Il Semantic web nelle biblioteche e nel patrimonio culturale

Oreste Signore, < os@orestesignore.eu>

Summer School LDA
Libraries in the digital age: linked data technologies for a global knowledge sharing
Pula (Cagliari), 29 agosto - 1° settembre 2016

Presentazione: http://www.orestesignore.eu/education/lda/slides/lda3.html
Documento pdf: http://www.orestesignore.eu/education/lda/slides/lda3.pdf

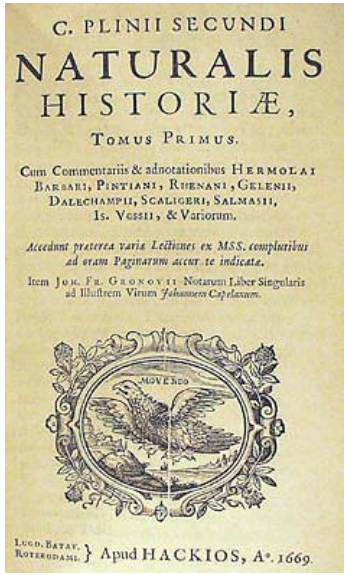
Formato XHTML realizzato usando il tool Slidy di Dave Raggett.
Slidy dovrebbe funzionare in tutti i browser moderni con Javascript abilitato. Usare freccia destra/sinistra per muoversi da una slide all' altra.
Vedi la pagina di aiuto di Slidy per ulteriori informazioni.

Contenuto

- Text retrieval
- Dublin Core
- Information integration
- CIDOC CRM
- Thesauri
- SKOS

Dublin Core in RDF

Il text retrieval


C. PLINII SECUNDI
NATURALIS HISTORIÆ,
TOMUS PRIMUS.
Cum Commentariis & adnotationibus HERMOLAI BARBARI, PINTIANI, RHENANI, GELENII, DALECHAMPII, SCALIGERI, SALMASII, Is. VOSSII, & Variorum.
Accedunt praterea varia Lectiones ex MSS. compluribus ad oram Paginarum accurate indicata.
Item JOH. FR. GRONOVII Notarum Liber Singularis ad Illustrem Virum Johannem Cuperum.
MOVEO
LUGD. BATAV. ROTERODAMI. } Apud HACKIOS, A°. 1669.

- La *Naturalis Historia* di Plinio il Vecchio
- In 37 libri
- Pubblicata nel 77 d.C.
- Il primo libro, pubblicato nel 79 d.C. dal nipote Plinio il Giovane, contiene il *sommario dei libri successivi* ed un *elenco delle fonti* per ciascun libro
- *I primi metadati?*

In principio fu lo scriptorium…

stat rosa pristina nomine, nomina nuda tenemus

…poi venne Gutenberg…

(http://upload.wikimedia.org/wikipedia/commons/b/b0/Gutenberg_Bible.jpg)

…e infine i calcolatori

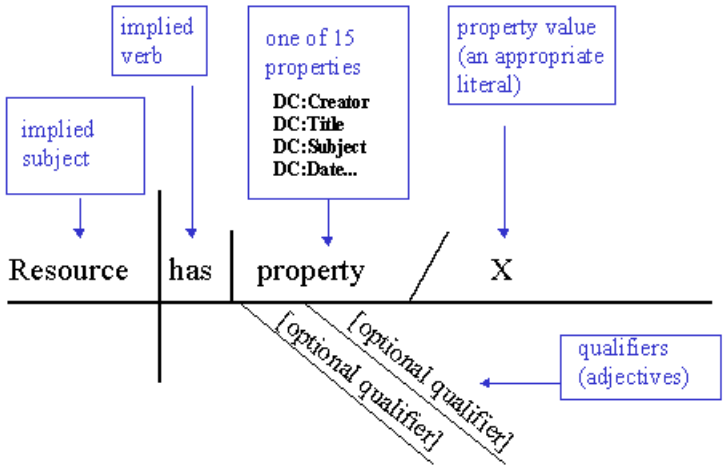
- Anni 1960: i grandi servizi di Information Retrieval
  - *organizzazione dei dati diversa*
  - *interfacce diverse*
- Anni 1970: Euronet Diane e il CCL
  - *un protocollo di comunicazione unico*
  - *Common Command Language: un insieme definito e comune di campi informativi (AU, TI, …) e di comandi (FIND, SHOW, …)*
- Il protocollo Z39.50
  - *lavori iniziati negli anni 1970, con successive variazioni nel 1988, 1992, e 1995*
  - *protocollo client-server*
  - *il profilo bib-1*
- 1985: Dublin Core
  - *15 "Property"*
  - *"qualifiers" (es. tipo e formato della data, vocabolario utilizzato)*

Dublin Core Metadata Initiative

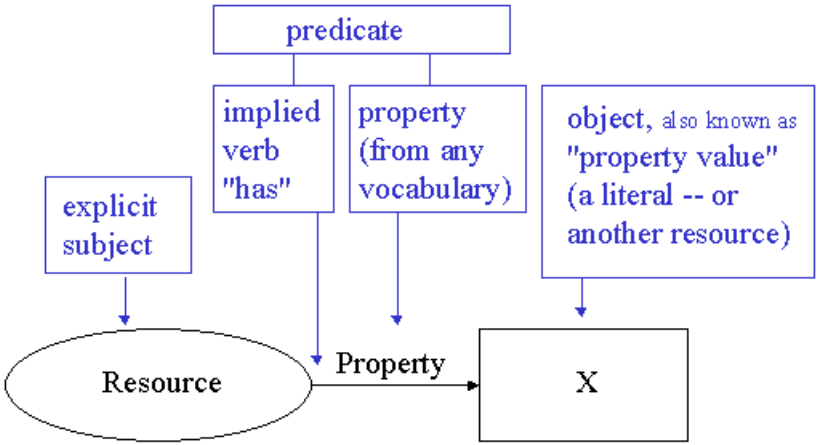
- Uno dei primi *vocabolari in RDF per metadati*
- Costituisce la *base* per i vocabolari per le *Digital Libraries distribuite*
- *Dublin Core Metadata Element Set*
  - *15 categorie generali (elements) per creare descrizioni semplici e facilmente comprensibili per la maggior parte delle risorse informative.*
  - *DCMES è solo elemento semantico di base per i metadati sul Web*
  - *le singole comunità hanno spesso bisogno di una semantica più ricca*
  - *altri metadati possono essere combinati con DCMES*

Il portale Dublin Core

Dublin Core: la grammatica



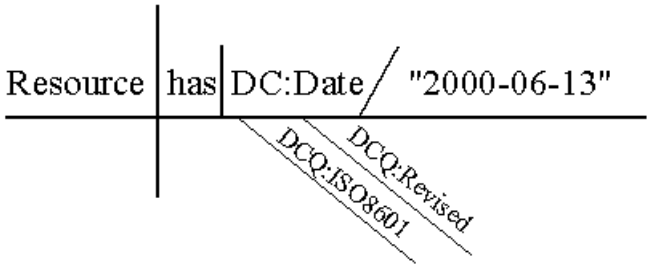
Dublin Core in RDF



Dublin Core: due esempi

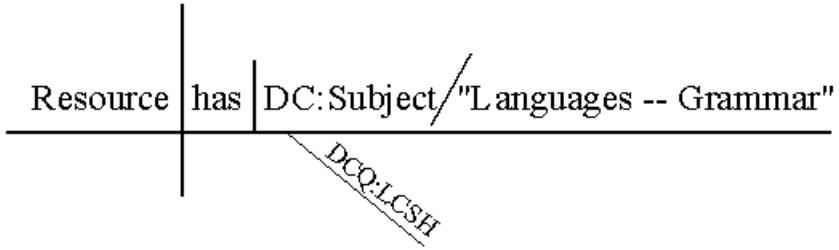
- Specifica del *tipo di data* (revised) e del *formato* (iso8601)

```
Resource has dcq:iso8601 dcq:revised dc:date '200-06-13'
```



- Specifica del *vocabolario* controllato utilizzato (Library of Congress Subject Headings)

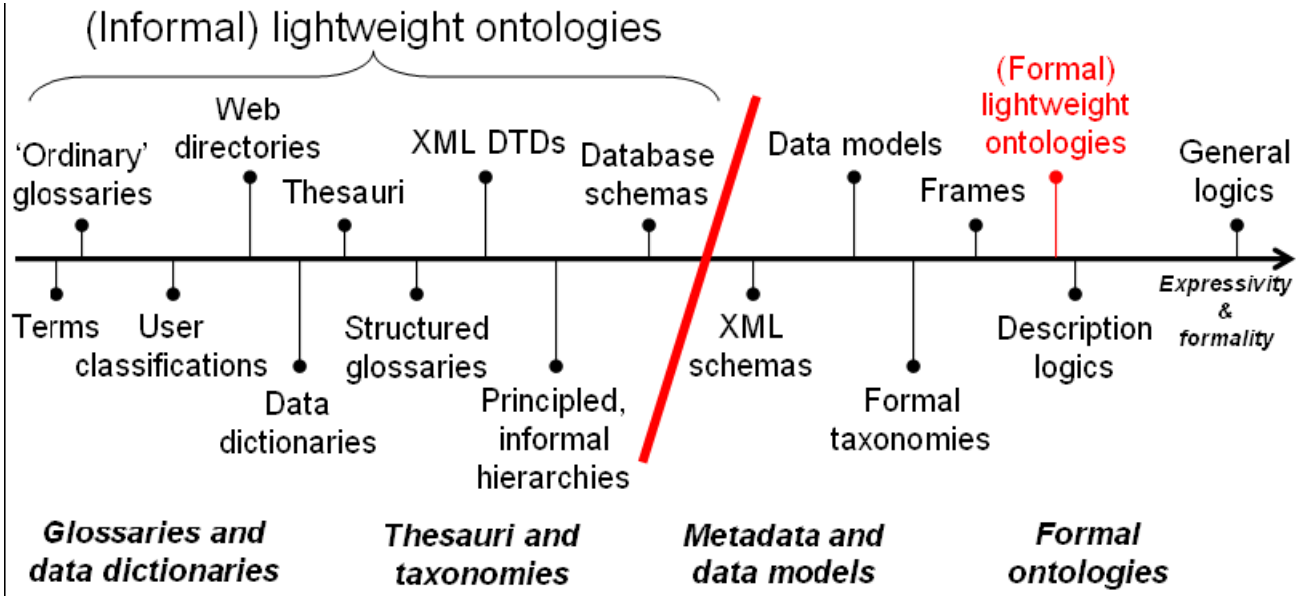
```
Resource has dcq:lcsh dc:subject 'Languages -- Grammar'
```



CIDOC-CRM

http://www.orestesignore.eu/education/lda/slides/cidoc.html

Thesaurus vs ontology



(from: Fausto Giunchiglia and Ilya Zaihrayeu: LIGHTWEIGHT ONTOLOGIES - October 2007 - Technical Report DIT-07-071)

Beware of false friends!

- Thesauri are often designed aiming to *more effective retrieval*, instead of formally *representing the knowledge*
- A thesaurus *is not automatically* an ontology
  - *Monumenti e siti archeologici*
    - Aree archeologiche
    - Monumenti archeologici
    - Parchi archeologici
    - Siti archeologici

    is a class with class-subclass relationships
  - *Prima età moderna*
    - Cinquecento
    - Seicento

    is not a class, but an instance, and cannot have sub-instances
- *Multiple inheritance* and *time dependent relationships* are also an issue
- See an example of temporal ontology and inference

Faceted thesauri

Faceted thesauri are similar in many ways to faceted classification systems. There are potentially differences in the intended use, as discussed in Sect. 4. However there is scope for using both in combination. Faceted thesauri can be used in both pre- and post-coordinated systems and can underpin both search and browsing applications.
Faceted thesauri belong to the family of KOS, which has been used by the library community in modelling for purposes associated with information retrieval applications. They provide a semantic structure at a suitable granularity for the general problem of search and retrieval. In such applications, where a fuzzy notion of "aboutness" is the basis for indexing or classifying a document, as opposed to an assertion of fact, the lightweight semantics of faceted thesauri and related KOS may be more suited than the formal semantics provided by AI ontologies, designed for precisely modelling the objects of interest in a domain. The SKOS standard representation, combined with other developments in standard identifiers and service protocols, now affords the combination of formal syntax and informal semantics, in Semantic Web applications and online applications generally. This offers a cost effective approach for annotation, search and browsing oriented applications that don't require first order logic.

(Douglas Tudhope & Ceri Binding: Faceted Thesauri, Axiomathes (2008) 18:211–222 DOI DOI 10.1007/s10516-008-9031-6)

Limitations of existing KOS

Lack of conceptual abstraction

thesauri and other traditional KOS are *collections of terms* (generic or domain-specific), ordered in a polyhierarchic lattice structure or a monohierarchic tree structure and interlinked with some very broad and basic relationships. The *distinction between a concept (meaning) and its lexicalizations (words) is not made consistently*, if at all, in such a system, and as such it does not reflect the ways humans understand the world in terms of meaning and language

Limited semantic coverage

most thesauri do not differentiate concepts into types or categories (such as living organism, substance, or process) and have a *very limited set of relationships between concepts*, distinguishing only between hierarchical relationships, i.e. NT/BT, and associative relationships, i.e. RT. These very *rudimentary relationships are not powerful enough* to guide a user in meaningful information discovery on the Web or to support inference. They d*o not reflect the conceptual relationships that people know* and that can be used by a system to suggest concepts for expanding the query or making it more specific.

Lack of consistency

since the *relationships* in thesauri *lack precise semantics*, they are applied inconsistently, both creating ambiguity in the interpretation of the relationships and resulting in an overall *internal semantic structure that is irregular and unpredictable*. Many of the NT/BT hierarchical relationships could, for example, be resolved to the non-hierarchical RT relationship, and vice versa

Limited automated processing

traditionally thesauri were *designed for indexing and query formulation* by people and not for a*utomated processing*. The *ambiguous semantics* that characterizes many thesauri makes them *unsuitable for automated processing*

Brian Vickery: A note on knowledge organisation, [web] [local]

SKOS

See other slides: [pdf] [ppt]

Conclusioni

- La rappresentazione della conoscenza è essenziale per una elaborazione automatica delle informazioni esistenti sul web
- Le *tecnologie del Semantic Web* (RDF, RDFS, OWL) consentono di *rappresentare*, *esportare* e *condividere* la conoscenza in maniera interoperabile
- Molte iniziative nel settore biblioteche beni culturali

Grazie per l'attenzione

Domande?

Se non è sul Web non esiste ...

... troverete sul sito (http://www.orestesignore.eu), sezione education
le *slide* (http://www.orestesignore.eu/education/lda/slides/lda3.html