsimili annotati del manoscritto o visionando la sola trascrizione, (2) esplorare i contenuti della base di conoscenza tramite una serie di viste (indici) sugli autori, le opere e le relative citazioni, (3) esplorare la ricostruzione della biblioteca di Bufalini e (4) interrogare la base di conoscenza tramite un'interfaccia grafica.

Lo sviluppo dell'applicazione è orientato alla valorizzazione degli aspetti caratteristici del testo, pertanto non è oggetto di valutazione del workflow. A partire dal risultato finale possiamo però generalizzare alcune considerazioni finali, evidenziando i costi affrontati per ottenere l'edizione digitale e i benefici apportati ad una ESD tradizionale.

Da un punto di vista del software richiesto per mettere in produzione l'applicazione, nuove tecnologie vengono aggiunte al workflow di pubblicazione tradizionale. Le tecnologie impiegate e sufficienti alla realizzazione dell'applicazione sono tutte open source, sviluppate, manutenute e adottate da una larga comunità. Perciò non si pongono problemi di obsolescenza a breve termine e i costi per la loro adozione sono nulli, quindi non impattano sui costi dell'edizione stessa. Le competenze richieste per lo sviluppo aumentano invece sensibilmente, essendo necessaria la figura di uno o più ontologisti con competenze filologiche e biblioteconomiche.

Come già accennato le tempistiche per lo sviluppo di una edizione semantica sono significativamente variabili, benché la variabilità non sia dettata dalle tecnologie per il Semantic Web ma, per così dire, dall'adattabilità del pregresso. Infatti, lo sviluppo di un modello di rappresentazione dei dati in RDF può essere sensibilmente ottimizzato riutilizzando ove possibile ontologie già esistenti e consolidate. Invece, la fase di revisione della codifica XML/TEI richiede tempistiche che variano in base alla complessità dell'edizione, degli obiettivi di ricerca e dello specifico modello di codifica XML/TEI stabilito dagli editori. Le fasi di sviluppo (la trasformazione dei dati e lo sviluppo dell'applicazione) sono facilmente preventivabili e non incidono significativamente sui costi di realizzazione dell'ESD. Pertanto il workflow proposto non si propone di risolvere i problemi noti (scientifici più che tecnologici) derivanti dalla codifica XML/TEI, ma si propone di formalizzare e ottimizzare la fase di sviluppo separando gli aspetti di organizzazione della conoscenza dagli aspetti di visualizzazione. La definizione di un workflow alternativo all'utilizzo della codifica XML/TEI è fuori dagli obiettivi di questo progetto, la cui priorità è massimalizzare il valore aggiunto dei prodotti della ricerca risultanti da pratiche consolidate nella filologia digitale utilizzando tecnologie per il Semantic Web.

Per valutare preliminarmente i benefici derivanti da questo approccio consideriamo due requisiti non funzionali che l'edizione semantica vuole assolvere, ossia il bisogno informativo e la capacità di risoluzione dei problemi.

Per bisogno informativo si intende il raggiungimento degli obiettivi prefissati dagli autori dell'ESD, tali che il risultato sia ottimale dal punto di vista dell'utente. Nel caso dell'edizione digitale del quaderno di Bufalini l'obiettivo è offrire diverse modalità di fruizione del testo che

6 https://w3id.org/bufalinis-notebook/

7 https://github.com/marilenadaquino/bufalinis-notebook

per sua natura si presta a modalità di lettura differenti. In particolare l'edizione del quaderno offre una lettura sequenziale per immagini annotate, una lettura sequenziale della sola trascrizione e la lettura non sequenziale dei contenuti del testo. Quest'ultima, ottenuta tramite l'interrogazione della base di conoscenza, consente al lettore di svincolarsi dalla lettura tradizionale, sequenziale appunto, e navigare i medesimi contenuti (p.e. citazioni, note, traduzioni) tramite percorsi intuitivi e alternativi. Inoltre, al di là dell'attuale sviluppo dell'applicazione, l'utilizzo di URI persistenti e la riconciliazione a dataset esterni consentirà di sviluppare ulteriori percorsi conoscitivi, dando vita a nuove forme di *storytelling*.

Per risoluzione dei problemi si intende l'effettivo vantaggio derivante dalla scelta di strategie o l'utilizzo di tecnologie per superare problemi che un approccio tradizionale non sarebbe in grado di affrontare. Nel caso specifico, l'utilizzo di tecnologie per il Semantic Web consente di superare un problema annoso nella filologia digitale: la definizione dei livelli di descrizione in XML/TEI e l'overlap XML ([11]; [29]). La separazione tra elementi di semantica - richiamati in authority file interni al documento XML ma rappresentati esternamente al documento in grafi RDF - ed elementi di formattazione del testo - registrati nel documento HTML - comporta un indubbio vantaggio in fase di codifica del testo in XML. Qui l'autore dell'edizione può dedicarsi agli aspetti di ricerca, utilizzando una codifica sintetica per puntatori a informazioni il cui reperimento e approfondimento è delegato ad altre tecnologie, i.e. gli authority di persone, bibliografia, concetti e fenomeni.

Altro problema a cui le edizioni semantiche possono assolvere è una effettiva interoperabilità semantica tra edizioni. Se le scelte in fase di codifica rimangono a discrezione delle decisioni autoriali e delle esigenze descrittive, la rappresentazione RDF di fonti diverse può uniformarsi, rappresentando fenomeni (classi e proprietà) ed elementi (individui identificati da URI) utilizzando gli stessi vocabolari e ontologie.

In ultimo, l'adozione delle nanopublication per rappresentare informazioni incerte, contraddittorie (p.e. le letture di varianti) e suscettibili di cambiamento nel tempo ci consente di garantire la citabilità di ogni elemento dell'edizione, smentendo le critiche che vedono nel medium digitale un supporto instabile, non meritevole di fiducia e di minor valore rispetto ad una tradizionale edizione cartacea.

8. Conclusioni

In questo contributo abbiamo sollevato l'attenzione sulle problematiche veicolate dalle edizioni scientifiche digitali e l'adozione dei Linked Open Data per realizzare una edizione semantica digitale. Abbiamo proposto un workflow di pubblicazione che sappia valorizzare le pratiche correnti nella filologia digitale (i.e. la codifica XML/TEI) piuttosto che riproporre il discorso su una alternativa alla codifica TEI. Di fatto, il workflow proposto si propone di recuperare il pregresso, ovvero il patrimonio di edizioni digitali già realizzate, per renderle effettivamente interoperabili. Lungi dall'aver formalizzato ogni singolo aspetto della rappresentazione della conoscenza veicolata da una qualsivoglia edizione digitale, abbiamo proposto l'infrastruttura

teorica per consentire ad edizioni anche molto diverse di dialogare sugli aspetti di alto livello (i.e. elementi del testo, paratesto e relazioni fondamentali tra persone e oggetti culturali, intervento ermeneutico dell'autore dell'edizione).

Il caso d'uso del quaderno di 'Appunti' di Paolo Bufalini ci ha permesso di prendere coscienza delle fasi di tale processo di destrutturazione del testo, degli approcci esistenti e delle esigenze informative dell'utente finale. Le tecnologie per il Semantic Web ci consentono di esplicitare le relazioni intratestuali e intertestuali tra frammenti del testo, persone, luoghi, opere citate e il quaderno stesso, estendendo gli orizzonti informativi del sistema chiuso dell'edizione a fonti esterne. Inoltre, ci permettono di esplicitare e identificare univocamente interpretazioni divergenti su di uno stesso fenomeno testuale, svincolandoli dall'edizione stessa e rendendoli citabili in altre edizioni.

La valutazione preliminare sui costi e i benefici merita indubbiamente di essere approfondita. In futuro esploreremo la possibilità di applicare tale workflow ad un numero consistente di edizioni, valutando quanto le tempistiche varino a seconda della complessità del testo e delle scelte di codifica, stabilendo un livello minimo di interoperabilità tra edizioni (al di là dei soli metadati descrittivi delle edizioni stesse). In questa sede esploreremo la possibilità di sostituire (parzialmente) il lavoro manuale di codifica degli authority e di riconciliazione dei dati estratti dal testo con metodi semi-automatici che garantiscano comunque un alto livello di accuratezza. Così facendo speriamo di abbattere ulteriormente i limiti dettati dalla codifica e permettere davvero al filologo digitale di concentrarsi sui suoi temi di ricerca e delegare gli aspetti meccanici alla tecnologia.

References

[1] Clark, James. 1999. "XSL Transformations (XSLT) Version 1.0." November 16, 1999. W3C Recommendation. https://www.w3.org/TR/xslt-10/.

[2] Ciotti, Fabio, Marilena Daquino, and Francesca Tomasi. 2016. "Text Encoding Initiative Semantic Modeling. A Conceptual Workflow Proposal." In *Communications in Computer and Information Science*, 48–60. Springer International Publishing. https://doi.org/10.1007/978-3-319-41938-1_5.

[3] Ciotti, Fabio, and Francesca Tomasi. 2016. "Formal Ontologies, Linked Data, and TEI Semantics." *Journal of the Text Encoding Initiative* 9. https://doi.org/10.4000/jtei.1480.

[4] Citti, Francesco. 2010. "I classici nelle carte di un politico. Traduzioni e appunti di Paolo Bufalini." *Aufidus* 70: 7–32.

[5] Citti, Francesco. 2008. "Paolo Bufalini and the Classics: Towards a Digital Edition of His 'Note-Book.'" *Conservation Science in Cultural Heritage*. https://doi.org/10.6092/issn.1973-9494/1396.

[6] Ciula, Arianna, Paul Spence, and Jose Miguel Vieira. 2008. "Expressing Complex

Associations in Medieval Historical Documents: The Henry III Fine Rolls Project." *Literary and Linguistic Computing* 23, 3: 311–25. https://doi.org/10.1093/llc/fqn018.

[7] Ciula, Arianna, and Øyvind Eide. 2014. "Reflections on Cultural Heritage and Digital Humanities." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage - DATeCH '14*. ACM Press. https://doi.org/10.1145/2595188.2595207.

[8] Cyganiak, Richard, David Wood, and Markus Lanthaler, eds. 2014. "RDF 1.1 concepts and abstract syntax." W3C Recommendation. https://www.w3.org/TR/rdf11-concepts/.

[9] Daquino, Marilena, and Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects." In *Communications in Computer and Information Science*, 424–36. Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37.

[10] Del Grosso, Angelo Mario, Salvatore Cristofaro, Maria Rosa De Luca, Emiliano Giovannetti, Simone Marchi, Graziella Seminara, and Daria Spampinato. 2018. "Le lettere di Bellini: dalla Carta al Web." *AIUCD 2018*: 60.

[11] DeRose, Steven J., David Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is Text, Really?" *Journal of Computing in Higher Education* 1(2): 3-26.

[12] Di Donato, Francesca, and Susanne Müller. 2013. "Biblioteche digitali semantiche. Il progetto burckhardtsource.org." *Bibliotime* 16, 1.

[13] Doerr, Martin. 2009. "Ontologies for Cultural Heritage." In *Handbook on Ontologies*, 463–86. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-92673-3_21.

[14] Driscoll, Matthew James and Elena Pierazzo. 2016. *Digital Scholarly Editing: Theories and Practices*. Digital Humanities Series.

[15] Eide, Øyvind. 2014. "Ontologies, Data Modeling, and TEI." *Journal of the Text Encoding Initiative* 8. https://doi.org/10.4000/jtei.1191.

[16] Franzini, Greta. 2012. "Catalogue of Digital Editions.". https://dig-ed-cat.acdh.oeaw.ac.at/.

[17] Groth Paul, Andrew Gibson, and Jan Velterop. 2010. "The Anatomy of a Nanopublication." *Information Services & Use* 30, 1–2: 51–56. https://doi.org/10.3233/ISU-2010-0613.

[18] Italia, Paola Maria Carmela and Claudia Bonsi (eds.). 2016. *Edizioni a confronto-comparing editions.* ROMA: Sapienza Università Editrice, pp. 198.

[19] Lebo, Timothy, Satya Sahoo, and Deborah McGuinness, eds. 2013. "PROV-O: The PROV Ontology." W3C Recommendation. https://www.w3.org/TR/prov-o/.

[20] Maiatsky, Michail, Alexey Boyarsky, Natalia Boyarskaya, Ekaterina Velmezova, and Michael Piotrowski. 2018. "VICOGLOSSIA: Annotatable and Commentable Library as a Bridge between Reader and Scholar (a Proof of Concept Study: Early Soviet Philological Culture)." *Umanistica Digitale* 2. https://doi.org/10.6092/issn.2532-8816/7253.

[21] Matteoli, Giovanni. 2002. *Paolo Bufalini: L'impegno politico di un intellettuale*. Soveria Mannelli: Rubbettino.

[22] Miles, Alistair, and Sean Bechhofer. 2009. "SKOS Simple Knowledge Organization System Reference." W3C Recommendation. https://www.w3.org/TR/skos-reference/.

[23] Orazio Flacco, Quinto. 1993. *A Leuconoe, e altre poesie*. Translated by Paolo Bufalini. Milano: All'insegna del pesce d'oro.

[24] Ore, Christian-Emil, and Øyvind Eide. 2009. "TEI and Cultural Heritage Ontologies: Exchange of Information?" *Literary and Linguistic Computing* 24, 2: 161–72. https://doi.org/10.1093/llc/fqp010.

[25] Peroni, Silvio, and David Shotton. 2018. "The SPAR Ontologies." In *Lecture Notes in Computer Science*, 119–36. Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_8.

[26] Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: Ontologies for Describing Bibliographic Resources and Citations." *Journal of Web Semantics* 17: 33–43. https://doi.org/10.1016/j.websem.2012.08.001.

[27] Pierazzo, Elena. 2015. "Digital Scholarly Editing: Theories, Models and Methods." In Digital Research in the Arts and Humanities.

[28] Pilgrim, Mark. 2010. *HTML5: up and running: dive into the future of web development*. "O'Reilly Media, Inc.".

[29] Renear, Allen, Elli Mylonas, and David Durand. 1993/1996. "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." In *Research in Humanities Computing*.

[30] Rossum, Guido. 1995. *Python reference manual*. Amsterdam: Centrum voor Wiskunde en Informatica.

[31] Sahle, Patrick. 2013. *Digitale Editionsformen*. Teil,1-3.

[32] Sanderson, Robert, Paolo Ciccarese, and Herbert Van de Sompel, eds. 2013. Open Annotation Data Model. W3C Community Draft. http://www.openannotation.org/spec/core/.

[33] Schloen, David, and Sandra Schloen. 2014. "Beyond Gutenberg: Transcending the Document Paradigm in Digital Humanities." *Digital Humanities Quarterly* 8, 4. http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html.

[34] Schmidt, Desmond. 2010. "The Inadequacy of Embedded Markup for Cultural

Heritage Texts." *Literary and Linguistic Computing* 25, 3: 337–56. https://doi.org/10.1093/llc/fqq007.

[35] Schmidt, Desmond. 2012. "The Role of Markup in the Digital Humanities." *Historical Social Research / Historische Sozialforschung* 37, 3: 125-46. http://www.jstor.org/stable/41636601.

[36] Tomasi, Francesca. 2013. "Digital Editions as a New Model of Conceptual Authority Data." JLIS 4. https://doi.org/10.4403/jlis.it-8808.