Le tecnologie informatiche per l'arricchimento dell'offerta museale: soluzioni adottate per il MUSEO GALILEO e per il progetto SMARTCITY

Eva Sassolini
Seminaio di Cultura Digitale
25 maggio 2015

Indice

Premessa

Prendendo spunto dal seminario “Il Museo di Storia Naturale dell'Università di Pisa: storia e prospettive”, tenuto dal prof. Roberto Barbuti il 20 maggio 2015, ho deciso di presentare una relazione su alcuni dei temi toccati durante il seminario e che riguardano le nuove possibilità offerte dalle tecnologie informatiche per l’arricchimento dell’offerta museale. Il concetto di museo si è trasformato nel tempo e si è aperto ad un nuovo approccio ai materiali esposti e ai percorsi da costruire. L’offerta museale deve sempre più rapportarsi con quello che sta fuori del museo, come il territorio e la rete museale ivi presente, ma anche guardare con occhi nuovi quello che c’è dentro. Arricchire il museo di nuove risorse a corredo delle varie esposizioni, spesso non è sufficiente. Quando è possibile si cerca di aggiungere risorse che sono nella disponibilità del museo, ma non sono parte delle sale espositive e studiare per queste le migliori modalità di accesso. Conosco il contesto per aver lavorato in progetti che avevano obiettivi molto simili a quelli descritti durante il seminario, limitatamente però all’ambito delle risorse digitali (il sito web del museo e le applicazioni software per audio-guide), che rappresentano il mio settore di interesse.

Introduzione

L'obiettivo di questa relazione è mostrare metodologie e tecniche applicate nei progetti ai quali ho lavorato e che riguardano i contenuti testuali che corredano, a vari livelli, le diverse attività del museo dell'era digitale. Le mie competenze molto specifiche mi pongono in una posizione relativamente marginale rispetto alle attività di un progetto. Allo stesso tempo però mi danno la possibilità di confrontarmi con aspetti diversi, calandomi una volta nella veste di possessore del bene museale a fianco di un'istituzione pubblica, ma anche in quella di proponente soluzioni museali, lavorando con aziende private del settore. Da questa mia prospettiva privilegiata ho potuto analizzare alcuni dei pregi e dei difetti più comuni dei diversi approcci. Per brevità mi concentrerò su due esperienze lavorative che guardano l'ambito museale da angolazioni diverse. Descriverò il lavoro fatto nell'ambito dell'accordo di collaborazione scientifica tra l'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) e il Museo Galileo - Istituto e Museo di Storia della Scienza di Firenze, per la conservazione, valorizzazione e sfruttamento del grande patrimonio testuale storico scientifico del museo. Proseguirò poi con la descrizione delle attività svolte nell'ambito del progetto "Smartcity: Nuove soluzioni di ingegneria dei contenuti e di *ambient intelligence* a supporto del turismo culturale di esperienza", il cui coordinatore era una società di *Information Technology* specializzata nei servizi per l'innovazione culturale.

Pur essendo il tema lo stesso l'approccio e le finalità sono diverse. Da un lato i musei e le istituzioni culturali hanno consapevolezza del valore dei propri beni, ma spesso non hanno il know-how per affrontare in modo innovativo le offerte culturali da proporre. Partendo da questa situazione, che può rappresentare un limite oggettivo, hanno però la possibilità di investire nella specializzazione delle soluzioni, realizzando sistemi *ad hoc* molto efficaci. Per contro le aziende che producono software per i beni culturali mirano alla replicabilità del modello, alla portabilità e riuso degli strumenti realizzati, ma spesso sviluppano prodotti ibridi che trattano contenuti culturali a fini turistici. Il loro scopo principale, infatti, è acquisire un migliore livello di competitività sul mercato e aggredire i settori di business nel campo dei supporti interattivi e dei servizi digitali, che il turismo culturale può offrire. Con l'intento di dare un contributo metterò in luce alcune differenze tra le strategie adottate nei due casi presi in esame, facendone un'analisi relativamente al mio specifico ambito di competenza.

Museo Galileo

Sono responsabile scientifico per ILC, presso il quale lavoro, della collaborazione scientifica tra lo stesso istituto e il Museo Galileo di Firenze. Le attività del museo sono diversificate e hanno molti fronti ambiziosi, da anni il museo sta lavorando all'arricchimento della sua offerta museale sia fisica che

virtuale. Le attività che mi hanno coinvolta riguardano la biblioteca di Galileo, che è accolta nello stesso edificio in cui ha sede il museo. Si tratta di oltre 500 opere che Antonio Favaro[1], il massimo studioso di Galileo, ha individuato come appartenute alla biblioteca privata del grande scienziato. La raccolta contiene tutte le sue opere, ma anche quelle di scienziati contemporanei che Galileo stesso ha commentato. La biblioteca include inoltre un immenso carteggio riguardante tutta la corrispondenza che Galileo ha intrattenuto con le più grandi personalità del suo tempo. Nel corso degli anni si sono susseguiti molti progetti di ricerca che hanno visto protagonista la figura e l'opera di Galileo e hanno permesso al museo di riversare in una banca dati, molti dei testi appartenenti al catalogo della biblioteca. Per offrire allo studioso la possibilità di accedere a questo grande patrimonio culturale, attraverso moderne modalità di accesso (per parola chiave, autore, titolo, editore, data ecc.), è stato necessario impostare un articolato progetto di recupero e valorizzazione dei testi. Questi ultimi, pur in formato digitale, avevano nella maggior parte dei casi formati e/o codifiche obsoleti. Redatti nell'arco di decenni con software diversi e ormai inutilizzabili, contenevano però importanti annotazioni morfo-sintattiche e un grande apparato critico. Lo studio del materiale ha sconsigliato una soluzione commerciale al

[1] A. Favaro, La libreria di Galileo Galilei, in «Bullettino di bibliografia e di storia delle scienze matematiche e fisiche», XIX, 1886, pp. 219-293 e successive appendici del 1887 e 1896.

problema, la varietà dei formati e la mancanza di specifiche per le codifiche inserite nei testi, non consentivano di pensare ad un processo di recupero che andasse bene per tutti i testi. Inoltre l'interpretazione delle diverse annotazioni richiedevano sperimentate competenze nel settore. Così è nata la collaborazione con ILC. L'Istituto fu tra i primi centri in Italia a sviluppare strumenti per l'annotazione dei testi e per la gestione di apparati critici. Già dagli anni '60, in tutto il mondo, iniziarono a svilupparsi tecnologie basate sull'uso del computer nel campo della filologia, della lessicologia, della lessicografia e delle scienze umane in genere. L'Italia, e Pisa in particolare, si dimostrarono all'avanguardia in questo settore, dapprima con la creazione di una sezione linguistica all'interno del primo e più grande centro di calcolo elettronico italiano, il CNUCE, e successivamente, in seno al Consiglio Nazionale delle Ricerche, con l'Istituto di Linguistica Computazionale. ILC si è sempre posto come punto di riferimento per la comunità scientifica nazionale ed internazionale per lo studio e la realizzazione di procedure per l'analisi automatica dei testi e di materiale lessicale. Effettivamente molte delle procedure di digitalizzazione di testi fatte all'epoca sono state sviluppate con il suo supporto tecnico-scientifico. Oggi, pur non potendo più disporre delle persone che fisicamente produssero quei testi, sono comunque ancora disponibili sia la memoria storica che le competenze tecniche per affrontare un lavoro di recupero così complesso.

Sin dalle mie prime collaborazioni con ILC ho lavorato in questo settore e ho potuto apprendere negli anni le procedure di decodifica dei complessi sistemi di

acquisizione utilizzati, quando la Linguistica Computazionale era un lavoro da pionieri [vedi rif.1].

La procedura di recupero

L'accordo di collaborazione scientifica prevedeva la definizione di un formato di rappresentazione XML/TEI che tenesse conto, da un lato della tipologia di annotazioni presenti nei testi di partenza, dall'altro delle analisi ed elaborazioni a cui i testi convertiti avrebbero dovuto essere sottoposti. Valutando le codifiche che erano state utilizzate per i testi del museo, abbiamo ideato un'articolata procedura ad hoc di recupero e standardizzazione dei testi, insieme al gruppo di ricercatori e tecnici del museo. Si è trattato di analizzare le caratteristiche dei testi e individuare, ove possibile, tipologie simili, in grado di essere trattate con una procedura semi automatica. Lo studio ha prodotto cinque grandi gruppi in cui suddividere il corpus testuale:

1. testi non lemmatizzati[2] con apparato critico, sia in lingua italiana che in lingua latina;
2. testi lemmatizzati in lingua italiana;
3. testi lemmatizzati in lingua latina;

---

[2] La fase di classificazione lessicale e morfo-sintattica di un testo è detta lemmatizzazione. Lemmatizzare significa associare ad ogni singola parola l'elemento lessicale (lemma) e la classificazione morfosintattica che descrivono il ruolo che la parola ha nel particolare contesto.

4. carteggio non lemmatizzato;
5. carteggio lemmatizzato

Le differenze maggiori tra i cinque gruppi sono dovute principalmente al periodo in cui sono stati prodotti e alle tecnologie informatiche coeve. Per i testi lemmatizzati in lingua latina, abbiamo riscontrato una maggiore rigidità nella classificazione morfo-sintattica e una maggiore percentuale di errore nella rappresentazione di fenomeni quali sigle, numeri e citazioni in altra lingua. Probabilmente il sistema di lemmatizzazione assistita per il latino, più vecchio di quello per l'italiano, risentiva di un approccio "artigianale" al testo. Non va dimenticato infatti che originariamente il file digitale non era un prodotto finale, ma rappresentava un formato intermedio di lavoro, propedeutico alla formazione del definitivo formato cartaceo, tipicamente destinato al controllo e alla revisione manuale, prima del passaggio alle procedure di stampa. Il carteggio infine ha richiesto procedure di recupero dedicate, perché unisce la complessità degli altri testi ad una struttura organizzata in campi, senza però avere la standardizzazione di un database. All'interno dei gruppi esistono ulteriori differenze, che dipendono sia dalla codifica adottata che dalla particolare tipologia dei testi: trattati, dialoghi, lettere, postille, frammenti, etc. Nella tabella 1 è sintetizzata la situazione dei testi appartenenti al secondo gruppo individuato (testi lemmatizzati in lingua italiana).

Testo in input	(FE) Fasi di elaborazione necessarie per il recupero	Tipo di codifica delle specifiche annotazioni	Metadati
Testi in un formato digitale obsoleto	FE>2	Tipo di codifica obsoleta	Recuperati grazie ai dati cartacei a disposizione del museo
Testi digitali con formato dei caratteri obsoleto	2<FE<3	Codifica legata ai specifici programmi software che hanno prodotto il file	Ottenuti incrociando le informazioni contenute nel testo digitale con quelle presenti sulla versione cartacea
Testi digitali	Una sola FE	Codifica storicamente utilizzata da ILC	Recuperate direttamente dai testi digitali

Tabella 1

Ad un livello più profondo di analisi sono state poi esaminate le annotazioni. Per esempio per i testi lemmatizzati sono state riconosciute tre casistiche ricorrenti per le associazioni forma-lemma:

1. ad una forma è associato un lemma (caso proto tipico) (Rapporto 1:1);
2. ad una forma sono associati più lemmi (Rapporto 1:n);
3. più forme sono associate a un unico lemma (Rapporto n:1)

Il caso 2 riguarda le parole ortografiche morfologicamente complesse (ad es. forme verbali con clitico, preposizioni articolate), in cui la parola ortografica è segmentata nei suoi elementi costitutivi. Ad esempio, nel caso di forme verbali

con clitico, due o più parole morfologiche fanno riferimento alla stessa forma ortografica, ne è un esempio la forma "inserirvelo". Il caso 3 è invece relativo alle espressioni polirematiche[3]. In questo caso, la sequenza di parole ortografiche che compongono l'espressione poliremática è annotata come un'unica parola morfologica. Nelle opere più antiche è forte la presenza di avverbi e locuzioni come composizione di due parole ancora separate. Per uniformità alla redazione di tutto il catalogo le lemmatizzazioni che si sono susseguite nel tempo hanno cercato di uniformare il lessico, inserendo sempre il lemma come unica parola. Per esempio "non ostante" con lemma "nonostante", "in somma" con lemma "insomma", "tal ora"/"talora", etc. Esistono poi casi in cui gli studiosi hanno indicato interpretazioni più elaborate per esempio:"eccetto che" come lemma di "eccetto".

Non è oggetto di questa relazione entrare nel dettaglio di tutte le valutazioni che hanno guidato la scelta del formato di rappresentazione dei dati, ma solo le modalità con le quali sono state affrontate di volta in volta le diverse problematiche. Eviterò quindi di soffermarmi oltre, sulla natura e complessità del compito di ricostruzione delle annotazioni. Va detto però che questo lavoro ha richiesto valutazioni e analisi provenienti da competenze diverse. Il progetto voleva salvaguardare e preservare i testi per il futuro, ma anche renderli

[3] Parole composte formate da più elementi che costituiscono un insieme non scomponibile, il cui significato complessivo è autonomo rispetto ai singoli costituenti.

navigabili con moderni sistemi di *Information Retrieval*. Il rigore filologico doveva essere mediato dalla necessità di consentire una consultazione efficace. Un caso tipico è stato quello legato alla sillabazione (hyphenation): in cui la volontà di mantenere il formato pagina originale, con la sillabazione a fine riga, avrebbe impedito al sistema di interrogazione di produrre risultati esaustivi, se il fenomeno non fosse stato riconosciuto e trattato in fase di indicizzazione.

Ogni fase del progetto è stata onerosa in termini di tempo e costo, ma l'aspetto più impegnativo è stato il riassetto continuo tra le diverse esigenze, ogni volta che emergeva un problema. Un gruppo di lavoro ampio e articolato nelle competenze, reagisce più lentamente a qualsiasi modifica alla tabella di marcia di un progetto, tanto è vero che il lavoro non è ancora terminato e molti dei nodi non sono ancora stati sciolti definitivamente: per esempio la lemmatizzazione con presenza di ambiguità nella codifica.

Analizzando quello che è stato fatto e quanto ancora resta da fare è facile capire che solo le forti motivazioni del museo potevano superare le enormi difficoltà del lavoro. Per me e per gli altri ricercatori coinvolti, questo progetto è diventato quasi una "missione" a servizio del patrimonio storico e culturale in generale. La prospettiva in cui possiamo inquadrare il progetto è di ampio respiro, tutti gli sforzi fatti per raggiungere l'obiettivo sono da vedersi come la ricerca di un modello di riferimento per operazioni di preservazione di dati di alto valore culturale.

Smartcity[4]

Descrivendo invece le attività relative al progetto Smartcity è chiaro come le prospettive cambiano radicalmente. Il progetto è nato come collaborazione tra un ente di ricerca e un'associazione di alcune aziende private, da sempre impegnate nel settore dei beni culturali. L'obiettivo primario era lo sviluppo di soluzioni per la progettazione e lo sviluppo di contenuti multimediali, destinati principalmente alle audio-guide di nuova generazione. Si trattava di ipotizzare scenari in cui fosse possibile una fruizione personalizzata di percorsi turistico-culturali sia fisici (nel contesto delle città d'arte) sia virtuali [vedi rif. 2, 3].

Nell'ambito del progetto sono stati realizzati prototipi di applicazioni software volte a migliorare la produttività e la versatilità dei contenuti erogabili mediante dispositivi mobili o fruibili in contesti di visita virtuale, supportata da ricostruzioni 3D VR. In generale il progetto prevedeva attività di ricerca e sviluppo, dirette alla definizione di modelli di fruizione delle informazioni, relative ai percorsi turistici in interni (musei, mostre, chiese, etc.) e in spazi esterni (percorsi cittadini, parchi archeologici, etc.) e alla definizione di metodologie per lo sviluppo, lo sfruttamento versatile e il riuso di contenuti

[4] Smartcity è un progetto finanziato dalla regione Toscana con fondi POR FESR 2007-2013 Attività 1.1 Linea di intervento "D", Bando Regionale 2008 per il sostegno a Progetti di Ricerca congiunti tra gruppi di imprese e organismi di ricerca in materia di scienze socio economiche e umane.

interattivi. Il programma di ricerca si spingeva oltre lo scenario delle attività museali e proponeva innovazioni metodologiche e tecnologiche, dirette a consentire lo sviluppo di servizi a valore aggiunto, per il comparto del turismo culturale. Strumenti che vedono quindi l'utente finale (oltre che il sistema delle imprese di settore) direttamente coinvolto, rispetto ad attività di innovazione i cui referenti espliciti o impliciti sono prevalentemente le Istituzioni culturali.

Il task[5] progettuale, al quale ho collaborato, era finalizzato alla realizzazione di strumenti per la preparazione di percorsi all'interno del museo e/o delle collocazioni territoriali ad esso connesse. Propedeutica allo svolgimento del task, era la creazione di un corpus testuale di riferimento[6] dal quale sarebbero stati estratti i contenuti per la produzione di itinerari commentati. L'individuazione di un adeguato dominio per il corpus di riferimento si è basata principalmente su valutazioni di ordine tecnico e pratico. E' infatti evidente che in questo particolare ambito, il rispetto delle tempistiche ha lo stesso valore della quantità e qualità delle informazioni da reperire. Inizialmente avevamo pensato a Firenze come località di riferimento, ma sebbene il territorio fiorentino sembrasse ad una prima riflessione il candidato, di fatto, la vastità del

[5] T2.1: Metodologie di "thorough indexing" descrittivo, semantico e topologico delle risorse contenutistiche

[6] Si definisce "corpus di riferimento" un insieme di testi raccolti con lo scopo di fornire un campione rappresentativo della lingua per quello specifico settore.

suo patrimonio storico-artistico e culturale, ha suggerito una scelta diversa e meno ambiziosa. La scelta definitiva per il dominio di riferimento è caduta quindi su un luogo certamente più piccolo e meno noto, ma ugualmente caratteristico e ricco di beni culturali di pregio: la città di Empoli e i suoi dintorni. Hanno giocato a favore di questa scelta soprattutto implicazioni di carattere temporale e logistico, infatti, la raccolta e selezione del materiale pertinente, avrebbe dovuto essere fatta velocemente e con l'impiego di poche mirate persone. Stabilito il dominio per il corpus di riferimento si è proceduto alla ricerca del materiale bibliografico necessario alla sua composizione.

Idealmente la creazione di un corpus di dominio è un'operazione che richiede attente valutazioni, se poi si ha l'esigenza di inserire diverse tipologie di testi, è necessario cercare il corretto bilanciamento. In generale la creazione di un corpus ha regole codificate in letteratura e precise distinzioni tra corpora chiusi e corpora aperti. La prima tipologia rappresenta lo standard tradizionale, in cui la quantità dei testi e delle parole è prefissata all'inizio del progetto. Questo approccio ci restituisce una sorta di "fotografia" di una lingua attraverso i testi selezionati. Si parla invece di corpus aperto quando l'intento è studiare la natura intrinsecamente dinamica di una lingua, estendendo la nozione tradizionale di corpus in uno strumento di monitoraggio linguistico. Un corpus di monitoraggio (*monitor corpus*) è una collezione "aperta" di testi che muta nel tempo, che può essere arricchito di nuovi testi solo se selezionati secondo gli stessi criteri usati per determinare la collezione iniziale. Questo tipo di corpus permette per

esempio di monitorare le dinamiche del lessico della lingua e dunque può essere usato in contesti lessicografici, come fonte di dati per mantenere aggiornati i dizionari [vedi rif. 6].

Nel nostro specifico caso questi intenti rigorosi si sono dovuti scontrare con la stretta organizzazione temporale dei task. Dovendo il corpus essere il presupposto grazie al quale si sarebbero sperimentate tutte le successive fasi, è stato necessario affrontare il problema in modo più pragmatico e riservare la definizione delle "buone pratiche" in fase di valutazione finale. Siamo quindi partiti dalle necessità alle quali il disegno del corpus doveva rispondere:

- ✓ generazione e estrazione della conoscenza linguistica, terminologica e semantica;
- ✓ compatibilità con le successive fasi di analisi di classificazione e produzione di contesti.

La selezione dei testi si è inizialmente concentrata su generi peculiari quali le guide turistiche e le monografie dedicate alla città, alla sua storia e al suo patrimonio storico-artistico ed etnografico. Abbiamo quindi dato una maggiore importanza a quei materiali che si ponessero lo stesso scopo dell'autore di una guida turistica. Questi materiali hanno il vantaggio di organizzare i contenuti in percorsi che il visitatore può realmente fare: la visita di un museo, una piazza con i suoi monumenti, etc. Questo approccio, se da un lato fornisce una buona quantità di informazioni rilevanti per i principali monumenti e opere, dall'altro non sempre propone un approfondimento storico delle notizie riportate. Nella

maggior parte dei casi ne è data solo una breve descrizione, sicuramente utile al visitatore curioso, ma non sufficiente per un autore che si appresta a creare egli stesso un itinerario turistico. Lo stesso approfondimento può essere necessario nella descrizione di un'opera d'arte, per esempio se voglio conoscere quali altre opere, dello stesso autore, trovo nel luogo che sto visitando. In questo caso sono necessarie informazioni più dettagliate di argomento storico,artistico e paesaggistico. Per rispondere a questa esigenza, accanto a testi che illustrano come visitare Empoli e i suoi monumenti, abbiamo incluso volumi sulla storia di Empoli e su quella del contado come per esempio: "Empoli il granaio della repubblica fiorentina", "La navigazione interna nella valle dell'Arno", "Le vie del vetro" etc. Oltre alle notizie storiche abbiamo selezionato anche testi di contenuto artistico e archeologico, come ad esempio "Il Pontormo", "La casa del Pontormo", "Chiese, cappelle, oratori del territorio empolese", "La città, il territorio, il porto: Empoli in età romana", etc. Sulla base di questi testi è stato generato un Training Corpus la cui dimensione, valutata in numero di parole, non superava le 300.000 occorrenze. Ad un primo esame, il materiale bibliografico raccolto si è dimostrato insufficiente a garantire un'adeguata consistenza del corpus, dato che le procedure di analisi statistico-linguistica offrono migliori risultati quando applicate a corpus di dimensioni consistenti. Ragioni di economia di progetto hanno consigliato di studiare nuove modalità di acquisizione, essenzialmente basate sulla ricerca di materiale reperibile sul web. Abbiamo quindi attinto a strategie di acquisizione di materiali testuali di dominio già sperimentate in altri progetti svolti nell'ambito dei Beni Culturali.

Per aumentare le dimensioni del corpus, una volta creato il primo nucleo di documenti di dominio, abbiamo specializzato alcuni strumenti software automatici e guidati (*crawlers* e *spiders*), per il recupero di altro materiale testuale di settore disponibile in internet. La metodologia seguita in questo caso, consente di catturare e immagazzinare notevoli quantità di contenuti digitali, orientati a diversi bisogni informativi e a vari livelli di dettaglio. Questi applicativi, sulla base di query prestabilite, recuperano documenti dal web e utilizzano filtri semantici per misurare la rilevanza dei materiali acquisiti al dominio di interesse. Si è trattato in pratica di creare dei filtri ad hoc (parole chiave e piccoli vocabolari di terminologia di dominio) per la ricerca di documenti inerenti i beni culturali nel territorio di Empoli e dintorni. L'individuazione delle "query" più significative da proporre ai vari motori di ricerca web disponibili ha richiesto fasi di studio e analisi sul training corpus [vedi rif. 4,5].

Le successive fasi di testing sui materiali raccolti hanno evidenziato difficoltà nel delimitare l'area geografica e valutare la pertinenza dei documenti acquisiti al dominio. Categorie quali sport o spettacolo talvolta non vengono ben identificate e producono del "rumore" nell'insieme dei documenti raccolti. Per migliorare la valutazione di pertinenza dei documenti, abbiamo sperimentato l'utilizzo simultaneo di più filtri semantici:

1. con funzione di attribuzione positiva di un documento al dominio;
2. con funzione di esclusione.

Per esempio, utilizzando il solo filtro per i Beni Culturali, si affiancano notizie relative alla rievocazione del Volo del Ciuco (evento folkloristico empolese) ad altre di ambito sportivo. E' seguito poi un controllo manuale sulle selezioni operate dai filtri e sulla bontà dei risultati ottenuti.

Figura 1:schermata del programma di controllo dei documenti selezionati in base ai filtri semantici impostati

A questo si sono poi aggiunti i documenti relativi alla produzione artistica del vetro di Empoli, cui abbiamo dedicato una fase di acquisizione separata.

Per poter analizzare sistematicamente i testi acquisiti, l'intero corpus è stato indicizzato con moduli e procedure DBT[7] e poi lemmatizzato utilizzando un componente linguistico, che utilizza un motore morfologico della lingua italiana. DBT è stato sfruttato nella sua grande potenzialità per fare una prima analisi del corpus e del suo contenuto. Nel caso specifico è stato utilizzato in tutte le fasi di testing, sia per valutare la composizione del corpus da costruire, sia per impostare successivamente la fase di arricchimento dei testi. La particolare eterogeneità dei materiali e la diversa origine degli stessi hanno richiesto un controllo molto capillare. Per i testi passati ad OCR il problema tipico era rappresentato dagli errori di cattiva interpretazione di alcuni caratteri per esempio:

l => per ll

i/î => per l/t

n => per r

In figura 2 è mostrata una lista di onomastiche estratte con DBT da una delle prime versioni del corpus. La possibilità di disporre di questo utile strumento ha facilitato il lavoro di normalizzazione di alcune terminologie e nomi propri, oltre a far emergere gli errori presenti nei materiali testuali acquisiti.

[7] Data Base Testuale, sistema proprietario di indicizzazione e analisi testuale ideato da Eugenio Picchi

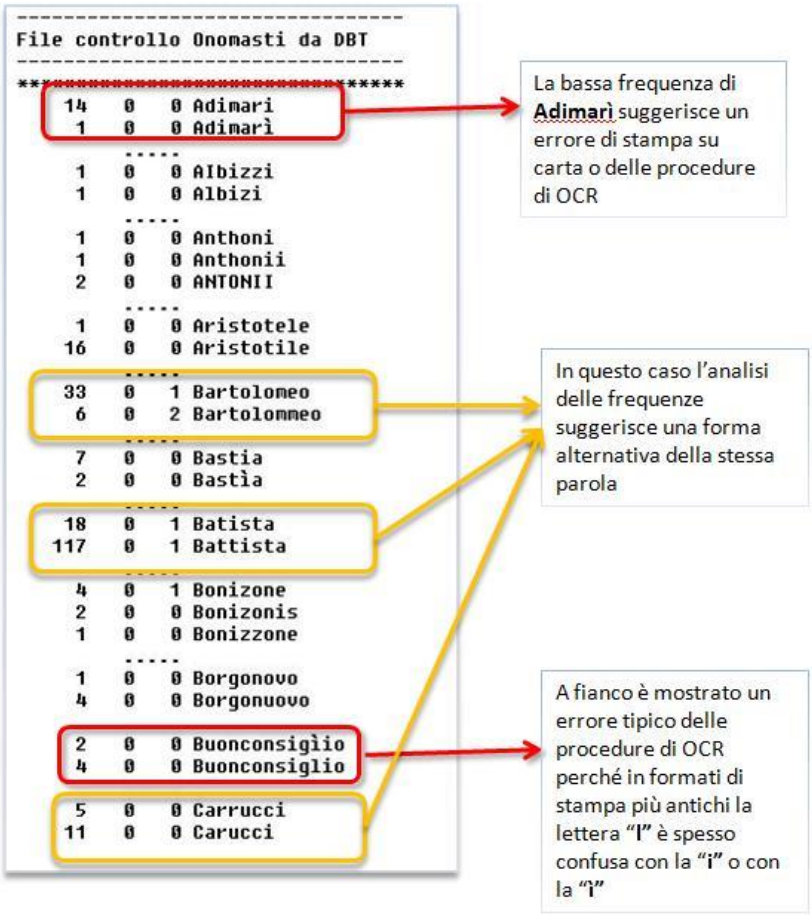

Figura 2: analisi di alcuni dati testuali

La procedura è stata necessaria per valutare la bontà degli strumenti di estrazione del testo, e laddove è stato possibile, apportare le giuste modifiche ai dati. Infatti ogni fase successiva di elaborazione del testo era condizionata dalla correttezza di quanto sottoposto ad analisi. In uno dei casi evidenziati in figura 2, dopo che le procedure DBT avevano evidenziato la presenza nei testi di entrambe le forme "Carucci" e "Carrucci", abbiamo cercato di normalizzare le due forme che evidentemente si riferivano alla stessa persona. Questo processo non è chiaramente automatizzabile per ogni parola del testo, ma la ricerca di questo famoso pittore era sicuramente rilevante e non poteva essere disattesa da

un sistema di costruzione di risorse legate ai Beni Culturali. In questo caso particolare siamo scesi nel dettaglio e abbiamo potuto appurare che, colui che universalmente è conosciuto come Pontormo, era indicato nei testi indifferentemente come:

Iacopo Carucci, Iacopo Carrucci, Jacopo Carucci, Jacopo Carrucci, Jacopo da Pontormo, Pontomo

La normalizzazione è stata quindi ottenuta riconducendo tutte le sequenze a quella più ricorrente (Pivot), in questo modo, cercando la sequenza pivot, sono individuabili nei testi anche tutte le altre forme alternative.

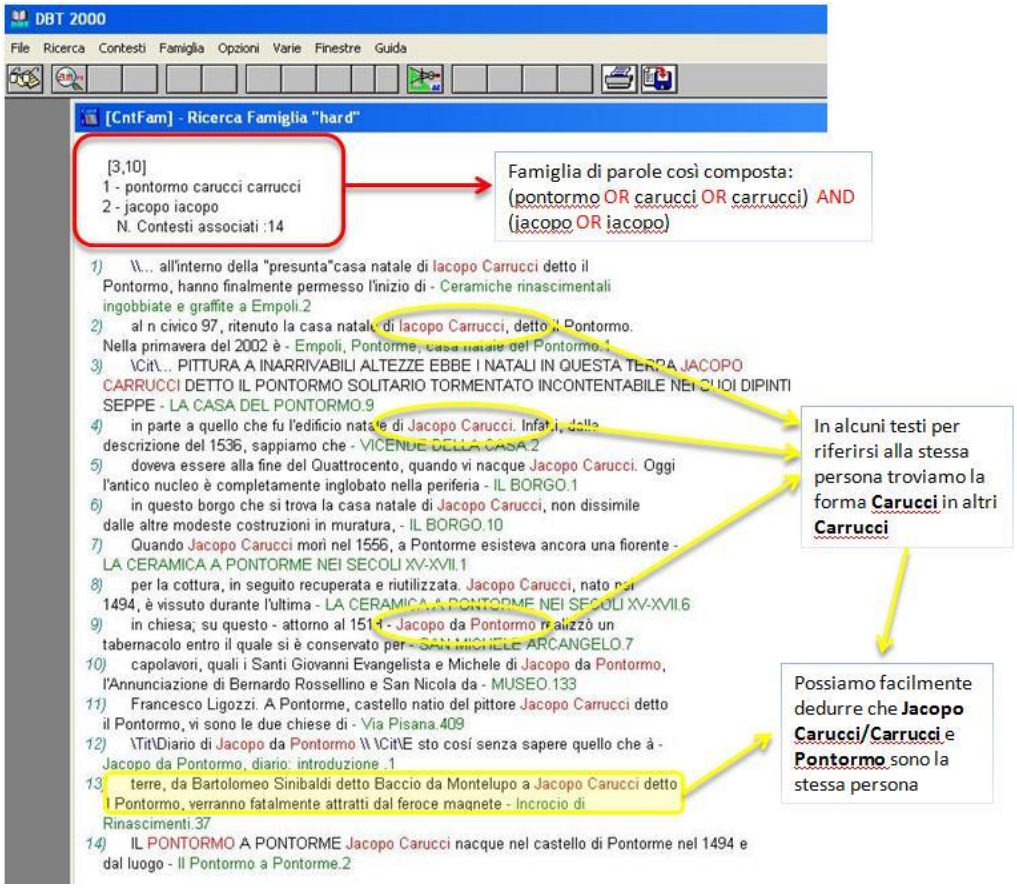

Figura 3: ricerche in DBT sul corpus

Strategie per individuare concetti rilevanti

Il riconoscimento di pattern linguistici a partire da testi lemmatizzati e disambiguati morfo-sintaticamente, costituisce un valido strumento per l'estrazione di terminologia di settore. Ponendo condizioni che valutano le categorie grammaticali vengono estratte automaticamente le sequenze di parole che rispettano tali regole. E' possibile, per esempio, analizzare le sequenze di sostantivi legati da preposizione o le coppie sostantivo aggettivo o semplicemente estrarre tutte le forme, i lemmi, le locuzioni, i verbi composti o i termini rilevanti in genere. Nel nostro caso sulla base dei testi che costituivano il corpus è stato estratto un vocabolario dedicato all'area geografica e ai temi culturali di interesse [vedi rif. 5].

Gli opportuni pattern linguistici sono stati quindi cercati all'interno del corpus, con l'intento di individuare terminologie e concetti rilevanti. In particolare per la lingua italiana i pattern più produttivi sono del tipo "sostantivo-preposizione-sostantivo" (S-E-S) o "aggettivo-sostantivo/sostantivo-aggettivo" (A-S/S-A). Successivamente i risultati di queste ricerche sono stati sottoposti ad analisi statistico-linguistiche. Una volta individuate, le connotazioni terminologiche polirematiche e monorematiche (costituite da una sola parola) sono state annotate all'interno dei testi. In una ulteriore fase di lavoro sono state poi

identificate le *named entities*[8]. A titolo di esempio ho riportato alcune delle tipologie di elementi semanticamente rilevanti estratti dal corpus. Il sistema di analisi, così come è stato concepito, propone liste ordinate di parole/gruppi di parole che sono stati considerati ottimi candidati a rappresentare terminologia di dominio:

- "campanile a vela"
- "facciata romanica"
- "scavi archeologici"
- "edificio sacro";
- "moto risorgimentale"
- "Filippo Lippi"
- "Battista di Donato Benti"
- "Signoria di Firenze"
- "pasta di vetro verde"

La terminologia individuata unisce quella tipica relativa ai beni culturali con quella specifica del territorio empolese come la "pasta di vetro verde" o il "volo del ciuco".

[8] Nomi di persona, di istituzioni, di luoghi geografici, sigle, etc.

Sempre a proposito della fase di disambiguazione morfo-sintattica sono state sperimentate strategie di riconoscimento di locuzioni e sintagmi. Per esempio il riconoscimento delle locuzioni temporali viene effettuato a partire da termini pivot e si avvale di strumenti di *pattern matching recognition* che agiscono in un intorno variabile dei termini pivot. Un altro esempio di riconoscimento di fenomeni linguistici sono i "gruppi monetari" o le "onorificenze" (con questo termine includiamo i titoli onorifici propriamente detti, i titoli nobiliari, le cariche politiche, i titoli di studio etc.). Le onorificenze assumono particolare valore quando sono unite alle terminologie poliematiche o a *named entities* ad esempio:

Professore associato di {(P)filologia classica } dell' {(C)Università di Siena}

L'esempio mostra che abbiamo a che fare con forme flessibili, infatti il titolo professore può comparire solo, preceduto o seguito da attributi. Lo stesso approccio agglomerativo è stato utilizzato per i sintagmi, cioè una combinazione di due o più elementi linguistici, che uniti hanno una specifica funzione nella struttura della frase. Diversamente dalle locuzioni molti sintagmi vengono ricondotti ad una categoria grammaticale. Ad esempio i sintagmi verbali sono ricondotti al lemma e connotati con la categoria grammaticale corrispondente al verbo. Nel caso di verbi composti verrà riconosciuto e classificato all'interno del sintagma anche questo elemento, ad esempio:

ANDARE A BUON FINE {(V_C)ANDARE# è andato} a buon fine

In merito all'esempio sopra esposto per i sintagmi verbali, dobbiamo specificare che sono state approntate anche regole per il trattamento dei verbi composti retti dai verbi ausiliari (essere e avere) che consentono di ricondurre il verbo composto al lemma del verbo reggente.

In sintesi è stato possibile annotare nel testo informazioni quali:

- ✓ Indirizzi
- ✓ Nomi propri di persona di enti o istituti
- ✓ Luoghi geografici
- ✓ Nomi di Santi
- ✓ Monumenti opere architettoniche parchi etc.
- ✓ Polirematiche
- ✓ Gruppi monetari
- ✓ Locuzioni temporali
- ✓ Sintagmi
- ✓ Verbi composti

Il corpus XML

Al suo completamento e dopo vari adeguamenti il corpus ha raggiunto le 2.219 unità testuali, di cui 1.634 provenivano da materiali cartacei e digitali, recuperati grazie al supporto delle biblioteche di Empoli e Cerreto Guidi. La dimensione complessiva del corpus ha raggiunto poco meno di 2.000.000 (1.854.654) di occorrenze, delle quali poco meno della metà (857.355), sono relative a 585 unità testuali recuperate dal web. Le unità testuali sono state annotate con le

informazioni rilevanti individuate e riorganizzate in 650 documenti in formato XML.

L'unità minima che costituiva il corpus era quindi un file di testo, acquisito a partire da un testo cartaceo, digitale o estratto da siti internet da appositi spider, e di seguito arricchito da una serie di informazioni, descrittive e strutturali, e da annotazioni che associavano a porzioni di testo particolare rilevanza semantica.

Le informazioni inserite nel file si possono in sostanza classificare in tre tipologie:

- ✓ Informazioni strutturali sul testo: queste annotazioni permettono di ricavare la struttura fisica che compone il testo.
- ✓ Informazioni descrittive sull'unità testuale: queste annotazioni permettono di associare ad una unità strutturale, o in alternativa a tutto il testo-file nel suo insieme delle informazioni amministrative o gestionali come ad esempio titolo, autore, etc.
- ✓ Informazioni semantiche e topografiche: queste annotazioni permettono di associare ad una porzione univocamente delimitata di testo un valore semantico o topografico, eventualmente riferendosi a domini esterni dove queste caratteristiche sono formalizzate e possono essere correlate.

Partendo da un formato prototipale condiviso, sono stati stabiliti i tag necessari alla codifica delle informazioni rilevanti contenute nei documenti, e alla definizione della specifica del formato XML di scambio, corredato dalla opportuna DTD. Come per le informazioni strutturali anche per quelle

semanticamente rilevanti sono stati studiati appositi tag di rappresentazione. Tipicamente sono state individuate due tipologie di fenomeni:

Terminologia:

<termine tipo='O/P/A/T'> ... contenuto del tag... **</termine>.**

Indica che il testo racchiuso fra i tag <termine> e </termine> ha un significato rilevante ai fini dell'estrazione della conoscenza contenuta nel testo. L'attributo "tipo" indica la tipologia dell'informazione contenuta nel tag, ovvero se si tratta di una *named entity* (indicata dal carattere 'O'), una poliremática (indicata dal carattere 'P'), un indirizzo (indicato dal carattere 'A') o un espressione temporale semplice (indicata dal carattere 'T'). Per esempio:

<termine tipo="O">Vincenzo Salvagnoli **</termine>**;

<termine tipo="T">Secolo XV **</termine>**;

<termine tipo="P"> Romanico fiorentino **</termine>**;

Terminologia con lemma associato:

<termineL>

<lemma> ..lemma.. **</lemma>**

<termine tipo="P/O"> ..faccetta.. **</termine>**

</termineL>

Il lemma associato rappresenta nella maggior parte dei casi una sorta di normalizzazione del termine, per esempio:

<termineL>

<lemma> marmo bianco e verde **</lemma>**

<termine tipo=”P”> marmi bianche e verdi **</termine>**

</termineL>

Questa codifica indica che nel testo è presente la stringa "marmi bianchi e verdi", ma è riconducibile al lemma "marmo verde e bianco". In questo modo le occorrenze della forma flessa plurale e singolare sono individuate come varianti dello stesso termine.

<termineL>

<lemma> bottega robbiana **</lemma>**

<termine tipo=”P”> bottega di Andrea della Robbia **</termine>**

</termineL>

In questo secondo esempio l’indicazione del lemma ha un valore esteso e indica la normalizzazione della stringa “bottega di Andrea della Robbia”, ad una forma più ampia, che comprende tutta la numerosa famiglia dei della Robbia: Luca, Giovanni, Girolamo, etc.

Analisi de risultati

Un primo elemento di confronto tra i due progetti presi in esame è rappresentato dall’arco di tempo impiegato nel loro svolgimento. Infatti, mentre il progetto Smartcity è terminato nell’arco dei due anni previsti dal progetto, con la realizzazione di un prototipo, per la biblioteca di Galileo i tempi non sono

ancora maturi per uscire con il catalogo online. Sicuramente gli sforzi messi in campo dal Museo Galileo in termini di persone, competenze e finanziamenti rendono questo progetto poco vantaggioso economicamente, ma non sarebbe corretto vederne solo il lato produttivo. Un progetto di ricerca va letto anche come sviluppo delle competenze delle persone che vi hanno lavorato, in questo senso l'esperienza fatta rappresenta certamente una crescita professionale. Le sinergie create hanno richiesto un lungo periodo di incubazione. Anche se era chiara la condivisione degli obiettivi, mancava la piena consapevolezza dello stato in cui testi digitali versavano, del loro contenuto, e del lavoro necessario al loro pieno recupero. Le prime riunioni di progetto si risolvevano in resoconti infiniti di problemi che ognuno chiamava in modo diverso.

Anche per il progetto Smartcity i problemi iniziali sono stati simili ma la tabella di marcia era ferrea e le decisioni avevano tempi certi. L'ambiente di lavoro che si realizza collaborando con aziende private è più snello e dinamico, ma spesso non c'è tempo per approfondire temi e prospettive, che restano sulla carta, pur con la migliore buona volontà di tutti.

Con una metafora possiamo dire che l'auto "azienda" è una sportiva che deve arrivare velocemente, l'auto "istituzione" è un diesel che stenta a partire ma poi difficilmente si arresta. Queste diverse filosofie ho potuto sperimentarle in casi specifici. Per esempio inizialmente con il Museo Galileo era stato scelto lo standard XML TEI P4 per la rappresentazione dei testi, ma di lì a poco si è affermato il TEI P5. Nonostante la conversione al nuovo standard non fosse

indolore, si è deciso di rimappare tutti i materiali già convertiti e reimpostare tutte le procedure di conversione per i nuovi testi ancora da trattare. Come se non bastasse alcuni tag erano stati modificati nelle loro caratteristiche e attributi ed è stato necessario ripensarne il loro uso nel nostro contesto. Per contro in Smartcity non si è adottato un formato XML standard per la rappresentazione dei testi. Le ragioni possono essere ricondotte a due motivi principali:

- ✓ non tutte le informazioni che volevamo "taggare" trovavano un corrispettivo in TEI. Lo studio di come mediare le esigenze di progetto e gli standard internazionali avrebbe richiesto un tempo troppo lungo o una competenza ulteriore, non prevista dal budget;
- ✓ il formato XML prodotto non era pensato per la conservazione e il riuso del testo. Ogni file del corpus doveva essere elaborato da una procedura software già esistente, con proprie specifiche di input. L'esperienza maturata e il grande impegno per produrre e annotare il corpus avrebbe probabilmente trovato applicazioni in altri ambiti del progetto, ma esulava dai propositi iniziali.

Esistono poi differenze nell'approcciarsi ai problemi quali errori, ripensamenti, riallocazioni. Nella mia esperienza personale nelle attività svolte per il Museo esiste quasi sempre il "passo indietro" rispetto a una decisione che si è rivelata sbagliata. Si fa nuovamente il punto e si riparte da prima dell'errore. Va sempre tenuto conto che in progetti di ricerca il confine tra una scelta giusta e una sbagliata è molto labile, spesso non esistono esperienze alle quali rifarsi per

capire dove le nostre decisioni ci porteranno. Nel lavoro fatto nell'ambito di un progetto che ha tra i suoi obiettivi il business abbiamo esigenze diverse, un cambio di direzione nel piano delle attività è mal tollerato. Tutto il progetto deve procedere alla stessa velocità, se c'è un intoppo va risolto in un tempo ragionevole. Occorre grande capacità manageriale nell'affrontare il problema ed esperienza nel saper trovare rapidamente soluzioni alternative ed efficaci. Un esempio emblematico è stato la scelta di Empoli e dintorni come corpus di riferimento. Quando ci siamo accorti che con tutta la documentazione raccolta, non riuscivamo ad ottenere dimensioni accettabili per il corpus, il cambio di obiettivo non è stato preso neanche in considerazione. La soluzione è stata cercata altrove, reperendo materiale testuale da Internet, anche se la tipologia dei materiali testuali scaricati dal web è in generale più eterogenea e frammentata, difficilmente riconducibile a categorie predefinite. A differenza dei volumi cartacei perdiamo la possibilità di mantenere una proporzione tra materiali di contenuto storico, artistico o paesaggistico. Mentre nel caso di un sito istituzionale posso ancora individuare una struttura nei contenuti, navigando nel web troviamo una serie di pagine disarticolate che provengono dalle più svariate fonti, che solo grazie ai filtri semantici individuati abbiamo potuto attribuire al dominio.

La scelta è stata consapevole, era infatti certo che, le problematiche connesse al metodo sperimentale non del tutto collaudato, all'affidabilità delle informazioni e alla mancanza di omogeneità dei contenuti, avrebbero influito sui risultati, ma la natura prototipale del progetto possiede anche questa caratteristica.

Conclusioni

Le attività in cui sono stata coinvolta sono state per me molto interessanti. Personalmente trovo che l'informatica applicata alle scienze umane sia un ambito lavorativo tra i più stimolanti. Sin da quando ho avuto il primo assegno di ricerca presso l'ILC sono stata coinvolta in progetti e attività che hanno arricchito le mie competenze di informatico, mi hanno permesso di avere una visione più ampia e articolata dei problemi che dovevo affrontare. Ho imparato che le soluzioni più efficaci, per l'economia di un progetto, sono quelle che si avvalgono delle competenze di persone che hanno profili professionali diversi.

Nella mia carriera di informatico prestato alla ricerca ho affrontato di volta in volta le richieste dello storico, del linguista, dell'umanista in generale, con malcelata diffidenza, che fortunatamente ho superato negli anni. Anzi ho sempre più apprezzato il parere e i consigli di chi vedeva le cose da punti di vista diversi dai miei. D'altro canto spesso ho avuto a che fare con filologi e studiosi che hanno avuto lo stesso atteggiamento iniziale con me, ma che poi sono diventati i migliori collaboratori. Trovare un terreno comune, dove le richieste di ognuno sono valutate e comprese non è facile: non parlare la "stessa lingua" comporta un grande forzo di adattamento.

Riferimenti

1. E. Sassolini, S. Cucurullo, M. Sassi, “Methods of textual archive preservation”. In: CLiC-it 2014 – Prima Conferenza Italiana di Linguistica Computazionale (Università di Pisa e CNR, Pisa, Italia, 9-10 dicembre 2014). Proceedings, vol. I, ISBN 978-886741-472-7, pp. 334 – 338. Pisa University Press., Pise, 2014.
2. E. Sassolini, A. Cinini, S. Sbrulli, e E. Picchi, “Tools and Resources Supporting the Cultural Tourism”. In: GL14 - Fourteenth International Conference on Grey Literature (Consiglio Nazionale delle Ricerche, Roma, Italia, 29-30 Novembre 2012). Proceedings, vol. (GL conference series, ISSN 1386-2316 ; No. 14) pp. 177 - 180. D.J. Farace, J. Frantzen, GreyNet (eds.). TextRelease, Amsterdam, 2013.
3. Spadoni F., Tariffi F., Sassolini E., “SMARTCITY: Innovative Technologies for customized and dynamic multimedia content production for Tourism applications”. In: EVA 2011 Florence Electronic Imaging and the Visual Arts. (Firenze, 4-5-6 maggio 2011). Proceedings, pp. 130 - 135. Cappellini Vito (ed.). Pitagora Editrice Bologna, 2011.
4. E. Picchi, E. Sassolini, “Text power: Tools for the cultural heritage”. In: CHC 2010 - 4-th Intl. Congr. Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Il Cairo, 6-7--8-12 2009). Proceedings, vol. 1 pp. 435 - 439. Fondazione Roma Mediterraneo, 2010.
5. Sassolini E., Cinini A., “Cultural Heritage: Knowledge Extraction from Web Documents”. In: LREC 2010 - Seventh International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010). Proceedings, pp. 3363 - 3368. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente

Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, Daniel Tapias (eds.). European Language Resources Association (ELRA), 2010.

6. E. Picchi, “Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian”. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex ‘94, Amsterdam, The Netherlands.