# Ontologie per la rappresentazione della conoscenza inclusa nei testi letterari

SEMINARIO DI CULTURA DIGITALE
CORSO DI LAUREA IN INFORMATICA UMANISTICA

PAOLA ANDRIANI

# In origine fu l'ipertesto

Il rapporto tra informatica e scienza umane è sempre stato interessante e produttivo di nuove sfide. Nel tempo sono infatti stati prodotti e affinati nuovi strumenti digitali per il trattamento di "materiale" umanistico, con lo scopo di ricavare nuova conoscenza da un ambito di per sé ostico al trattamento "informatico".

In origine fu l'ipertesto, con il fondamentale testo *Hypertext*[1] di George Landow, in cui si auspicava tramite l'utilizzo della nuova tecnologia, una felice convergenza tra la tradizione degli studi letterari e il dominio delle discipline computazionali.
Da quel primo intervento sono passanti più di vent'anni, l'ipertesto è ormai considerato come superato e nuove tecnologie si sono succedute. Quella che all'inizio era una disciplina di nicchia, definita *Humanities Computing,* si è rafforzata e consolidata nel tempo, prendendo il nome di *Digital Humanities*.
Gli ambiti di competenza delle *Digital Humanities* (o Informatica Umanistica) sono numerosi e diversificati.

Uno dei principali è quello relativo al trattamento dei testi. Inizialmente forte è stata la spinta verso una massiva campagna di digitalizzazione di fonti primarie e secondarie, attraverso l'utilizzo di diversi linguaggi standard (in particolare XML-TEI) e lo sviluppo di infrastrutture software utili sia per il reperimento e la visualizzazione off e on line dei dati così codificati, sia per realizzare nuove analisi testuali dei testi trattati.

Queste sfide hanno prodotto negli anni numerosi e interessanti risultati e in particolare:
1. una nuova consapevolezza teorica e metodologica,
2. il concetto di modellizzazione di "dati letterari", per un migliore utilizzo di questi dati da parte degli strumenti informatici,
3. la predisposizione e l'utilizzo di linguaggi standard e condivisi per la modellizzazione, rappresentazione, condivisione e disseminazioni di risorse digitali di qualità.

Quello che è stato maggiormente lamentato da parte degli esperti di dominio informatici, è la mancanza di una formalizzazione delle teorie del testo, laddove la rappresentazione informatica necessita di una formalizzazione e una strutturazione più puntuale di teorie che per definizione non lo sono.

I tentativi di colmare questo divario tra l'informatica stretta e il mondo letterario sono stati molteplici, a partire da quella branchia definita come "codifica dei testi". Tuttavia gli esperti di dominio ritengono non ancora sufficiente il grado di formalizzazione raggiunto e spingono verso nuovi sistemi di codifica della conoscenza.

Le nuove frontiere di formalizzazione della teoria letteraria hanno aperto così nuove strade, focalizzando l'attenzione sopratutto su due binari:
- uno include tutta la conoscenza che ruota intorno ai cosiddetti *Big Data*, quindi lo sviluppo di strumenti per l'analisi automatica di ingenti quantità di dati testuali/documentali disponibili, attraverso l'uso di metodologie e tecnologie di *text mining* e *knowledge extraction*,
- l'altro è il campo trattato in questa relazione, definito *Semantic Web,* che include la sperimentazione di nuovi linguaggi e modelli di dati per la rappresentazione dei livelli semantici nelle risorse informative attraverso l'utilizzo di tecnologie e architetture legate ai paradigmi del *Semantic Web* e dei *Linked Data*, di cui parleremo più avanti.

[1] G. P. Landow, *Hypertext: the convergence of contemporary critical theory and technology*, Baltimore, Johns Hopkins University Press, 1992.

# Che cos'è il *Semantic Web*?

> *"The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries" (Tim- Berners Lee, 2001)* [2]

La definizione di cosa è il *Semantic Web* è da attribuire a Tim Berners Lee . L'idea di fondo è quella di associare alle risorse informative sul Web una descrizione formale del loro significato, attraverso la sovrapposizione di uno o più livelli di metadati semantici.

Tali metadati sono espressi in formalismi che fanno parte della famiglia dei sistemi di rappresentazione della conoscenza sviluppati nell'ambito dell'intelligenza artificiale e permettono un'elaborazione su diversi livelli. È possibile infatti effettuare dalle semplici visualizzazioni e consultazioni dei dati attraverso indici strutturati, o delle interrogazioni più complesse fino ad arrivare alla derivazione di nuova conoscenza mediante inferenze logiche.

Una volta definite e descritte queste risorse, lo scopo è quello di condividere e riutilizzare le stesse risorse in ambiti e progetti differenti senza doverle descrivere nuovamente ogni volta.

L'architettura generale del Web Semantico è raffigurata solitamente come un diagramma a pila, che ne mostra i vari componenti a vari livelli di astrazione. La base di partenza sono gli oggetti informativi a cui si applica, definiti appunto come risorse.

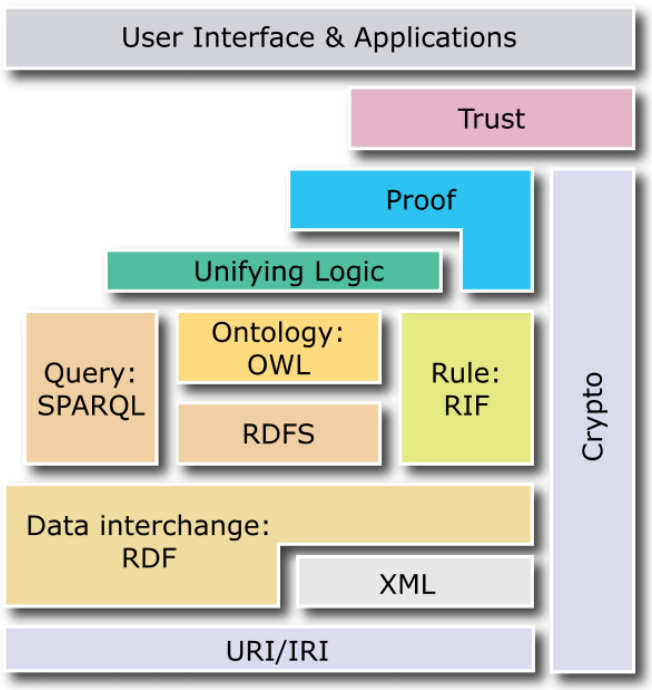



Fig. 1 - *Semantic Web*: diagramma a pila, da http://www.w3.org/2007/03/layerCake.png (agosto 2015)

Il primo problema che si pone è l'identificazione univoca delle risorse in modo non ambiguo e indipendente.

[2] "W3C Semantic Web Activity". World Wide Web Consortium (W3C), <http://www.w3.org/2001/sw/>, agosto 2015

L'identificazione dell'oggetto che può essere un oggetto informativo, un oggetto reale o un oggetto astratto, avviene quindi tramite URI[3] (*Uniform Resource Identifiers*), un identificativo univoco e persistente che permette di individuare e collegarsi ad una data risorsa nel Web.
Le forme più comuni di URI sono gli indirizzi delle pagine web (URL - Uniform Resource Locator), un sottoinsieme degli URI che svolgono anche la funzione di localizzatori di una risorsa.
Una volta identificata la risorsa in maniera univoca, è possibile allegare alla risorsa i metadati semantici che la riguardano, ossia proprietà, relazioni e una descrizione dell'oggetto in questione.

DBPedia[4], la versione formalizzata della nota enciclopedia collaborativa Wikipedia, è un ottimo repository di URI. Per esempio l'URI del frutto mela è così formato: http://dbpedia.org/resource/Apple[5] e mostra oltre ad una breve descrizione, una serie di proprietà collegate il cui valore può essere un ulteriore URI o un valore letterale, ossia i metadati semantici di cui abbiamo scritto poco sopra.

About: Apple
An Entity of Type : eukaryote, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

La manzana es una fruta pomácea comestible, fruto del manzano doméstico (Malus domestica), otros manzanos (especies del género Malus) o híbridos de aquel. En algunas zonas de España se le suele llamar pero cuando tiene forma alargada, aunque en muchas partes de Andalucía se llama indistintamente "pero" a la manzana.

| Property | Value |
|---|---|
| dbo:abstract | La mela è il frutto (più precisamente si tratta di un falso frutto a pomo) del melo.Il melo ha origine in Asia centrale e l'evoluzione dei meli botanici risalirebbe al Neolitico.La specie è pres… è il frutto più destagionalizzato (lo si trova tutto l'anno) e ciò richiede la presenza di impianti che provvedono alla conservazione e ne distribuiscano la disponibilità su di un ampio arco di… elevati contenuti in acidi organici, di norma la conservazione va da uno a quattro mesi.Nella conservazione industriale sono importanti le condizioni fisiche in cui questa avviene.Dopo il … controllata (più ricca di CO2). La mela ha un potere antiossidante (ORAC) con un indice di valore 4275 poiché contiene vitamine importanti come provitamina A, vitamine B1, B2, B6, E … anche in cucina per la preparazione di primi, secondi e diversi dolci. Inoltre si presta anche ad essere utilizzata per preparare in casa maschere di bellezza. La mela è da sempre alleata produzione di succhi, sidro, olio di semi di mela (molto utilizzato nei paesi del nord Europa ed ottenuto come sottoprodotto dalla produzione del succo e del sidro), creme, fette di mela e… difficilmente contrastabili, al colpo di fuoco batterico, alla ticchiolatura, oidio e afidi. Si punta anche all'ottenimento, per le varietà commerciali più note, di cloni autofertili. |
| dbo:binomialAuthority | dbr:Moritz_Balthasar_Borkhausen |
| dbo:class | dbr:Eudicots |
| dbo:division | dbr:Flowering_plant |
| dbo:family | dbr:Rosaceae |
| dbo:genus | dbr:Malus |
| dbo:kingdom | dbr:Plant |
| dbo:order | dbr:Rosids<br>dbr:Rosales |
| dbo:synonym | Malus communis (Desf.)<br>Malus pumila (auct.)<br>Pyrus malus (L.) |

Fig.2 - Link alla risorsa "Mela", da dbpedia.org (agosto 2015)

Tali metadati, detti asserti, vengono definiti in un linguaggio trattabile computazionalmente, l'RDF[6] (*Resource Description Framework*), definito dal W3C come lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati.

L'RDF è un metalinguaggio dichiarativo per formalizzare asserti che esprimono proprietà e relazioni tra risorse il cui modello dei dati è basato su tre elementi:
- risorse,
- proprietà,
- relazioni.

Le risorse, come abbiamo visto, possono essere pagine Web, documenti, persone, oggetti reali o astratti, concetti, identificate appunto da un URI.
Le relazioni sono specificate da un asserto espresso tramite una struttura formata da soggetto, predicato e oggetto, definito "tripla". Le proprietà sono delle relazioni che legano tra loro risorse e valori (soggetto e oggetto), e sono anch'esse identificate da URI, hanno un significato specifico, una serie di valori leciti e sono associabili ad uno o più tipi di risorse.

[3] URI, http://www.w3.org/TR/uri-clarification/, agosto 2015

[4] DBPedia, http://wiki.dbpedia.org, agosto 2015

[5] Definizione di una mela in DBPedia, http://dbpedia.org/page/Apple, agosto 2015

[6] RDF, http://www.w3.org/RDF/, agosto 2015

Un valore, invece, è un tipo di dato primitivo, che può essere una stringa o l'URI di una risorsa.

Il modello di dati RDF è quindi formato da risorse, proprietà e valori ed è rappresentabile da un grafo orientato sui cui nodi ci sono risorse o tipi primitivi e i cui archi rappresentano le proprietà

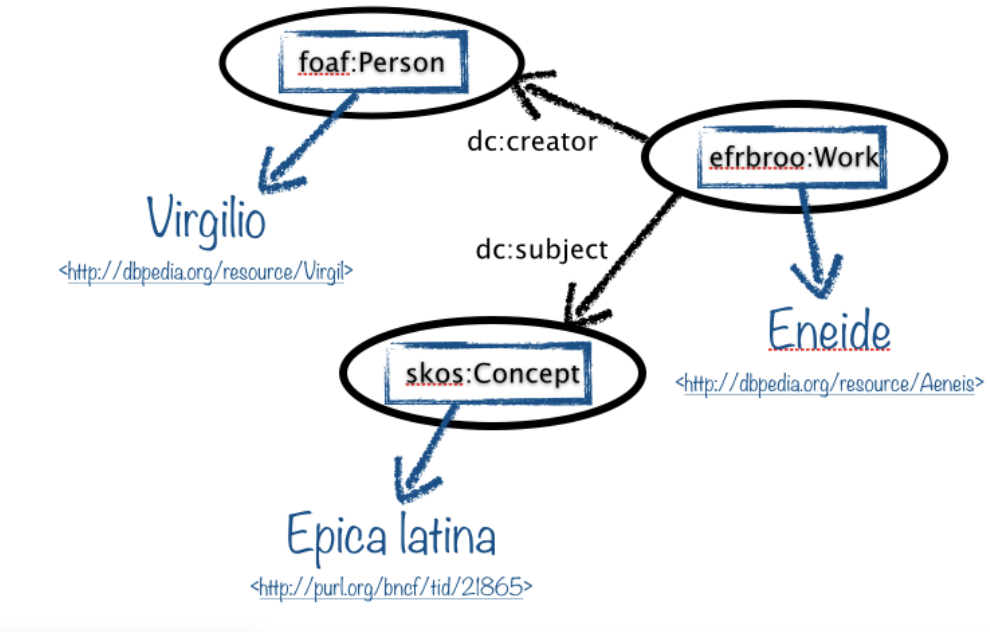



Fig. 3 - Grafo RDF relativo alla risorsa "Eneide"

```
<http://dbpedia.org/resource/Aeneid>
        a                       efrbroo:Work , rdfs:Resource ;
        dnt:workCreator         <http://dbpedia.org/resource/Virgil> ;
        dc:creator              <http://dbpedia.org/resource/Virgil> ;
        dc:subject              <http://purl.org/bncf/tid/21865> ;
        dcterms:alternative     "Aeneis" .
```

Fig. 4 - Triple RDF relative alla risorsa "Eneide"

Nelle immagini 3 e 4 possiamo vedere l'esempio di alcune triple RDF e del grafo ad esse associato.
Possiamo infatti notare che il testo Eneide(il soggetto, rappresentato dall'URI http://dbpedia.org/resource/Aeneis) ha un autore (predicato, rappresentato dalla proprietà dc:creator) il cui valore è Virgilio (oggetto, rappresentato dall'URI http://dbpedia.org/resource/Virgil).
Le triple così costruite seguono il paradigma dei *Linked Data*.

Una volta create le triple RDF, esse possono essere immagazzinate in un *triple store*, per esempio *Virtuoso*[7], e interrogate attraverso SPARQL[8], un linguaggio di interrogazione per grafi RDF.

[7] Virtuoso triple store, https://www.w3.org/2001/sw/wiki/OpenLink_Virtuoso e http://virtuoso.openlinksw.com, agosto 2015

[8] SPARQL, http://www.w3.org/TR/rdf-sparql-query/, agosto 2015

## Cosa sono i Linked Data?

Con il termine Linked Data si indicano un insieme di soluzioni per la pubblicazione e l'interconnessione di dati strutturati sul Web mediante le tecnologie del Web Semantico.
Sono delle indicazioni di buona pratica per pubblicare e collegare dati strutturati in rete, nello specifico:
- Usare URI per identificare gli oggetti.
- Usare HTTP URI per fare in modo che sia possibile reperire questi oggetti in rete.
- Utilizzare linguaggi standard come RDF e SPARQL per fornire e reperire informazioni utili collegate agli URI prima definiti.
- Includere link ad altri URI in modo da scoprire più cose quando si fa una ricerca su una risorsa.

Come abbiamo visto negli esempi precedenti, sia il repository DBPedia, sia il linguaggio RDF si avvale delle buone pratiche prescritte dal W3C e dai principi dei *Linked Data.* Questo facilita il riuso delle risorse in rete, che è uno degli scopi primari del Web Semantico.

## Che cos'è un ontologia?

Il termine, ereditato dalla metafisica classica, denota varie classi di oggetti, dai vocabolari controllati, ai thesauri, fino alle ontologie formali vere e proprie.
Come abbiamo già detto, gli informatici richiedono un alto grado di formalizzazione dei dati umanistici, per poterli elaborare computazionalmente. Abbiamo visto finora come formalizzare le risorse oggetto dei nostri studi e come formalizzare le descrizioni, le proprietà e i valori che le riguardano.

Il successivo livello di formalizzazione è quello che prevede l'inserimento delle triple base RDF in un modello più ampio: un'ontologia formale.

La definizione di questo concetto è di T.R. Gruber che definisce un'ontologia come *"an explicit specification on a conceptualization"*[9].
È di fatto una rappresentazione formale attraverso un modello che nomina e definisce gli elementi che ne fanno parti ossia:
- classi (o tipi) di risorse
- classi di proprietà
- relazioni tra classi di risorse e proprietà (es: classe –> sottoclasse)
- domini e range di proprietà

L'RDF è un linguaggio che può essere utilizzato per definire delle ontologie formali. Non permette però di specificare le relazioni logico-semantiche tra oggetti e proprietà di un medesimo schema o tra schemi diversi. Ad un livello più complesso si pone quindi OWL (Web Ontology Language)[10] che permette un secondo livello di formalizzazione, più completo, introducendo la possibilità di effettuare sui dati deduzioni e inferenze logiche.

Riassumendo:
- i *Linked Open Data* forniscono un metodo di pubblicazione di dati strutturati in modo che possano essere collegati facilmente tra di loro,

---

[9] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993, on line all'indirizzo http://tomgruber.org/writing/ontolingua-kaj-1993.htm, agosto 2015

[10] OWL, http://www.w3.org/2001/sw/wiki/OWL, agosto 2015

- le tecnologie del web semantico, come RDF, OWL e SPARQL unite alle tecnologie standard del web (HTTP, URI) ampliano la possibilità di formalizzazione delle risorse e permettono di creare e interrogare delle basi di conoscenza strutturate, dette ontologie formali,
- le ontologie formali rappresentano il modello da popolare con le risorse così codificate e permettono di collegare tra di loro dataset diversi, aumentando la possibilità di condivisione e riutilizzo delle risorse esistenti in rete.

I paradigmi del *semantic web* offrono le basi per la struttura formale, attraverso l'uso di metadati semantici.
Il paradigma dei linked data sono delle *best practice* per pubblicare e collegare dati in rete, attraverso l'uso delle tecnologie web standard come HTTP, RDF e URI.
Le ontologie formalizzano e organizzano la conoscenza su un dato ambito e vengono popolate con dati che seguono il paradigma dei *Linked Data*.

## Ontologie esistenti per il trattamento della conoscenza inclusa nei testi letterari

La rappresentazione della conoscenza inclusa nei testi letterari è complessa.
In letteratura esistono numerose ontologie, sviluppate nel corso degli anni da enti, istituzioni e progetti di ricerca che si focalizzano su aspetti differenti dell'informazione testuale, ma non esiste una sola ontologia che li rappresenti tutti.

È possibile quindi, in base alle informazioni che vogliamo formalizzare, selezionare una o più ontologie già esistenti, o una parte di esse, per creare il modello della base di conoscenza che vogliamo popolare.
In questo modo viene incentivata la logica del riuso per le parti già modellate da altri, estendendo in un secondo momento il modello nel caso in cui siano necessarie categorie e proprietà non presenti nelle ontologia di riferimento.

Tra le ontologie più note ci sono:
CIDOC-CRM, FRBR, FRBRoo, SAWS, DM2E Model, Dublin Core, SKOS, FOAF, DoCO, FaBIO, CiTO, Annota0on Ontology, Open Annotation Core Data Model.
Queste ontologie sono state sviluppate nel campo della letteratura e dei beni culturali e codificano nozioni legate al dominio testuale.

### CIDOC-CRM

L'ontologia CIDOC *Conceptual Reference Model* (CRM)[11] fornisce le definizioni e la struttura formale per rappresentare la conoscenza implicita ed esplicita in ambito museale e dei beni culturali.
Per la sua ampia copertura e autorità (è uno standard ISO) è considerata come vocabolario comune per rappresentare informazioni pubblicate in archivi, musei, gallerie, librerie e altre istituzioni culturali e per mapparle in una equivalente rappresentazione digitale. Gli istituti culturali sono incoraggiati nell'uso di questa ontologia per migliorare l'accessibilità alle informazioni e alla conoscenza relativa ai materiali culturali che gestiscono.

### FRBR e FRBRoo

FRBR (*Functional Requirements for Bibliographic Records*)[12] è un modello concettuale entità-relazione, realizzato dalla International Federation of Library Associations and Institutions (IFLA) con lo scopo di dare una rappresentazione semi-formale delle informazioni bibliografiche.
FRBR include una descrizione del modello concettuale (entità, proprietà e attributi o metadati), un esempio

[11] CIDOC-CRM, http://www.cidoc-crm.org e http://www.cidoc-crm.org/cidoc_core_graphical_representation/hierarchy.html, agosto 2015

[12] FRBR, http://www.ifla.org/publications/functional-requirements-for-bibliographic-records, agosto 2015

di record bibliografico per tutti i tipi di materiali, e task utente associati alle risorse bibliografiche descritte in cataloghi, bibliografie e altri strumenti bibliografici. FRBR contiene molte classi utili per descrivere domini di tipo testuale. Per esempio la classe Work, utile per rappresentare una creazione intellettuale, o la classe Expression, le cui istanze sono un'espressione di un singolo lavoro, solitamente in formato fisico.

**FRBRoo** (**"FRBR-object oriented"**)[13] è la versione di FRBR espressa in un modulo compatibile con l'ontologia CIDOC-CRM.

**SAWS**

L'ontologia SAWS (*Sharing Ancient Wisdom*)[14] è stata sviluppata all'interno del progetto SAWS, il cui scopo è quello di presentare e analizzare il sapere della tradizione letteraria greca e araba. In particolare l'obiettivo è quello di pubblicare diverse collezioni di testi on line, usando la codifica XML-TEI per codificare i testi e RDF per esprimere e visualizzare le relazioni fra i materiali codificati. Tecnicamente SAWS riusa ed estende l'ontologia FRBRoo.

**EDM**

*Europeana Data Model*[15] è un'ontologia creata all'interno del progetto Europeana, che raccoglie dati relativi al patrimonio culturale europeo, a partire dalla modellizzazione dei dati originaria, trasformandola secondo i principi del web semantico. In questo modo è diventato un modello di dati più sviluppato che aggiunge collegamenti più ricchi e significativi al patrimonio culturale europeo. I dati provenienti da partner o risorse informative esterne con riferimenti a persone, luoghi, oggetti , ecc., sono così collegati e condivisi fra diverse istituzioni e iniziative differenti.

**DM2E Model**

DM2E[16] è un'ontologia sviluppata nell'ambito del progetto "*Digitised Manuscripts to Europeana Project*" ed è l'attuale specializzazione del modello di dati Europeana ( EDM ) per la gestione dei manoscritti . Il modello è strato utilizzato per trasferire e trasformare tutti i record di metadati e oggetti digitali forniti al progetto DM2E dal loro formato locale in risorse RDF . Il Modello DM2E riutilizza il più possibile, proprietà e classi provenienti da altre ontologie.

**Dublin Core Metadata Element Set**

Il data set *Dublic Core Metadata Element Set*[17] è un vocabolario di quindici proprietà utilizzabile per la descrizione di qualsiasi tipo di risorsa accessibile via web. Questi quindici elementi fanno parte di un più grande dataset di metadati sviluppato dalla *Dublin Core Metadata Initiative*. Le proprietà RDF, le classi, lo schema di codifica e i tipi di dati dichiarati e mantenuti dalla *Dublin Core Metadata Initiative* sono conosciuti come *DCMI Metadata Terms* e possono essere utilizzati per descrivere un ampio spettro di risorse web (video, immagini, pagine web, ecc.) e oggetti fisici, come libri o manufatti.

**SKOS**

*Simple Knowledge Organization System* (SKOS)[18] è un modello per condividere e collegare sistemi di

---

[13] FRBRoo, http://www.cidoc-crm.org/frbr_inro.html, agosto 2015

[14]SAWS, http://www.ancientwisdoms.ac.uk/method/ontology/, agosto 2015

[15] Europeana, http://www.europeana.eu/portal/ e http://pro.europeana.eu/page/edm-documentation, agosto 2015

[16] DM2E Model, http://wiki.dm2e.eu/Main_Page e http://dm2e.eu, agosto 2015

[17] Dublin Core, http://dublincore.org/documents/dces/, agosto 2015

[18] SKOS, http://www.w3.org/TR/skos-reference/, agosto 2015

conoscenza organizzata come thesauri, schemi di classificazione, tassonomie. È stato per esempio utilizzato dalla Biblioteca Nazionale di Firenze per realizzare il soggettario nazionale[19], strumento rivolto alle biblioteche, i musei, le mediatiche, gli archivi e i centri di documentazione, creato con lo scopo di indicizzare per soggetto risorse di varia natura. È aderente ai principi stabiliti dall'*International Federation of Library Associations and Institutions* (IFLA) e alle indicazioni degli standard internazionali.

**FOAF**

*Friend of a friend* (FOAF)[20] è un ontologia creata con lo scopo di collegare persone e informazioni attraverso il web. In particolare FOAF permette di descrivere persone, attività e relazioni con altre persone e oggetti.

**SPAR**

Le ontologie SPAR (*Semantic Publishing and Referencing*)[21] sono un insieme di otto ontologie complementari e ortogonali usate per descrivere tutti gli aspetti delle pubblicazioni bibliografiche sotto forma di metadati comprensibili per un computer.
FaBiO, CiTO, BiRO, C4O, DoCO, PRO, PSO e PWO sono utili per descrivere oggetti bibliografici, record bibliografici e riferimenti, citazioni, conteggi delle citazioni , contesti delle citazioni e le loro relazioni con sezioni rilevanti relative ai documenti citati, liste ordinate di riferimenti e cataloghi di biblioteche, parti di un documento, ruoli editoriali, stato della pubblicazione e flussi di lavoro editoriali.

Per esempio, per descrivere la struttura di un testo, i tipi di pubblicazioni testuali e le citazioni bibliografiche sono interessanti le tre ontologie DoCO, FaBiO e CiTO, sviluppate all'interno del progetto SPAR.

- **DoCO,** *Document Components Ontology*, permette di descrivere le parti di un documento in RDF. Questi componenti sono sia strutturali (blocchi, capitoli, intestazioni, paragrafi, sezioni) sia retorici (abstract, introduzione, risultati, discussioni, conclusioni, ringraziamenti, bibliografia).

- **FaBiO**, è un'ontologia allineata con FRBR ed è stata creata con lo scopo di registrare e pubblicare nell'ambito del *Semantic Web* record bibliografici creati dagli studiosi. Le entità di FaBiO sono principalmente pubblicazioni testuali come libri, riviste, quotidiani, giornali e elementi in essi contenuti come poemi, articoli di conferenze e editoriali.

- **CiTO**, *Citation Typing Ontology* è un ontologia per le citazioni bibliografiche, sia fattuali che retoriche creata per la loro pubblicazione in rete.

**Annotation Ontology**

L'*Annotation Ontology*[22] è un vocabolario per la codifica di diversi tipi di annotazioni, commenti, tag semantici, annotazioni testuali, note, ecc, su ogni tipo di documento elettronico (testi, immagini, audio, video, ecc) e parti di documento. AO non fornisce alcuna ontologia di dominio, ma favorisce il riutilizzo di quelli esistenti al fine di non interrompere il principio della scalabilità del Semantic Web.

---

[19] Soggettario Nazionale, http://thes.bncf.firenze.sbn.it, agosto 2015

[20] FOAF, http://www.foaf-project.org, agosto 2015

[21] SPAR, http://sempublishing.sourceforge.net, agosto 2015

[22] Annotation Ontology, https://code.google.com/p/annotation-ontology/wiki/Homepage, agosto 2015

**Open Annotation Core Data Model**

L'*Open Annotation Core Data Model*[23] ha un approccio volto a collegare le annotazioni con le risorse, utilizzando una metodologia conforme all'architettura del *World Wide Web* e l'iniziativa *Linked Data*. Nell'*Open Annotation Model* un'annotazione è considerata come un insieme di risorse connesse, che include un corpo e un target e la relazione tra di essi. Tale ontologia è particolarmente interessante perché rappresenta in dettaglio la struttura delle note ai testi.

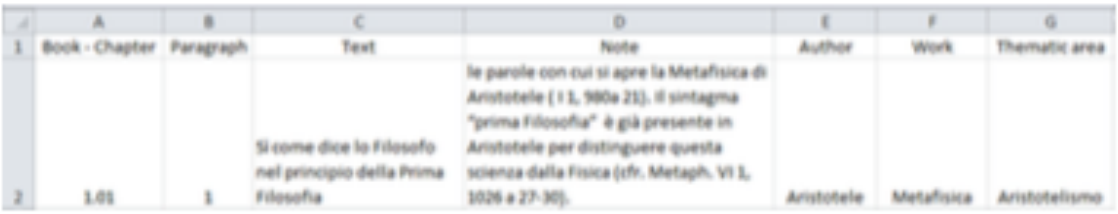
**Ontologia per la codifica delle fonti primarie**

L'ontologia creata nell'ambito del progetto DanteSources "Per una enciclopedia dantesca digitale" riutilizza parte delle ontologie precedentemente descritte, integrandole con altre proprietà create ex novo.
È stata realizzata con lo scopo di codificare la conoscenza contenuta nei commenti relativi ai testi danteschi e costruire una base di conoscenza da poter interrogare su quelle che sono presumibilmente le fonti primarie citate da Dante durante la stesura delle sue opere.Il modello si intende aperto e riutilizzabile, disponibile sia per essere esteso che per codificare la conoscenza sulle fonti primarie relative ad altri autori.

## Il progetto DanteSources, un caso di studio

Come abbiamo visto in letteratura esistono molte ontologie che si concentrano su differenti aspetti dell'informazione testuale. Alcune di esse rappresentano un insieme di possibili interpretazioni del testo di partenza. Tra queste, mancava un'ontologia specifica che analizza il rapporto tra un testo e le sue fonti primarie. Il progetto "Per una enciclopedia dantesca digitale" ha creato una base di conoscenza che memorizza informazioni relative alle fonti primarie a cui Dante si riferisce nelle sue opere. Il modello creato è stato popolato con le informazioni relative a Dante e alle fonti primarie da lui citate, ma l'applicabilità di questo lavoro va oltre lo specifico autore e può essere utilizzato anche per autori differenti.

Il punto di partenza è stato un foglio di lavoro Excel, il cui contenuto è stato elaborato da uno studioso di Dante. A partire dai dati organizzati nel foglio di calcolo, è stato sviluppato un modello concettuale, ossia un insieme di classi e proprietà che mappano il contenuto del file.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Book - Chapter | Paragraph | Text | Note | Author | Work | Thematic area |
| 2 | 1.01 | 1 | Si come dice lo Filosofo nel principio della Prima Filosofia | le parole con cui si apre la Metafisica di Aristotele ( I 1, 980a 21). Il sintagma "prima Filosofia" è già presente in Aristotele per distinguere questa scienza dalla Fisica (cfr. Metaph. VI 1, 1026 a 27-30). | Aristotele | Metafisica | Aristotelismo |

Fig. 5 - Foglio di calcolo relativo al testo "Convivio" di Dante Alighieri

Sono state riviste le ontologie esistenti con lo scopo di creare un vocabolario per esprimere le classi e le proprietà identificate. Molti termini sono stati ripresi dalle ontologie analizzate in modo da massimizzare l'interoperabilità della rappresentazione creata. Sono infine state aggiunte le classi e le proprietà mancanti. L'ontologia così realizzata è stata espressa in RDF.

[23] Open Annotation Core Data Model, http://www.openannotation.org/spec/core/, agosto 2015

```
<!-- Classes -->

<!--ExpressionFragment-->
<rdfs:Class rdf:about="efrbroo:ExpressionFragment">
<rdf:type rdf:resource="dctypes:Text"/>
<rdf:type rdf:resource="dctypes:ContentAsText"/>
</rdfs:Class>

<!-- Work. Inferenza: ordinamento con cui le fonti primarie vengono citate da Dante-->
<rdfs:Class rdf:about="efrbroo:Work">
</rdfs:Class>

<!-- Expression. Inferenza: ordinamento con cui le fonti primarie vengono citate da Dante-->
<rdfs:Class rdf:about="efrbroo:Expression">
</rdfs:Class>

<!-- Annotation -->
<rdfs:Class rdf:about="oa:Annotation">
</rdfs:Class>

<!-- Body -->
<rdfs:Class rdf:about="oa:Body">
<rdf:type rdf:resource="dctypes:Text"/>
<rdf:type rdf:resource="dctypes:ContentAsText"/>
</rdfs:Class>

<!-- Paragraph. Inferenza: ordinamente di paragrafi, capitoli e libri -->
<rdfs:Class rdf:about="doco:Paragraph">
</rdfs:Class>

<!-- Chapter -->
<rdfs:Class rdf:about="doco:Chapter">
</rdfs:Class>

<!-- Book -->
<rdfs:Class rdf:about="fabio:Book">
</rdfs:Class>
```

```
<!-- Properties -->

<!--hasNote -->
<rdf:Property rdf:about="crm:hasNote">
<rdfs:domain rdf:resource="efrbroo:ExpressionFragment"/>
<rdfs:range rdf:resource="oa:Body"/>
</rdf:Property>

<!--hasTarget -->
<rdf:Property rdf:about="oa:hasTarget">
<rdfs:domain rdf:resource="oa:Annotation"/>
<rdfs:range rdf:resource="efrbroo:ExpressionFragment"/>
</rdf:Property>

<!--hasBody -->
<rdf:Property rdf:about="oa:hasBody">
<rdfs:domain rdf:resource="oa:Annotation"/>
<rdfs:range rdf:resource="oa:Body"/>
</rdf:Property>

<!--hasDate
<rdf:Property rdf:about="dc:hasDate">
<rdfs:domain rdf:resource="oa:Body"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>
-->
```

Fig. 6 - Classi e proprietà estratte dal foglio di calcolo e definite in uno schema RDF

Nel foglio Excel fornito dallo studioso sono stati identificate le seguenti colonne:

- il numero del libro, del capitolo e del paragrafo a cui la nota del commento al testo si riferisce,
- il frammento del testo dantesco a cui la nota si riferisce,
- un frammento della nota,
- la fonte primaria citata, strutturata in autore, titolo e area tematica,

a cui sono state aggiunte in corso d'opera l'intero testo del commento al frammento di testo dantesco e il grado di citazione (esplicita, stringente o generica).

In sintesi, quasi il 90% delle classi e proprietà dell'ontologia creata deriva da ontologie standard. Sono state riutilizzate classi e proprietà rispettando i vincoli stabiliti nelle ontologie di origine, vale a dire che sono state conservate le definizioni delle classi e i domini e range delle proprietà scelte, questo per evitare problemi di natura semantica.
Nello specifico, ecco una riassunto del mapping che è stato effettuato. Per convenzione, è stato apposto un prefisso alle proprietà e le classi riprese da altre ontologie (per esempio *frbroo:hasFragment* è ripreso dall'ontologia FRBRoo) mentre sono state lasciate senza prefisso le classi e le proprietà create ex novo.

Per la descrizione della risorsa a cui si riferisce il testo dantesco e la nota dello studioso, ossia il paragrafo (o verso nel caso di un testo poetico), il capitolo (o il poema), il capitolo e il libro a cui la nota del commentario riferisce, sono state selezionate alcune classi dalle ontologie FaBiO e DoCO.
Nello specifico:

- *doco:Paragraph*, che rappresenta un'unità di discorso auto-contenuta incentrata su un particolare punto o un'idea;
- *fabio:Poem*, che rappresenta un'opera artistica scritta con un intensità di linguaggio carattestico della poesia piuttosto che della prosa;
- *doco:Line*, che rappresenta il verso in poesia, per esempio l'unita di linguaggio in cui un poema è suddiviso;
- *doco:Chapter*, che rappresenta una delle principali suddivisioni del corpo di un testo di grande dimensione;
- *fabio:Book*, che definisce un documento completo in un volume o in un numero finito di volumi.

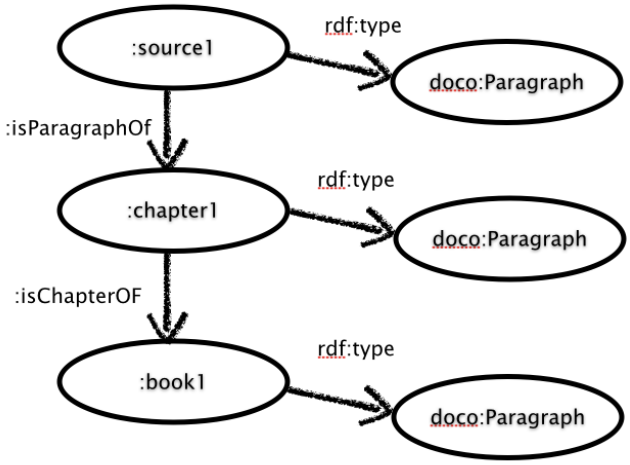



Fig. 7 - Descrizione della risorsa a cui si riferisce il testo dantesco sotto forma di grafo RDF

Per collegare i rispettivi paragrafi (o versi) con i corrispondenti capitoli (o poemi) e libri, sono state definite tre sottoproprietà della proprietà *:isPartOf* dell'ontologia FRBR: *:isParagraphOf*, *:isChapterOf* e *:isBookOf*.

Dalle ontologie CIDOC-CRM e FRBR, sono state riprese le classi:

- *efrbroo:Work*, per rappresentare una specifica opera, citata dai commentatori (per esempio la Metafisica di Aristotele), senza che sia specificata una particolare edizione;
- *efrbroo:Expression*, per mappare una precisa edizione di un testo;
- *efrbroo:ExpressionFragment*, per definire sia il frammento del testo dantesco a cui la nota si riferisce, sia il frammento della nota che punta ad una specifica fonte primaria.

In aggiunta, seguendo le raccomandazioni del W3C rispetto alle specificazione su come rappresentare il contenuto degli elementi codificati in RDF[24], sono state aggiunte le sotto-proprietà *dc:format* (dal Dublin Core) e *cnt:chars*[25] per definire il formato e il tipo di dato (letterale nello specifico) con cui sono espresse le istanze della classe ExpressionFragment.

Come riportato in precedenza, per descrivere un'opera a cui il commento dello studioso si riferisce, è stata utilizzata la classe *frbroo:Work*. Nel modello, un'opera ha un autore e un'area tematica di riferimento (per esempio: Retorica, Astronomia, Aristotelismo, ecc.).
Per rappresentare questa conoscenza sono state selezionate le classi *foaf:Person* e *skos:Concept* per mappare rispettivamente l'autore dell'opera e l'area tematica. Per collegare le classi tra di loro, sono state utilizzate due proprietà del *Dublic Core*: *dc:creator* per mettere in relazione l'opera all'autore e *dc:subject* per l'area tematica all'opera.

[24]Representing Content in RDF 1.0, http://www.w3.org/TR/Content-in-RDF10/, agosto 2015

[25] "cnt" è il namespace utilizzanto per rappresentare il contenuto in RDF.

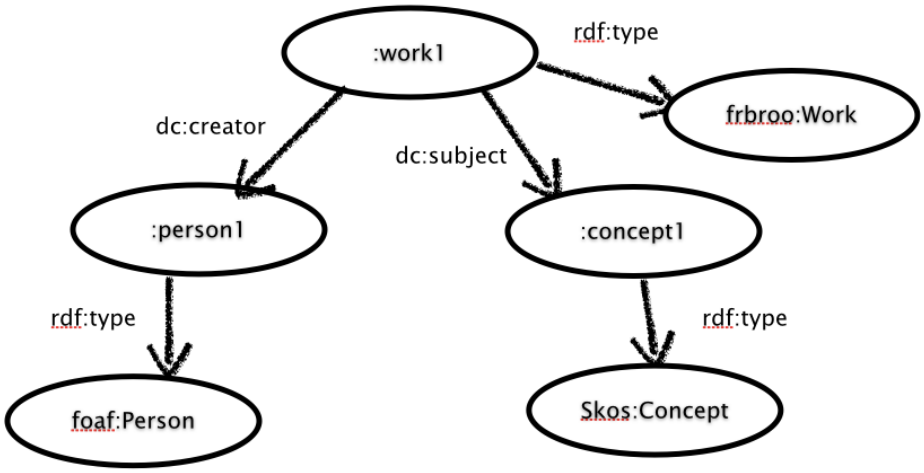



Fig. 8 - Descrizione delle relazioni tra opera, autore e area tematica sotto forma di grafo RDF

In ultimo per codificare la conoscenza relativa alle note dei commentari, sono state seguite le raccomandazioni del W3C che suggerisce l'adozione come standard *de facto* dell'*Open Annotation Model*. Quindi sono state selezionate le proprietà *oa:hasBody* e *oa:hasTarget*, per mappare il collegamento fra il corpo della nota (classe: *oa:Body*) e la classe *frbroo:ExpressionFragment*.

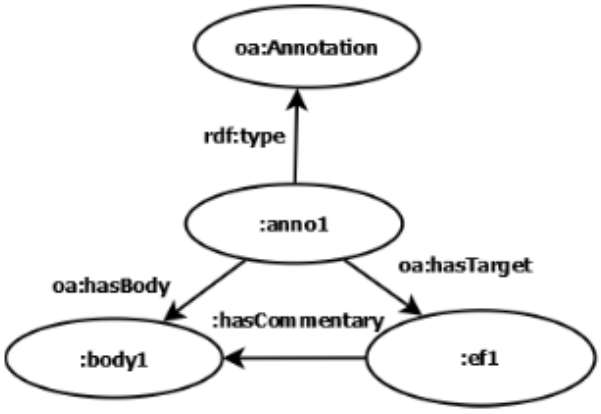



Fig. 9 - Descrizione delle relazioni tra nota e frammento della nota, sotto forma grafo RDF

Una volta finita la parte di modellizzazione della base di conoscenza, il passo successivo è stato quello di popolarla con i dati relativi alle fonti primarie citate da Dante nelle sue opere e segnalate dagli studiosi nelle loro note al testo.

Per farlo, è stato realizzato un programma in Java per trasformare il data set originario in formato DOC, in un corpus annotato in RDF. Il tool permette di estrarre automaticamente alcuni dati (numero del libro, del capitolo, del paragrafo e il testo dantesco) dai file in formato DOC trasformandoli in formato CSV. Successivamente, altre informazioni (autore della fonte primaria, titolo della fonte primaria, area tematica, frammento del testo citato) sono state aggiunte al file da tre studiosi dantisti. Il tool permette un ulteriore passaggio, ossia la trasformazione del file CSV così annotato in un file XML, modellato sull'ontologia.

L'ultima trasformazione riguarda il file XML che viene quindi triplificato in un file RDF pronto per essere immagazzinato nel triple store Virtuoso.
Nel triple store è stato aggiunto anche il Nuovo Soggettario, come repository di URI per le aree tematiche, il cui formato RDF ha permesso un più facile inserimento nella base di conoscenza.
In ultimo nello stesso triple store sono stati anche immagazzinati i testi danteschi, anch'essi trasformati tramite un apposito tool prima in XML e successivamente in triple RDF.

Per estrarre e visualizzare la conoscenza memorizzata nella biblioteca digitale , è stata sviluppata una applicazione web[26] con lo scopo di supportare gli studiosi nel realizzare una più completa enciclopedia dantesca.


Fig. 10 - Home page della web application Dante Sources

L'applicazione estrae conoscenza dall'ontologia attraverso interrogazioni SPARQL ed è in grado di produrre istogrammi al fine di mostrare i dati relativi alle fonti primarie citate da Dante. Una funzione Javascript permette inoltre di poter scaricare direttamente in dati in formato CSV, per poterli trattare direttamente.


Fig. 11 - Il grafico rappresenta la distribuzione delle fonti primarie citate nel Convivio

[26] DanteSources, http://perunaenciclopediadantescadigitale.eu/dantesources/index.html, agosto 2015

Inoltre uno SPARQL End Point fornisce agli studiosi interessati un accesso diretto all'ontologia così da poter effettuare ulteriori query e soddisfare direttamente curiosità e ricerche non implementate nell'applicaizone.

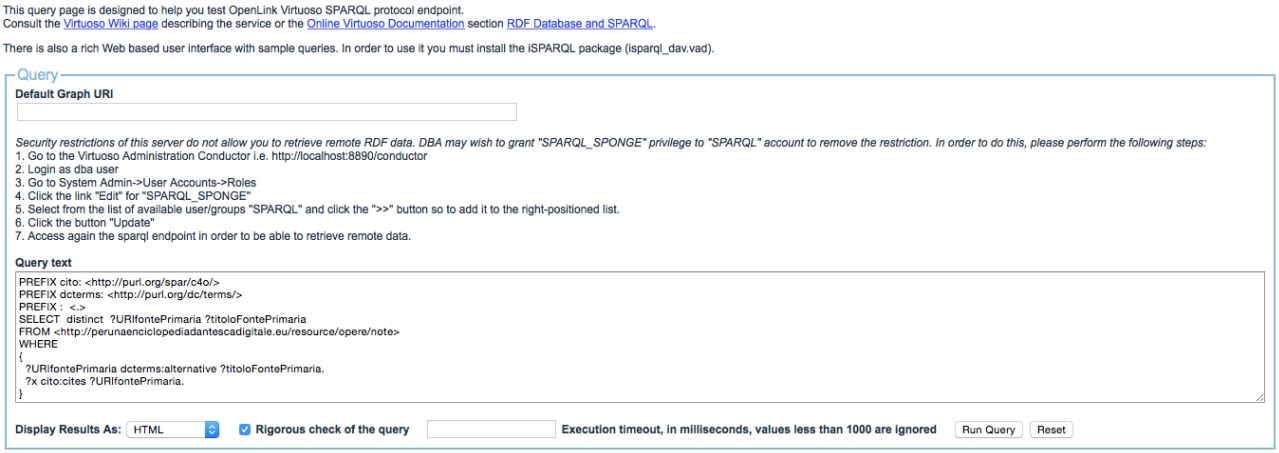



Fig. 12 - SPARQL End Point con interrogazione SPARQL relativa agli URI delle fonti primarie citate

| URIfontePrimaria | titoloFontePrimaria |
|---|---|
| http://perunaenciclopediadantescadigitale.eu/resource/Rhetorica_Novissima | Rhetorica Novissima |
| http://dbpedia.org/resource/Digest_(Roman_law) | Digesta |
| http://dbpedia.org/resource/Distichs_of_Cato | Disticha Catonis |
| http://dbpedia.org/resource/Consolation_of_Philosophy | De consolatione philosophiae |
| http://perunaenciclopediadantescadigitale.eu/resource/Enarrationes_in_psalmos | Enarrationes in Psalmos |
| http://dbpedia.org/resource/De_finibus_bonorum_et_malorum | De finibus bonorum et malorum |
| http://perunaenciclopediadantescadigitale.eu/resource/De_summo_bono | De summo bono |
| http://dbpedia.org/resource/Chronicon_(Jerome) | Chronicon |
| http://dbpedia.org/resource/Pauline_epistles | Lettere di Paolo |
| http://dbpedia.org/resource/Gospel_of_John | Vangelo di Giovanni |
| http://dbpedia.org/resource/Laelius_de_Amicitia | De amicitia |
| http://dbpedia.org/resource/Gospel_of_Luke | Vangelo di Luca |
| http://dbpedia.org/resource/Book_of_Wisdom | Sapienza |
| http://perunaenciclopediadantescadigitale.eu/resource/Facta_et_dicta_memorabilia | Facta et dicta memorabilia |
| http://dbpedia.org/resource/Summa_contra_Gentiles | Summa contra Gentiles |
| http://perunaenciclopediadantescadigitale.eu/resource/Lettera_a_Gianni_Bentivegna | Lettera a Gianni Bentivegna |
| http://dbpedia.org/resource/Book_of_Isaiah | Libro di Isaia |
| http://dbpedia.org/resource/Nicomachean_Ethics | Ethica Nicomachea |
| http://dbpedia.org/resource/Confessions_(St._Augustine) | Confessiones |
| http://dbpedia.org/resource/Physics_(Aristotle) | Physica |

Fig. 13 - Risultato della query SPARQL precedentemente lanciata

## Verso altre formalizzazione e oltre...

In futuro il gruppo di ricerca si occuperà di trasformare l'ontologia da RDF a OWL, in modo da aggiungere nuove informazioni logiche e portare l'ontologia ad un più completo livello di formalizzazione, così da permettere inferenze logiche e deduzioni che potrebbero apportare nuova conoscenza a partire dai dati codificati. Inoltre lo stesso gruppo, sta lavorando alla realizzazione di un nuovo modello, focalizzato sulla gestione delle narrazioni e degli eventi, con lo scopo di gestire in maniera automatica una nuova parte di conoscenza, continuando così la ricerca di nuove vie per formalizzare la conoscenza "umanistica" e renderla così disponibile per nuove ricerche informatiche.

# Bibliografia

Bartalesi, V., Locuratolo, E., Versienti, L., & Meghini, C. (2013, September). A preliminary study on the semantic representation of the notes to Dante Alighieri′s Convivio. In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities* (p. 4). ACM.

Berners Lee, T. (2006). Linked Data–Design Issues. *2006-07-27]. http://www. w3. org/DesignIssues/ LinkedData. html, 20/08/2015*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 28-37.

Binding, C., May, K., & Tudhope, D. (2008). Semantic interoperability in archaeological datasets: Data mapping and extraction via the CIDOC CRM. In *Research and Advanced Technology for Digital Libraries* (pp. 280-290). Springer Berlin Heidelberg.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the world wide web*, *7*(3), 154-165.

Burnard, L. The Text Encoding Initiative. (1995). An Overview. Geoffrey Leech, Greg Myers, and Jenny Thomas (eds.) *Spoken English on Computer: Transcription, Mark- up and Application*. London: Longman.

Buzzetti, D. (2002). Digital representation and the text model. *New Literary History*, *33*(1), 61-88.

Ciccarese, P., Ocana M., Das S., and Clark T. (2010). AO: An Open Annotation Ontology for Science on the Web. *Bio Ontologies 2010*. Boston MA, USA.

Ciotti, F. (2011). La rappresentazione digitale del testo: il paradigma del markup e i suoi sviluppi. *Lorenzo Perilli, Domenico Fiormonte (a cura di). In La macchina nel tempo: studi di informatica umanistica in onore di Tito Orlandi, Firenze, Le lettere*.

Ciotti, F. (2012). Web semantico, linked data e studi letterari: verso una nuova convergenza. *Quaderni DigiLab*, *2*(1), 243-276.

Della Valle, E., Celino, I., & Cerizza, D. D. (2008). *Semantic web: modellare e condividere per innovare*. Pearson Addison Wesley.

Doerr, M. (2002). The CIDOC CRM: an ontological approach to semantic interoperability of metadata. *AI Magazine, Special Issue*, *24*(3), 75-92.

Doerr, M., & LeBoeuf, P. (2007). Modelling intellectual processes: the FRBR-CRM harmonization. In *Digital libraries: Research and development* (pp. 114-123). Springer Berlin Heidelberg.

Eide, O., Felicetti, A., Ore, C. E., D'Andrea, A., & Holmen, J. (2008, February). Encoding cultural heritage information for the semantic web. procedures for data integration through cidoc-crm mapping. In *Open Digital Cultural Heritage Systems Conference* (p. 47).

Eide, O., & Ore, C. E. S. (2007). From TEI to a CIDOC-CRM conforming model: Towards a better integration between text collections and other sources of cultural historical documentation. *Digital Humanities*.

Erling, O., & Mikhailov, I. (2009). RDF Support in the Virtuoso DBMS. In *Networked Knowledge-Networked Media* (pp. 7-24). Springer Berlin Heidelberg.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

Gruber, T. (2009). Ontology. *Encyclopedia of database systems*, 1963-1965.

Jordanous, A., Lawrence, K. F., Hedges, M., & Tupman, C. (2012, June). Exploring manuscripts: sharing ancient wisdoms across the semantic web. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (p. 44). ACM.

Landow, G. P. (1991). *HyperText: The Convergence of Contemporary Critical Theory and Technology (Parallax: Re-visions of Culture and Society Series)*. Johns Hopkins University Press.

Madison, O. M. (2011). The IFLA Functional Requirements for Bibliographic Records. *Library Resources & Technical Services*, *44*(3), 153-159.

Miles, A., Matthews, B., Wilson, M., & Brickley, D. (2005, September). SKOS core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications* (pp. pp-3).

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, *17*, 33-43.

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013, May). Designing the W3C open annotation data model. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 366-375). ACM.

Shotton, D., and Peroni, S. (2015). DoCO, the Document Components Ontology. 20/08/2015. IRI=http://purl.org/spar/doco/.

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *J. Biomedical Semantics*, *1*(S-1), S6.

Sinclair, P., Addis, M., Choi, F., Doerr, M., Lewis, P., & Martinez, K. (2006). The use of CRM core in multimedia annotation.

Tillett, B. (2005). What is FRBR? A conceptual model for the bibliographic universe. *The Australian Library Journal*, *54*(1), 24-30.

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text c.org/Guidelines/P5/ Encoding and Interchange. TEI Consortium. http://www.tei-c.org/Guidelines/P5/, 20/08/2015