Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Titolo	**D 1.2: Modelli dati e contesti d'uso**
Deliverable	D1.2 Documento che analizzi le specificità di vari contesti d'uso esistenti e dei modelli di dati utilizzati.
Release	Versione 1.0
Partner	CELI, Università degli Studi di Torino
Autori principali	Andrea Bolioli (CELI), Matteo Casu (CELI) Roberto Rosselli del Turco (Unito)

Indice

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

Introduzione

La trattazione e la scelta di modelli dati adeguati per annotazioni semantiche di edizioni digitali non può non basarsi sui casi d'uso previsti, e sugli attori coinvolti in essi: a chi è rivolta l'edizione? A una categoria di utenti (ad esempio, i filologi) o a più d'una (ad esempio, studiosi, pubblico generalista, studenti di scuola superiore) o ancora a più di una, permettendo collaborazioni tra le diverse categorie di utenti? Infatti, mentre una semplice *pubblicazione digitale* di un'edizione cartacea può limitarsi a fornire uno sfogliatore che permetta di leggere il testo, una vera e propria *edizione digitale*, per di più annotata semanticamente, richiede di poter arricchire il testo con annotazioni di varia granularità e di vario tipo (a seconda, appunto, dei casi d'uso previsti) che insistano su porzioni di testo, e opzionalmente di poter tracciare relazioni tra le annotazioni. Il testo e gli altri contenuti devono quindi essere rappresentati in modo strutturato, e dando ragione di diversi livelli di analisi.

In questo report assumiamo che il testo dell'opera sia sempre presente nell'edizione digitale[1] . A seconda dei casi il testo può essere visualizzato con lo stesso layout dell'edizione cartacea (tipicamente nelle edizioni diplomatiche digitali) oppure essere visualizzato con un layout slegato da edizioni cartacee, ad esempio in pubblicazioni digitali rivolte al pubblico generalista, tipicamente più improntate a presentare testo e contenuti in modo più accattivante.

I tipi di annotazione da prevedere sul testo si possono suddividere lungo due assi:

- annotazioni tipografiche o strutturali: servono a specificare la struttura del testo (suddivisione in righe, versi, paragrafi capitoli), ed eventualmente la sua resa grafica (stili ed enfatizzazioni di porzioni di testo)
- annotazioni semantiche: servono per collegare porzioni di testo a "informazioni semantiche", ad es i link tra porzioni di testo e entità nominate (luoghi, persone, ecc), ma anche annotazioni linguistiche di porzioni di testo (metadati linguisticici)

I modelli dati usati dalle varie applicazioni web che pubblicano edizioni digitali si distinguono soprattutto per le scelte concettuali e tecnologiche inerenti la memorizzazione delle annotazioni lungo questi due assi concettuali. Alcune scelte portano a distinguere questi tipi di annotazione anche nel formato di memorizzazione, mentre altre trattano tutte le annotazioni con lo stesso modello dati.

[1] Assumiamo di trattare edizioni digitali dove il testo delle opere/edizioni sia presente in una qualche forma. Esistono applicazioni web che pubblicano i contenuti di opere senza riportarne il testo -- un esempio è *DanteSources* -- si veda D1.1.

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Annotazioni

Annotazioni strutturali e semantiche

La prima fondamentale scelta sulla rappresentazione e memorizzazione delle annotazioni, che si può far risalire alla linguistica computazionale, è quella tra annotazioni *inline* e *stand-off*. Le annotazioni stand-off, usate nel Natural Language Processing, hanno raggiunto l'ambito delle pubblicazioni digitali negli anni '90 del '900: il formato TEI, nella sua quinta revisione, ne prevede l'uso accanto alle annotazioni inline (vedi sezione 3 di Banski 2010)[2].

Annotazioni *inline* e *stand-off*

La differenza fondamentale tra i due tipi di annotazione (o meglio, tra i due metodi per salvare annotazioni) consiste nel fatto che le annotazioni inline sono presenti sottoforma di markup (nel caso di XML, come elementi XML) direttamente nel testo annotato:

Lorem **<annotazione1>***ipsum dolor***</annotazione1>** *sit amet, consectetur adipiscing elit*

mentre la stessa annotazione, rappresentata in modalità stand-off, può essere salvata separatamente dal testo, riportandone la posizione:

annotazione1 [start=2, end=3]

Le annotazioni inline hanno il vantaggio di essere leggibili facilmente da un umano direttamente nel testo: un caso d'uso abbastanza diffuso nel caso si abbia un annotatore umano che opera sul testo. Questa esigenza ha però meno peso nel caso si operi con strumenti informatici di editing delle annotazioni su testo, che rendano trasparente all'operatore le operazioni di salvataggio delle annotazioni. Le annotazioni inline possono chiaramente essere gerarchiche, ma ovviamente non possono sovrapporsi. Proprio questo punto costituisce il vantaggio della metodologia stand-off, la quale permette di trattare qualunque tipo di annotazione a diversi livelli: le annotazioni diventano oggetti trattabili in modo indipendente, per quanto legato, al testo.
La metodologia stand-off segue inoltre nello spirito la separazione, avvenuta negli anni '90, tra markup e contenuto nella presentazione grafica delle pagina web: il contenuto, reso in HTML, è separato dal layout, espresso in fogli CSS[3].

[2] Vedi sezione 3 di Bański, Piotr. "Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless", in *Proceedings of Balisage: The Markup Conference 2010*. Balisage Series on Markup Technologies, vol. 5 (2010). URL: http://www.balisage.net/Proceedings/vol5/html/Banski01/BalisageVol5-Banski01.html

[3] In realtà l'HTML è comunque costituito da tag inline, il cui stile è però definito nei CSS. L'analogia citata è imprecisa, ma rende l'idea di una tendenza nel settore a separare il puro contenuto testuale da come il contenuto è presentato.

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Una disamina critica di come il TEI permetta l'uso di annotazioni stand-off, anche se nell'ambito più specifico delle risorse linguistiche, è fornita dal già citato Banski 2010.

Il dibattito sulle metodologie inline e stand-off si interseca con il dibattito sull'uso tout court di TEI nel caso delle edizioni digitali: secondo alcuni l'approccio "all in one" di TEI andrebbe abbandonato per optare per approcci in cui testo e annotazioni siano contenuti in più asset (files). Un difensore di questa linea è Schmidt 2014[4].

Passiamo ora a esaminare i modelli dati e gli standard usati dalle applicazioni web e dagli strumenti citati nel D1.1.

Modelli dati e standard nelle edizioni digitali

TEI/XML

Come ci si potrebbe aspettare, e come è evidente dai numerosi esempi discussi nel D1.1, lo standard TEI è molto usato nelle edizioni digitali prese in considerazione, il che garantisce una certa interoperabilità delle opere digitali, con alcuni casi in cui però i dati non sono rilasciati al pubblico e altri casi in cui la rappresentazione dei dati è custom. È da osservare però che l'interoperabilità di TEI/XML può essere messa in discussione (si veda il già citato Schmidt 2014): in particolare TEI/XML garantisce un'alta interoperabilità del formato ma una scarsa interoperabilità semantica, per via delle eccessive possibilità di scelta che l'annotatore umano ha rispetto al complesso sistema di tag TEI. Schmidt 2014, Renear 2000 [5] e Durusau 2006[6] mostrano come, oltre all'enorme numero di modi possibili di usare il TEI per l'encoding di semplici annotazione, un problema ancor più cogente è che il TEI, includendo le annotazioni nel testo, non porta a rappresentare adeguatamente la loro soggettività.
Queste considerazioni non vogliono portare a una demonizzazione di TEI, che è uno standard molto usato e garantisce una base comune a cui molti progetti di editoria digitale si riconducono, ma serve a introdurre un secondo importante data model per annotazioni (in particolare per annotazioni sul web) nato nel contesto delle comunità afferenti al World Wide Web Consortium (W3C) e al Semantic Web: il Web Annotation Data Model.

[4] Desmond Schmidt, «Towards an Interoperable Digital Scholarly Edition», Journal of the Text Encoding Initiative [Online], Issue 7 | November 2014, Online since 01 January 2014, connection on 22 September 2015. URL : http://jtei.revues.org/979
[5] Renear, Alan. 2000. "The Descriptive/Procedural Distinction Is Flawed." Markup Languages: Theory & Practice 2(4): 411–20.
[6] Durusau, Patrick. 2006. "Why and How to Document Your Markup Choices." In Electronic Scholarly Editing, edited by Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, 299–309. New York: MLA.

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



TEI Core

Il formato TEI, che attualmente si rifà alle linee guida denominate P5[7], è diventato sempre più complesso nel corso degli anni per venire incontro a molte esigenze diverse. Per questo motivo le linee guida individuano un core, comune a tutti i documenti TEI, e dei moduli, utilizzabili solo per casi specifici.

Il core TEI[8] definisce innanzitutto dei tag elementari, in grado di delimitare paragrafi, punteggiatura, annotazioni che mettono in risalto aspetti del testo (quali enfasi, informazioni tipografiche, …) o citazioni di parti del discorso.
Si trovano metodi standard per marcare liste, note a margine e indici.
Una parte interessante per il nostro studio riguarda anche l'annotazione di semplici interventi editoriali (quali errori apparenti, aggiunte, cancellazioni, …). Per interventi editoriali più complessi, tipici delle edizioni critiche, invece si rimanda al modulo specifico.
Il core TEI comprende anche alcune marcature per delimitare aree del testo contenenti nomi propri, date, numeri abbreviazioni o indirizzi.
Inoltre si trovano delle marcature per segnalare collegamenti esterni, riferimenti incrociati, note bibliografiche e riferimenti ad altre opere.
Completano il core anche delle marcature di tipo più specifico, utilizzate per denotare versi di poesia e frammenti di opere teatrali.
Tutte le marcature sono corredate da metadati caratteristici che aiutano a cogliere aspetti più profondi del testo.

TEI Fonti Primarie

Il modulo chiamato *Representation of Primary Sources*[9] è utilizzato per collegare alla trascrizione digitale la fonte primaria, come per esempio un manoscritto. Questo modulo è pensato per le edizioni digitali diplomatiche.
Il metadato più importante di questo modulo è probabilmente il *digital facsimile*, cioè quella risorsa, tipicamente un file fotografico digitale, che raffigura l'opera o una sua parte.
Solitamente vengono fornite a corredo di un'edizione digitale la scansione delle pagine della sua controparte cartacea.
Un altro strumento importante per lavorare con i facsimile è l'identificazione di alcune porzioni di immagine, delle quali può essere fornita una trascrizione digitale (embedded trascription).
In questo modulo sono previste anche delle marcature per specificare l'impaginazione del testo, tra le quali intestazione e piè di pagina.
Completano il modulo una serie molto specifica di marcature e metadati che consentono di segnalare nel migliore dei modi tutto ciò che riguarda le fonti primarie, oltre a quelli già

[7] P5: Guidelines for Electronic Text Encoding and Interchange - http://www.tei-c.org/Vault/P5/2.9.1/doc/tei-p5-doc/en/html/
[8] Elements Available in All TEI Documents - http://www.tei-c.org/Vault/P5/2.9.1/doc/tei-p5-doc/en/html/CO.html
[9] Representation of Primary Sources - http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

presenti nel core: abbreviazioni ed espansioni, correzioni e congetture, aggiunte e cancellazioni, sostituzioni, scritture a mano, frammenti danneggiati o illeggibili, e molti altri. Come si può osservare, questo modulo TEI si presta in particolare anche per edizioni genetiche, poiché fornisce strumenti di marcatura per individuare cambiamenti e revisioni del testo.

TEI Apparato Critico

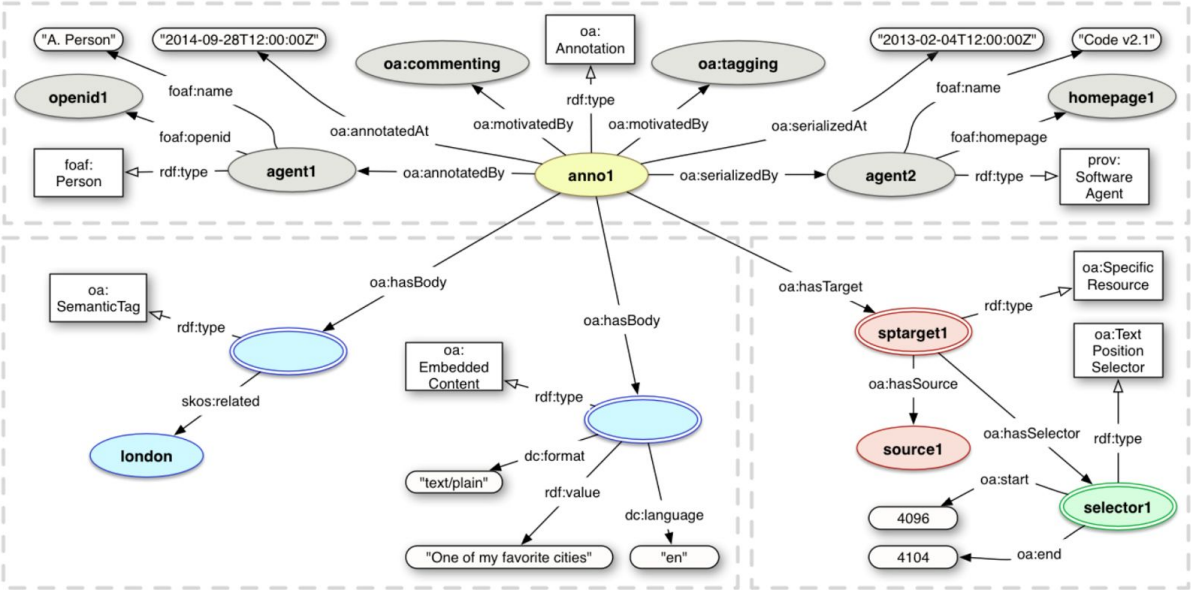
Il modulo TEI chiamato *Critical Apparatus*[10] è specializzato per la gestione di quelle informazioni che riguardano l'evoluzione del testo nelle sue varianti, o edizione critica.
Il modo proposto da TEI per codificare queste informazioni richiede di definire una serie di testimonianze (witnesses), cioè le diverse varianti che si stanno confrontando.
Quindi si può comporre un apparato (apparatus) come un elenco di letture (reading) provenienti dalle diverse testimonianze, anche frammentarie. Ogni lettura può essere differente dalle altre, seppure sia definita una relazione molto forte tra di esse.
Il modulo prevede diverse modalità per collegare il testo alle sue trascrizioni per consentire la trattazione compatta di casi anche relativamente complessi.

Web Annotation Data Model

Il Web Annotation Data Model[11] (WADM) nasce sostanzialmente dall'Open Annotation Data Model[12], elaborato da uno specifico Working Group in seno al W3C. Il WADM fa parte di una serie di specifiche per definire un'architettura per le annotazioni su web coerente con il modello Linked Data, e che al momento si trovano nello status di "public working draft", e stanno percorrendo la strada che normalmente porta le specifiche emanate da questi gruppi di lavoro a diventare raccomandazioni W3C.
Il WADM Permette di trattare annotazioni stand-off in RDF (Resource Description Framework[13]). Uno dei pregi di questo modello è che è pensato per rappresentare annotazioni non solo su testo, ma anche su immagini: le annotazioni sono infatti legate al loro "target" tramite "selettori", coordinate che nel caso del testo saranno due numeri (le posizioni su testo dell'annotazione), mentre nel caso di un'immagine saranno i punti del poligono indicante la porzione di immagine da annotare. Essendo basato su RDF (anche se il data model in quanto tale può ispirare altri formati di salvataggio di annotazioni) si possono usare le metodologie del semantic web per linkare le annotazioni a entità linked data. È inoltre semplice rappresentare la provenance dell'annotazione, ossia chi l'ha emessa e quando, raggiungendo quindi la piena interoperabilità semantica.
La figura mostra un esempio completo dalla specifica WADM, in cui una persona, usando un software, ha salvato un commento su una porzione di testo riguardante Londra dicendo che è una delle sue città preferite.

[10] Critical Apparatus - http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html
[11] http://www.w3.org/TR/annotation-model
[12] http://www.openannotation.org/spec/core
[13] http://www.w3.org/TR/rdf11-concepts

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



METS ALTO

Lo schema XML[14] che prende il nome di *Analyzed Layout and Text Object*[15] (ALTO) ha come obiettivo la descrizione dell'impaginazione (o layout) e il contenuto di risorse testuali con una controparte fisica, come pagine di un libro o di giornale. Solitamente è usato come estensione dello schema *Metadata Encoding and Transmission Schema*[16] (METS), pensato per mantenere un insieme di metadati riguardanti una libreria di oggetti digitali.

Un documento conforme al formato METS è composto da sette componenti principali.
Il primo è l'*header*, che indica metadati quali la data di creazione del documento, lo stato di completamento ed eventualmente i suoi autori.
Quindi si trovano una o più sezioni di *metadati descrittivi* riguardanti gli oggetti della libreria. Questi metadati possono essere sia riferimenti esterni ad altri documenti che dati inclusi direttamente (embedded) nel documento, siano essi in formato binario o testuale.
La sezione di *Administative Data* è composta di quattro sottosezioni: metadati tecnici (informazioni riguardo la creazione del file, dei formati usati, …); proprietà dei diritti intellettuali; informazioni sulla libreria fisica da cui provengono i dati; informazioni sulla provenienza digitale dei dati.
La sezione di *File* contiene l'elenco e i metadati relativi ai documenti che rappresentano in formato digitale gli oggetti della libreria. Questi file possono essere raggruppati all'interno della sezione, in base a qualche criterio d'interesse (per esempio: tutti i file pdf, tutti i file TEI, …).
La *Structural Map* è la sezione che contiene informazioni riguardanti la struttura gerarchica dei documenti descritti. Ogni nodo della struttura può contenere un numero a piacere di

[14] XML Schema - http://www.w3.org/XML/Schema
[15] About ALTO - https://www.loc.gov/standards/alto/about.php
[16] METS: An Overview & Tutorial - http://www.loc.gov/standards/mets/METSOverview.v2.html

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

puntatori a dei metadati descrittivi o puntatori a file. Un puntatore a file può anche indicare un frammento del file.
La *Structural Links* una sezione simile alla Structural Map, che consente di collegare documenti correlati in modo non gerarchico.
Si può inoltre trovare una sezione chiamata *Behaviour*, nella quale è possibile definire un comportamento predefinito con il quale interagire con i file descritti (per esempio si può specificare di aprire i file HTML con un browser).

ALTO, che presuppone la scansione di un documento e la sua trascrizione con tecniche OCR[17], definisce una nuova sezione per un documento METS. La radice della sezione principale contiene una o più *Description*, una o più sezioni di tipo *Style* e almeno una sezione di tipo *Layout*, ognuna delle quali corrisponde a una pagina.
La *Description* è utile per informazioni generali, tra le quali unità di misura, metadati sul file immagine di partenza e sugli strumenti di OCR utilizzati.
Lo *Style* è una sezione che contiene regole di formattazione di default per il testo.
Per ogni *Layout* si può definire una pagina, della quale vengono descritti sia la struttura che il testo con degli stili associati. Ogni pagina a sua volta è suddivisa in aree, utili per esempio per distinguere i margini dal corpo principale.
Le aree solitamente contengono uno o più blocchi di testo, ciascuno dei quali può essere posizionato con riferimenti assoluti rispetto alla pagina che lo contiene. Inoltre ogni singolo blocco di testo può portare con se informazioni riguardanti la formattazione e metadati riguardanti il riconoscimento OCR. Un aspetto interessante è l'ordinamento dei blocchi, che è implementato attraverso la possibilità, per ciascun blocco, di dichiarare un successore, tramite riferimento a un identificativo XML.

In questa breve descrizione si è avuto modo di vedere come un formato molto specifico, come METS ALTO, riesca a raggiungere l'obiettivo di rappresentare scansioni di opere in formato digitale in maniera semplice e pulita. I limiti di questo formato sono tuttavia la scarsa flessibilità a diverse esigenze, ambito nel quale TEI mostra tutte le sue peculiarità.

Architettura CITE/CTS

L'architettura CITE[18] definisce un framework per citare in modo universale documenti testuali. L'architettura è pensata per essere indipendente da tecnologie specifiche ma al tempo stesso per favorire il processamento automatico.
All'interno dell'architettura è definito un protocollo, chiamato Canonical Text Service[19] (CTS), che permette a un client di recuperare un testo, memorizzato in un repository gerarchico, a partire da un URN.

[17] Optical character recognition - https://en.wikipedia.org/wiki/Optical_character_recognition
[18] http://cite-architecture.github.io/about/
[19] http://cite-architecture.github.io/cts/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

Ispirandosi al modello gerarchico del testo, proposto dal formato OHCO2[20], lo schema di un URN CTS è così definito:

```
urn:cts:CTSNAMESPACE:WORK:PASSAGE@SUBREFERENCE?
```

Dove:

- urn: rappresenta l'identificativo del namespace (NID)
- cts: rappresenta la stringa specifica del namespace (NSS)
- CTSNAMESPACE: rappresenta un sotto-namespace all'interno di CTS
- WORK: rappresenta l'identificativo del lavoro
- PASSAGE: indica il passaggio di un testo, oppure un intervallo di testo generico (non necessariamente all'interno dello stesso passaggio)
- SUBREFERENCE: è un elemento opzionale, che serve per individuare intervalli di testo indicizzati all'interno di un passaggio

Del framework esistono diverse implementazioni, sia su database relazionale, che gerarchico (XML) o a grafo (triple store / SPARQL endpoint).

L'importanza del framework è data dall'essere considerato un modello di riferimento per referenziare in modo universale un qualunque testo nell'ambito della Linked Open Data cloud.

Architettura IIIF

La specifica International Image Interoperability Framework[21] (IIIF) è da considerarsi l'analogo dell'architettura CITE/CTS nell'ambito delle immagini digitali. Anche in questo caso, a partire da un repository di immagini, è possibile definire lo schema di un URI del tipo:

```
{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}
```

Dove:

- schema, server, prefix: sono informazioni di partenza tramite le quali localizzare le immagini su un server
- identifier: è un identificativo univoco, universale rispetto al server
- region: identifica l'area dell'immagine referenziata
- size, rotation, quality, format: sono informazioni che riguardano la scala, l'effetto di rotazione, la qualità e il formato desiderati per accedere all'immagine

[20] http://cite-architecture.github.io/ctsurn/ohco2/
[21] http://iiif.io/technical-details.html

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Anche in questo caso è possibile referenziare in modo univoco e quanto più possibile stabile immagini o porzioni di esse, favorendone perciò il riuso.

Sono disponibili alcune implementazioni dell'interfaccia, cioè dei server in grado di memorizzare le immagini digitali e rispondere alle chiamate definite dallo schema URI in modo efficiente, ma sono anche disponibili dei client in grado di sfruttare le specifiche per rendere migliore l'esperienza di navigazione di immagini. Infatti se si vuole visualizzare tutta l'area di un'immagine ad alta risoluzione risulta più efficiente richiederne al server una copia a bassa risoluzione e successivamente recuperare porzioni di essa ad alta risoluzione, solo se effettivamente richieste (tramite funzioni di ingrandimento o zoom).

Modelli dati e formati custom

Vi possono essere comunque più strategie per separare testo e annotazioni. In particolare se il formato dei dati è trasparente agli utenti, e se non è necessario, o non è voluto, importare ed esportare i dati in formati standard, anche un modello e un formato custom possono essere usati -- questo approccio può avere una giustificazione in particolare in contesti commerciali. Applicazioni verticalizzate su particolari casi d'uso potrebbero ad esempio aver bisogno di un potere espressivo molto limitato, e in questi casi conformarsi a formalismi relativamente pesanti come TEI/XML potrebbe costituire un peso più che un aiuto. Rimane altresì vero che, anche in applicazioni commerciali, rappresentare i dati in modelli e formati standard dovrebbe garantire un più semplice approvvigionamento di dati per l'applicazione stessa. È quindi bene, anche in contesti dove ha senso usare modelli e formati *custom*, porsi il problema di rendere i dati, almeno in linea di principio, importabili ed esportabili nei formati più diffusi.

La tabella seguente mostra i modelli / formati per i dati usati nei vari progetti di edizioni digitali citati in D1.1.

Cotton Nero	XML custom
Exeter Anthology	HTML
Vercelli Book	TEI/XML
Codice Pelavicino	TEI/XML
Codex Sinaiticus	TEI/XML
CBook	custom XML + OWL
Parzival Project	custom
Dante's Commedia	TEI/XML

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Mark Twain Project	TEI/XML
Petrus Plaoul	TEI/XML
Vespasiano da Bisticci lettere	TEI/XML + OWL
Proust Prototype	TEI/XML
Shelley-Godwin	TEI/XML

In un'altra tabella mostriamo i formati gestiti da alcuni tool utilizzati nella creazione di edizioni digitali (o simili), alcuni dei quali citati in D1.1.

Pundit	Open Annotation Model
Annotation Studio	TEI
Textus	custom stand-off
TextLab	TEI/XML
EVT	TEI/XML
Open Convert	da vari formati a TEI

Come è evidente e prevedibile, molti dei progetti di edizioni digitali presi in considerazione (circa la metà) usano nativamente TEI/XML, che deve essere quindi tenuto in considerazione per qualunque strumento di pubblicazione/editing su edizioni digitali che voglia essere interoperabile con il materiale già esistente.

Per quanto riguarda i tool, è interessante notare come il TEI sia presente in tool di conversione tra formati, mentre i tool che trattano l'Open Annotation Model (o equivalentemente il WADM) lo usino nativamente. Questo fatto sembrerebbe suggerire (anche se la previsione potrebbe essere azzardata, basandosi su una platea di utenti e sviluppatori relativamente di nicchia) che al momento l'Open Annotation Model si trovi nel suo momento emergente, mentre il TEI potrebbe trovarsi nel momento tra maturità e declino.

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Textus

Il modello dati di Textus[22][23] prevede una netta separazione tra testo e annotazioni. Il testo viene memorizzato nel repository principale come un semplice flusso di caratteri, senza alcun tipo di markup o annotazione. Le annotazioni sono quindi di tipo stand-off e hanno tutte un modello di riferiemento comune. In questo modo dato un testo, oppure una sua porzione, è possibile recuperare le sue annotazioni e utilizzarle per renderizzare una pagina HTML che si presenti come ci si aspetta.
In Textus le annotazioni sono divise in tre tipologie principali:

- tipografiche: veicolano informazioni su come renderizzare il testo (grassetto, corsivo, …)
- semantiche: veicolano informazioni sul significato del testo (bibliografia, provenienza, interpretazioni, …)
- strutturali: veicolano informazioni sulla scomposizione strutturale del testo (capitoli, senzioni, …)

Le annotazioni semantiche sono quelle di natura più eterogenea, che rispecchia in qualche modo l'eterogeneità delle opere letterarie, e sono arricchite da metadati supplementari: un sottotipo, la data di inserimento nel sistema e l'utente che ha inserito l'annotazione. Il contenuto specifico definito dal sottotipo di annotazione è invece incapsulato in un payload in formato JSON. I sottotipi delle annotazioni semantiche sono diversi e, alcuni esempi sono:

- BibJSON: note bibliografiche, secondo lo schema definito dal formato BibJSON[24]
- source: un URL per indicare la fonte dell'informazione, per esempio l'immagine che rappresenta la scansione di una pagina, oppure la registrazione audio
- scene: delimitano una porzione di testo identificata da un luogo, una collocazione temporale e un insieme di attori (solitamente usata per testi teatrali)
- letter: utilizzato per annotare porzioni di testo che corrispondono a una lettera

Anche le annotazioni di testo libero da parte degli utenti rientrano nelle annotazioni semantiche.
Rispetto al modello definito da Open Annotation emergono due limitazioni abbastanza importanti. Innanzitutto le annotazioni semantiche di Textus non possono fare riferimento a entità, rendendo le annotazioni prodotte significative solo per un lettore umano e non da agenti software come previsto dalla vision di Semantic Web. Inoltre il target di un'annotazione Textus, cioè l'oggetto annotato, può essere solamente una porzione di testo composto da caratteri, escludendo perciò la possibilità di annotare documenti multimediali, contenenti per esempio immagini.

Le annotazioni strutturali sono modellate principalmente attraverso un attributo rappresentato da un numero naturale, che indica il livello di annidamento all'interno del testo

[22] Textus Text Format - http://okfnlabs.org/textus/doc/textus-format.html
[23] Textus JSON Import Format and Textus Basic Profile - https://github.com/okfn/textus/blob/master/docs/json_import_format.md
[24] BibJSON Data Format - http://okfnlabs.org/bibjson/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



(depth). Per esempio un libro può essere rappresentato dal livello zero (0), un capitolo dal livello uno (1) e una sezione dal livello due (2). Il numero di livelli è da considerarsi arbitrario. In questo caso la relazione gerarchica tra annotazioni strutturali non è esplicita nei dati (si ricordi che Textus non rappresenta relazioni) ma è inferita a partire dalla porzione di testo identificata e dall'attributo di profondità.

CBook

Il formato utilizzato nel CBook è stato fortemente influenzato da TEI XML, che si è scelto di non implementare nella sua interezza per diversi motivi.
Il primo criterio guida per il formato CBook è stato quello della semplificazione del numero di tag a disposizione e di una nomenclatura quanto più possibile vicina al dominio dell'applicativo sviluppato. Questo aspetto ha velocizzato l'apprendimento del formato da parte di collaboratori che hanno digitalizzato le opere.
In particolare sono stati esplicitati in modo chiaro i concetti di Sequenza e di Capitolo, due suddivisioni strutturali di un'opera non in relazione tra di loro. Il limite attualmente più evidente di questo approccio è che la struttura di un CBook può basarsi soltanto su questi due livelli.
Ogni elemento contiene frammenti di testo che possono essere definiti di lunghezza arbitraria e, grazie a un sistema di tipizzazione, possono essere associati sia a informazioni tipografiche (font, allineamento del testo, …) che di tipo semantico (battute di dialogo, date, fime, …). Sia i frammenti testuali che quelli strutturali possono essere associati a relazioni semantiche, cioè aventi per oggetto altre entità del dominio: per esempio le battute di dialogo sono relazionate con il personaggio che le pronuncia.
Le informazioni tipografiche sono contenute in un file separato nel quale è possibile definire regole CSS[25]. Questo permette una separazione molto chiara tra ruolo di annotatore, solitamente un esperto letterario, e addetto all'impaginazione.
Vi è inoltre un'altro file, conforme al formato OWL[26], nel quale è possibile definire le entità di dominio dell'opera, come personaggi o luoghi, e le relazioni che intercorrono tra essi. Gli individui definiti all'interno di questo file possono essere referenziati dal file xml principale. Anche questa separazione tra annotazione del testo ed entità dell'opera consente di suddividere il lavoro in due fasi: una macroscopica, nella quale vengono individuati gli attori e le location significative, e una più dettagliata, nella quale ogni componente del testo viene arricchito di informazioni utili per migliorare l'esperienza del lettore dell'edizione digitale.
Il formato prevede inoltre altri tre file in formato tabulare (CSV) che serializzano altre informazioni inserite dagli utenti del sitema e pertanto sempre associate a uno userid e a un timestamp: le annotazioni testuali, i segnalibri e i contributi. Di questi le annotazioni e i segnalibri sono da considerarsi come annotazioni stand-off sul testo principale, mentre i contributi rappresentano metadati per collegare un URL a entità dell'opera (capitoli, personaggi, …) oppure a una breve descrizione testuale.

[25] Cascading Style Sheets - http://www.w3.org/Style/CSS/
[26] OWL Web Ontology Language - http://www.w3.org/TR/owl-features/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



TEI e Linked Data

Clavius on the Web

Il progetto Clavius on the Web ha un modello dei dati a granularità molto fine, che arriva fino ad individuare ogni singolo token del testo come un'individuo.
A partire da un testo TEI, tramite strumenti automatici, viene assegnato un URN conforme allo standard CTS[27] per ogni token. Ogni token viene quindi annotato da un punto di vista linguistico assegnando cioè una categoria lessicale, un lemma di riferimento e delle feature morfologiche a partire da un lessico conforme al modello SIMPLE.
Quindi viene fatta un'annotazione semantica, intesa come una classificazione in alcune classi d'interesse (Person, Place e Letter) e, quando possibile, un collegamento con un individuo della Linked Open Data cloud. La risorsa utilizzata per il linking dei personaggi è dbpedia[28] e, nel caso si tratti di autori, VIAF[29].
Oltre ai token linguistici sono identificati come individui anche i documenti del corpus, che corrispondono a delle lettere. Di ogni lettera vengono espressi in formato RDF alcuni metadati, come la data di stesura, l'autore, il destinatario, il luogo di provenienza e il titolo. Questi metadati sono tutti espressi utilizzando vocabolari standard, come BIBO[30], FOAF[31] e Dublin Core[32]. Per quanto non chiarito esplicitamente, anche questi metadati risultano estraibili in modo automatico a partire da un documento TEI annotato correttamente.

Codice Pelavicino Edizione Digitale

Questa edizione digitale è da considerarsi come uno dei punti di riferimento per quanto riguarda l'annotazione esclusivamente con XML TEI.
Le risorse annotate sono strutturate come un insieme di file, a partire da un file XML in grado di definire la struttura principale dell'opera richiamando, tramite il meccanismo di xinclude[33], altri file contenenti i componenti testuali codificati in formato XML TEI P5.
La libreria è caratterizzata da alcuni metadati testuali riguardanti informazioni di tipo generale ed è suddivisa in libri.
Ogni libro a sua volta è suddiviso in pagine, per le quali è fornita una trascrizione e la scansione a partire da un documento originale. Ogni libro è inoltre corredato di un regesto in formato testuale.
Le pagine sono descritte dal loro numero e dall'informazione di recto/verso, cioè dritto o rovescio. Il testo all'interno di ciascuna pagina può essere annotato con informazioni semantiche, che riguardano in particolare: persone, luoghi, ruoli (o mestieri) e unità di misura (principalmente monete considerata l'opera specifica).

---

[27] A Gentle Introduction to CTS & CITE URNs - http://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html
[28] http://wiki.dbpedia.org/
[29] https://viaf.org/
[30] http://bibliontology.com/
[31] http://xmlns.com/foaf/spec/
[32] http://dublincore.org/specifications/
[33] http://www.w3.org/TR/xinclude/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Vespasiano da Bisticci, Lettere

Si tratta di un corpus di lettere manoscritte delle quali vengono modellati i tipici metadati d'interesse, quali materiali, copisti, autori e via dicendo.
Ogni documento è codificato in formato XML TEI.
La particolarità è data dall'utilizzo dell'attributo ref in diversi elementi TEI. Grazie a questo attributo infatti è prevista la possibilità di collegarsi ad altre risorse, che possono essere anche risorse RDF, o individui in un'ontologia della Linked Open Data cloud. Ove possibile sono state usate risorse open, altrimenti sono stati definiti dei file RDF specifici contenenti i dati necessari. I dati attualmente non sono pubblicati e pertanto non è possibile fornire ulteriori informazioni sul modello, oltre a quelle che si evincono dalla documentazione ufficiale.
In particolare i link utilizzati per il progetto riguardano: persone, luoghi, citazioni interne ed esterno, eventi storici, lessico tecnico (associabile a un tesauro).
Le singole lettere non sono dotate di URI ma solamente di un Permalink, cioè un URL con garanzia non cambiare nel tempo.

Burckhardtsource

La documentazione a riguardo del modello dei dati della libreria digitale Burckhardsource è molto ricca[34]. Vengono forniti dettagli riguardo a tutti gli elementi TEI utilizzati per descrivere sia l'edizione filologica[35] (diplomatica) che quella semantica[36].

L'edizione filologica è caratterizzata, come ci si può aspettare, da tutti quegli aspetti tipici della marcatura di edizioni diplomatiche, come le regolarizzazioni (correzioni nel testo, sillabazione, …), dichiarazioni di caratteri illeggibili, e via dicendo. Inoltre si trovano informazioni riguardanti l'impaginazione: posizione del testo (indentazione, pedice, apice) e allineamento; caratteristiche grafiche (sottolineature, colorazioni del testo, …). Questa versione include anche le scansioni delle pagine dei documenti originali.

L'edizione semantica in particolare evidenzia aspetti quali l'autore del testo manoscritto, la presenza di menzioni: di luoghi, persone, opere d'arte e riferimenti bibliografici. L'aspetto interessante delle menzioni è che queste sono collegate a individui appartenenti alla Linked Open Data cloud che, nella maggior parte dei casi, permette di arricchire automaticamente le informazioni di descrizioni, o immagini. In particolare i dataset più usati risultano essere Freebase[37] (per luoghi e persone) e bildindex (per le opere d'arte). I riferimenti bibliografici invece risultano tutti definiti ad hoc per il progetto.

[34] Project Documentation - http://wiki.burckhardtsource.org/
[35] http://wiki.burckhardtsource.org/protocols/philological-protocol/
[36] http://wiki.burckhardtsource.org/protocols/semantic-edition-protocol/
[37] https://www.freebase.com/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Per ogni elemento sono forniti i dettagli in termini di trasformazioni XSLT per la conversione dei documenti XML in HTML, tramite software EVT.

Inoltre vengono fornite le specifiche dei metadati riguardanti ogni singolo documento della libreria[38] (nel caso specifico si ricorda che si tratta di lettere). I metadati sono così organizzati: informazioni specifiche sulla corrispondenza (mittente e destinatario, luogo di scrittura e di consegna, data di scrittura), sugli aspetti fisici del documento (dimensioni del foglio, stato di conservazione) e sull'archivio di provenienza (e informazioni di proprietà intellettuale). Anche la maggior parte dei metadati è arricchita di informazioni provenienti dalla Linked Open Data cloud.

Recogito

Il modello dei dati di Recogito è impostato a partire da un insieme di opere classificate in modo non esclusivo secondo la tradizione culturale di provenienza (greca, latina, cristiana, …).
Ogni opera è dotata di un titolo, di una data e di un autore ed è collegata alla Linked Open Data cloud tramite l'URL dell'opera corrispondente su Wikidata .
Ogni opera può essere presente nella libreria in più versioni linguistiche, ciascuna delle quali è ulteriormente scomposta in parti (genericamente una parte è un Libro oppure un Capitolo) che sono effettivamente i contenitori del testo digitale. Ogni versione può essere collegata a una sorgente di riferimento, nel caso specifico a URL dello store digitale Google Play Books.
All'interno di ogni documento si possono annotare i toponimi. Visto che potenzialmente qualunque utente di Internet può annotare il testo, per ogni annotazione viene mantenuto uno stato per indicare ad esempio che è stata controllata da un esperto di dominio, oppure che è stata riconosciuta in modo errato da un software di Named Entity Recognition. Ogni annotazione viene collegata ad un URI nel dataset di Pleiades[39], la risorsa Linked Data per mantenere informazioni riguardanti luoghi antichi e che a sua volta fornisce svariate informazioni la più importante della quale è la localizzazione geografica.

Contesti d'uso delle edizioni digitali

In questa sezione descriviamo i contesti d'uso delle edizioni digitali prese in considerazione nel presente studio (descritte nel documento D1.1). Essendo edizioni scientifiche create e/o rivolte a studiosi (cioè SDE), la maggior parte delle edizioni digitali analizzate in D1.1 rientrano nell'ambito della produzione editoriale scientifica online. Gli utenti a cui sono rivolte sono perciò in prima istanza i ricercatori, gli studiosi, gli studenti universitari (tra i quali filologi, storici, studiosi di scienze sociali, biblioteconomi, linguisti, ecc).
Una edizione online pubblica, in ogni caso, può essere utilizzata da chiunque abbia accesso al web, e quindi l'insieme degli utenti e dei collaboratori può ampliarsi notevolmente e

[38] http://wiki.burckhardtsource.org/protocols/metadata-protocol/
[39] http://pleiades.stoa.org/

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

comprendere un pubblico molto più ampio. In alcuni casi tra gli utenti sono previsti esplicitamente gli studenti e gli insegnanti della scuola.
Riportiamo ad es. una citazione dalla quarta edizione digitale di Beowulf:
“The Fourth Edition of Electronic Beowulf is designed to meet the needs of general readers, who require a full, line by line, translation; of students, who want to understand the grammar and the meter and still have time in a semester to study and appreciate other important aspects of the poem; and of scholars, who want immediate access to a critical apparatus identifying the nearly 2000 eighteenth-century restorations, editorial emendations, and manuscript-based conjectural restorations.” (http://ebeowulf.uky.edu/ 2015)

Si tratta di contesti d’uso nuovi potenzialmente interessanti, che iniziano a diffondersi grazie alla disponibilità della rete internet nelle scuole, ma che trovano per ora un freno, in Italia, nell’editoria scolastica tradizionale, legata ai libri di testo di carta o a esempi di ebook che riproducono in digitale i modelli cartacei.

Un contesto d’uso nuovo e molto vivace che citiamo solamente come esempio interessante è quello dei progetti di *crowdsourcing* e *citizen science* nell’ambito del *cultural heritage* (descritti ad es. in [Ridge 2014]). Alcune biblioteche, archivi, musei (ad es. New York Public Library e British Museum) e altri enti culturali stanno sperimentando forme di collaborazione con il grande pubblico per la creazione, l’annotazione, la trascrizione, la correzione di informazioni digitali online in vari ambiti.

Nel paragrafo seguente ci soffermiamo sul contesto d’uso delle SDE per studiosi, basandoci prevalentemente sul contesto italiano e sulle interviste fatte ai ricercatori.

Edizioni scientifiche per studiosi (SDE)

Una analisi dei metodi per la realizzazione e dei modi di fruizione delle SDE si trova in “Digital Scholarly Editing: Theories, Models and Methods” [Pierazzo 2015], già citato in D1.1, di cui riportiamo un passaggio dal capitolo 5 , dedicato appunto alle questioni metodologiche nella creazione di SDE :

“In a work of 2011, I have listed a series of features that an editor may want to include in her/his transcription in the creation of a documentary edition [...], and a list of parameters that that will help decide their inclusion or not in the final edition [....] Although the list of features has been modelled with a digital documentary edition in mind, the parameters apply to any type of digital editions, and are:

1. **The purpose of the edition**: which are the scholarly reasons for which an edition of a particular work or text is necessary; which gaps in the previous editions, if any, is the editor to fill? Which is the theoretical framework that guides the editorial work?

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte

2. **The needs of the users**: to which type of user, present and future, is the edition aimed, and what is the particular subset of user likely to need in order to achieve their aims?

3. **The nature of the documents**: are the documents to be edited clean, scribal manuscripts? Early printed material? Draft manuscripts? Which are the extension and the relevance of the non-verbal content? Is there more than one witness or many?

4. **The capabilities of the publishing technology**: which is the best technology, if any, able to fulfil the requirement of the edition? Which is its life expectancy? Is it sustainable?

5. **Costs and time**: to what degree are the desired outcomes actually achievable within the timeframe allocated by the funding available? Which features will require more resources than are actually worth in terms of desired output? What can be postponed to a successive phase? Which are the priorities?"

(Pierazzo 2015, Chapter 5 Work and Workflow of Digital Scholarly Editions)

Il **processo** di creazione di SDE è stato schematizzato in Rehbein and Fritze (2012), in uno schema ripreso da Pierazzo (2015), che ci sembra adeguato per la maggior parte delle edizioni analizzate:

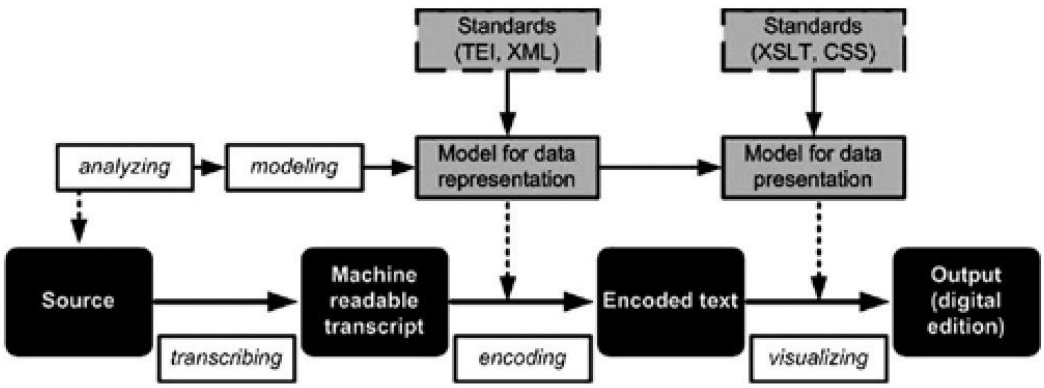


Figure 5.2 Workflow of editing (Rehbein and Fritze, 2012, p. 52)

Come trattato in D1.1, per le Scholarly Digital Editions sono richieste funzionalità di studio filologico del testo. Ci troviamo quindi nel caso d'uso del filologo digitale o Digital Humanist che crea una nuova edizione digitale solitamente partendo da un documento antico manoscritto o a stampa.
Un aspetto importante del processo è lo strumento utilizzato per creare l' edizione digitale, nelle fasi di **trascrizione ed encoding**. Abbiamo constatato che finora gli studiosi non hanno lavorato con strumenti collaborativi online, ma quasi esclusivamente con editor offline

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



e in modo asincrono. Il filologo digitale che gravita sugli schemi TEI usa, solitamente, Oxygen XMl Editor (offline), definito un “ottimo strumento dopo che lo si inizia a conoscere bene (l’UI, soprattutto delle preferenze, è un po’ confusionaria)”. La difficoltà ad usare un editor XML è un ostacolo non indifferente per chi si avvicina alle edizioni digitali e ha poche o scarse basi di un qualsiasi linguaggio di markup (o di programmazione: quest’ultimo caso ancora più improbabile per i filologi).
Se lo studioso ha un po' di confidenza con Office o con un altro editor di testo, sembra che sia più semplice farlo lavorare con gli strumenti che conosce piuttosto che fargli seguire un corso accelerato di XML. La codifica in XML in questo caso viene fatta da studenti universitari o altre persone con le competenze richieste.

Nella realizzazione dell’edizione digitale del Codice Pelavicino la codifica in XML è stata fatta dai ricercatori e dagli studenti universitari. Il responsabile del software ha fatto da "supervisore" della codifica in TEI XML: ha spiegato come dovesse funzionare il sistema delle *named entities*, come suddividere l'edizione in tanti documenti separati usando XInclude, etc. E’ stato usato Excel per le tabelle con tutte le liste di nomi di persona e le altre *named entities*, che poi sono state convertite in XML e importate nell'intestazione TEI.
Nel progetto Burckhardt Source è stato invece definito e usato un template Open Office apposito. “The idea is to use the tools of the software Open Office Writer (such as sections, styles …) to insert into the document the encoding informations to include in the xml file. There are two kinds of annotations included in the odt document:

- structural – regarding the general structure of the letter, such us pages, paragraphs, line breaks, opener and closer of a letter, notes;
- microstructures – all the particular constructions that are in the letter, such as corrections, abbreviations, text renderering and so on.

The way to show all these elements in a odt file is dividing text into sections, using styles to highlight some portion of text, adding comments.”
Il template e la procedura sono descritti nella pagina web “ODT Protocol” http://wiki.burckhardtsource.org/protocols/protocol-odt/

A chi è già esperto di edizioni digitali interessano **sistemi collaborativi online**, per la codifica dei testi. Molto spesso un approccio collaborativo è dettato/ispirato da considerazioni riguardo l’oggetto dell’edizione, ovvero se l’edizione riguarda ad es. poche poesie di un autore con buon tradizione testuale non ci sono grandi motivi per inserire feature di tipo collaborativo nell’edizione; viceversa, se la tradizione testuale è ampia, e/o i testi in questione molto lunghi, allora un edizione di tipo collaborativo/in crowdsourcing comincia a diventare molto interessante. Viene considerato interessante in particolare avere qualche strumento web-based che aggiunga strumenti collaborativi a un normale sito di edizione digitale. Viene considerato sicuramente interessante fare in modo di poter condividere annotazioni, correzioni, etc., anche fra progetti diversi.

Il filologo “tradizionale” è invece ancora alle prese con problemi generali che rendono un po’ complicato affacciarsi a questo campo da principianti assoluti: la creazione dell’edizione (per

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



mezzo di marcatura semantica o altri strumenti simili), la pubblicazione, la citabilità (necessaria per la carriera accademica), la durabilità (rispetto ad una edizione su carta, l'edizione digitale ha una durata più breve se viene abbandonata a se stessa).

Una domanda importante alla quale non siamo riusciti a rispondere con precisione è quanti sono gli **utenti** delle edizioni digitali: al solito dipende dal tipo di edizione, dal tipo di testi, dal periodo etc. In alcuni casi la risposta data dal filologo digitale è "molti meno utenti di quello che vorremmo"; già in un articolo di 10 anni fa (Robinson 2014 "Where we are with electronic scholarly editions, and where we want to be") si spiegava come un ottimo lavoro di edizione era pressoché ignorato dagli studiosi di quel settore se era disponibile soltanto in forma digitale. In altri casi ci sono migliaia di accessi al giorno, ma si tratta di grandi biblioteche digitali, archivi o musei che ospitano centinaia di testi o nuove forme di crowdsourcing.

Per quanto riguarda **i tempi e i costi** richiesti dai progetti di SDE (che sono due aspetti fondamentali per verificare la possibilità di "industrializzazione" del processo), non abbiamo raccolto purtroppo dati completi. Secondo le interviste effettuate, i tempi di realizzazione sono stati molto diversi da progetto a progetto: ad esempio la realizzazione del Codice Pelavicino ha richiesto circa un anno per arrivare a una buona edizione di più di cento documenti, mentre la realizzazione del Vercelli Book ha richiesto molto più tempo perché il realizzatore ha dovuto lavorare da solo (per la ricerca dei fondi, la scansione del manoscritto, la codifica, lo sviluppo del software di visualizzazione).
Anche i costi sono variabili: le scansioni sono ormai abbastanza economiche, in alcuni casi la codifica viene fatta con l'aiuto di studenti universitari, e quindi il budget per queste prime fasi può essere basso. Viceversa, se interessano feature particolari per la visualizzazione, la ricerca o altro, oppure se si vogliono fare scansioni multispettrali, restauro digitale, etc., i costi sono almeno di un ordine di grandezza più alto.

Pur tenendo conto delle possibili diverse esigenze di un progetto rispetto all'altro, è emersa l'esigenza di stabilire un "processo industriale" di creazione delle SDE, nel senso di un workflow efficiente, basato su condivisione e salvataggio nel cloud dei dati dell'edizione (versioning e backup), per arrivare alla fase finale di pubblicazione senza dover tornare indietro e correggere/cambiare i dati.

Per quanto riguarda l'**annotazione semantica**, secondo le nostre ricerche ai principianti del digitale sembra già un buon risultato riuscire a pubblicare qualcosa in HTML. A chi ha un po' più di confidenza va bene usare la TEI. Chi invece è al forefront della ricerca si occupa di quelle che saranno cose "normali" presumibilmente fra alcuni anni: Linked Data, ontologie, edizioni collaborative. Anche se dal punto di vista tecnologico sono strumenti disponibili e abbastanza consolidati, nella pratica toccano un numero ancora ristretto di persone (tra i filologi o altri studiosi). Nel documento D1.1 abbiamo presentato gli esempi più rilevanti in ambito non solo italiano.

Progetto finanziato nell'ambito del POR FESR 2007/2013 della Regione Piemonte con il concorso di risorse comunitarie del FESR, dello Stato Italiano e della Regione Piemonte



Per quanto riguarda i **dispositivi hardware** in cui le edizioni digitali esistenti sono usabili (PC, tablet, smartphone, LIM), la maggior parte funzionano bene solo con un PC, perché non sono state progettate e sviluppate per dispositivi *mobile*, anche se sono considerati un veicolo di pubblicazione e fruizione utile e interessante.

Le **funzionalità** necessarie per l'intero processo di creazione e pubblicazione sembrano essere:

- possibilità di vedere velocemente i risultati a partire da un documento TEI
    - senza competenze di web / programmazione
- possibilità di condividere velocemente i risultati tramite Web
- modalità di visualizzazione:
    - solo testo
    - solo immagini
    - confronto diretto tra immagine e testo
    - confronto diretto tra edizioni diverse dello stesso testo
- strumenti di analisi approfondita di immagini:
    - lente d'ingrandimento
    - annotazione grafica delle immagini
    - visualizzazione delle annotazioni direttamente sulle immagini
- supporto per le edizioni: diplomatiche, critiche, e interpretative

Le funzionalità, il modelli dei dati e l'architattura funzionale verranno descritti nei documenti successivi (D2.1 e D2.2).