ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

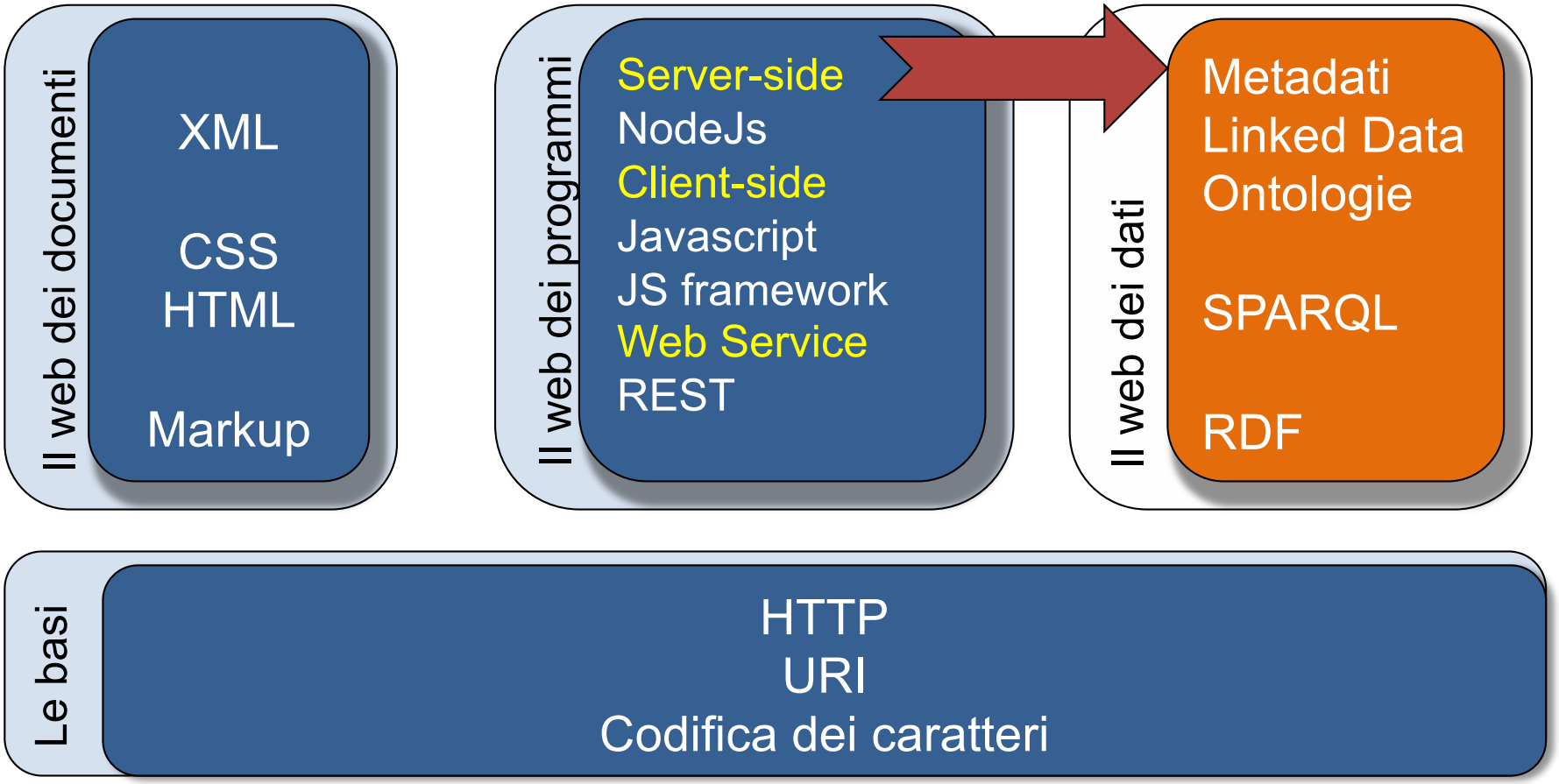
Verso il Semantic Web: metadati e vocabolari

Fabio Vitali

Corsi di laurea in Informatica

Alma Mater – Università di Bologna

Argomenti delle lezioni



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Indice degli argomenti

Oggi parliamo di:

- La necessità del Semantic Web
- PICS
- L'organizzazione delle informazioni
- La struttura del Semantic Web

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

L'inconfrontabilità del sapere

La difficoltà della ricerca ha molto a che fare con alcuni problemi specifici:

- Differenza tra termini usati nella ricerca e termini usati nei documenti
  - Io cerco "mal di testa", il documento contiene "emicrania"
- Molteplicità di termini usati per stile o abitudine
  - Università degli Studi di Bologna, Alma Mater, l'ateneo bolognese, il rettorato, in Via Zamboni, gli Istituti di Ricerca regionali, le Istituzioni di Istruzione Superiore italiane, ecc.
- Ambiguità intrinseca di alcuni termini
  - "L'importanza della pesca nell'economia della provincia": a Parma (coltivano frutta) interpretano la frase diversamente che a Ravenna (c'è un porto).

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

I vantaggi della serializzazione

I linguaggio di markup (XML, HTML) o di dati strutturati (JSON o YAML) forniscono una soluzione chiara e indipendente dall'applicazione ad alcuni problemi nella gestione di dati:

- **Codifica**: esiste oggi un modello chiaro, non ambiguo e universale per la codifica dei caratteri (UTF-8), e un meccanismo di ordinamento dei dati privo di dipendenze dalle architetture hardware su cui il documento viene memorizzato.
- **Delimitazione**: i tag sono chiaramente e visibilmente diversi dai dati. C'è un supporto evidente per dati che non sono contenuto (metadati) e i delimitatori hanno chiari meccanismi di escaping.
- **Etichettatura**: i documenti sono organizzati gerarchicamente e aggiungono ai dati etichette che permettono di dare senso ai frammenti e permettono di includere e gestire anche le eccezioni.
- **Semantica**: ???

Dov'è il significato?

- Non nei dati
- Non nel markup (i tag)
- Non nel documento che specifica il vocabolario ristretto e i suoi vincoli (il DTD o il modello delle classi)
- Non nello strumento HTML

Ma:

- nell'applicazione che gestisce il contenuto del documento di markup…
- … e nella mente dell'essere umano che scrive o legge il contenuto del documento di markup, sia nel formato sorgente che nel formato reso in un browser.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

La risposta: il semantic web

L'arma definitiva per l'appassionato fan della dichiaratività.

Ancora più astratto, ancora più sintattico, ancora più privo di significati e comportamenti predefiniti di SGML e XML.

Nel semantic web l'universo è formato da affermazioni su qualche classe o proprietà del dominio del discorso. Queste affermazioni, ovviamente, non hanno significato predefinito, e richiedono strumenti software o cervelli per fornirsene.

RDF è un meccanismo generico per esprimere affermazioni, OWL è un meccanismo generico per organizzarle e strutturarle.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Verso il Semantic Web

Dalla pornografia all'intelligenza artificiale

L'estensione del WWW ha uno scopo: generare informazioni che non siano soltanto destinati alla lettura, ma che possano essere riutilizzati per applicazioni automatiche.

Non c'è niente in un documento HTML che indichi l'argomento trattato o la fonte delle informazioni. L'unico tipo di ricerca che si può fare su un documento è la ricerca sul contenuto. Questo non è sufficiente nella maggior parte delle volte: usando un motore di ricerca si ottiene un qualche migliaio di hit, la maggior parte dei quali non serve assolutamente a niente.

Le meta informazioni permettono agli autori di specificare informazioni sui loro documenti che siano non soltanto leggibili, ma anche interpretabili in maniera intelligente dalle applicazioni di rielaborazione, e soprattutto dai motori di ricerca.

L'utilizzo sistematico di meta-informazioni, ci dicono, ci porterà alla prossima generazione di Web: il Web semantico

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

La responsabilità autoriale sul Web

Nel 1995 il W3C era preoccupato perché l'esistenza di siti con materiale “sconveniente” (pornografia, hacking, estremismo politico, ecc.) stava danneggiando lo sviluppo della rete.

Il Web, per sua natura, non permette meccanismi di controllo e valutazione preventiva dei contenuti, né meccanismi seri di identificazione degli autori di un documento. Questo significa che chiunque voglia inserire dati e documenti “opinabili” può farlo senza sforzo, rischio personale e ricusabilità.

In particolare si temeva in molte sedi l'intervento pesante dei governi, la nascita di un organismo centrale (governativo o meno, comunque controllabile politicamente) che decidesse cosa fosse lecito mettere e cosa no.

L'idea di un organismo centrale è radicalmente opposta alla filosofia dell'end-to-end argument alla base del Web e di Internet in generale, in quanto collo di bottiglia per l'uso e la distribuzione di informazioni sulla rete.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PICS (1)

PICS (Platform for Internet Content Selection) fu la risposta del W3C alle lamentele sulla qualità delle informazioni sul web poste da associazioni di genitori, associazioni di educatori, associazioni religiose e politiche, ecc.

PICS era un brillante meccanismo per coinvolgere le stesse associazioni. PICS richiedeva il contributo attivo e indispensabile delle associazioni di categoria.

PICS non invocava nessun meccanismo di censura sui contenuti (peraltro impossibile in una realtà veramente globale), ma proponeva l'uso di meccanismi di valutazione (rating) dei contenuti del sito, e individuava nei singoli utenti il compito di attivare o meno un controllo.

Inoltre, non forniva un sistema fisso di categorie e valori, ma permetteva a ciascuna associazione di definire le categorie rilevanti e le voci di ciascuna categorie.

PICS (2)

Infine, non forniva un meccanismo centralizzato per memorizzare i rating di ciascun sito, ma un'architettura distribuita secondo la quale il seguace, il genitore, la scuola si abbonava ad un particolare sito di rating, stabiliva le politiche di selezione del contenuto visibile, e era autorizzata a visitare soltanto quel sottoinsieme di pagine innocuo per il fornitore di rating.

Come detto, PICS non stabiliva categorie pre-definite, ma un meccanismo generale per identificare categorie. Era quindi possibile usare PICS non solo per fornire censure locali a categorie, ma in generale per esprimere concetti di qualunque tipo sui siti.

Ad esempio, l'associazione delle biblioteche universitarie americane forniva attraverso PICS un rating dell'autorevolezza delle informazioni tecniche contenute sui siti: non per fare censura, ma per guida e valutazione delle informazioni.

PICS (3)

Questo era in definitiva un meccanismo funzionante e sofisticato per esprimere meta-informazioni (cioè valutazioni) su siti Web.

Tuttavia PICS richiedeva di identificare in anticipo tutte le categorie rilevanti e tutti i valori da assegnare a qualunque categoria.

Questa limitazione, più il successo di XML, più la nascente idea di Web semantico, hanno portato alla creazione di RDF.

Tuttavia PICS non ha avuto il successo sperato. Prima è stato sostituito da POWDER (*Protocol for Web Description Resources*) esplicitamente basato su RDF e poi anche POWDER non è mai stato veramente adottato dalla comunità, che invece ha creato mille modelli diversi e incompatibili tra loro .

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Ma prima: l'organizzazione delle informazioni

L'organizzazione delle informazioni

Per capire a cosa serve il Semantic Web, dobbiamo prima introdurre alcuni termini:

- Metadati e metainformazioni
- Vocabolario controllato
- Tassonomia
- Thesaurus (o tesauro)
- Classificazione a faccette
- Ontologia
- Folksonomia

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

I metadati

Letteralmente: dato a proposito di qualcosa (e.g., un documento o un altro dato)

Ci sono due possibili visioni del concetto di metadato su un documento:

- Una qualunque informazione a proposito del documento nella sua interezza.
  - Ad esempio, il titolo o l'autore di un documento
- Una qualunque informazione che non appartenga originariamente al documento
  - Ad esempio, che il documento parli di calcio o che la terza frase sia una menzogna.

Le due definizioni coesistono, creano confusione e questa confusione ce la terremo per un bel po'.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Termini dei metadati

Esterni vs. interni vs. riflessivi

- Un metadato esterno parla di una risorsa terza, a cui dà un significato.
- Un metadato interno è contenuto nel documento stesso,
- Un metadato riflessivo è interno e parla del documento ***e pure*** di se stesso.

Autoriale vs. redazionali (o editoriali)

- I metadati autoriali che vengono forniti dall'autore del documento (N.B.: è diverso da *autorevole*).
- I metadati redazionali (*editorial* in inglese) vengono forniti da membri della catena di produzione del documento, redattori o *editor* del documento.

Oggettivi vs. soggettivi (o interpretativi)

- I metadati su cui non c'è discussione o dubbio, sono detti oggettivi.
- I metadati che rappresentano un'interpretazione personale, su cui ci può essere dissenso o critica, sono detti soggettivi o interpretativi.

Dissettivi o anti-dissettivi

- Un metadato dissettivo si applica a ogni parte del documento (se Fabio Vitali è l'autore del documento, è autore di ogni sua frase).
- Un metadato anti-dissettivo si applica al documento ma non a tutte le sue parti (in un documento che parla di calcio non è vero che ogni frase parla di calcio)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Termini dei metadati

Esterni vs. interni vs. riflessivi

- Un metadato interno è contenuto nel documento stesso,
- Un metadato esterno parla di una risorsa terza, a cui dà un significato.
- Un metadato riflessivo parla del documento ***e pure*** di se stesso.

Autoriale vs. redazionali (o editoriali)

- I metadati autoriali che vengono forniti dall'autore del documento.
- I metadati redazionali (*editorial* in inglese) vengono forniti da membri della catena di produzione del documento, redattori o editor del documento.
- *Ad esempio "I Promessi Sposi" è un titolo autoriale, mentre "Divina Commedia" è un titolo redazionale.*

Oggettivi vs. soggettivi (o interpretativi)

- I metadati su cui non c'è discussione o dubbio, sono detti oggettivi.
- I metadati che rappresentano un'interpretazione personale, su cui ci può essere dissenso o critica, sono detti soggettivi o interpretativi.

Dissettivi o anti-dissettivi

- Un metadato dissettivo si applica a ogni parte del documento (se Fabio Vitali è l'autore del documento, è autore di ogni sua frase).
- Un metadato anti-dissettivo si applica al documento ma non a tutte le sue parti (in un documento che parla di calcio non è vero che ogni frase parla di calcio)

Termini dei metadati

Esterni vs. interni vs. riflessivi

- Un metadato interno è contenuto nel documento stesso,
- Un metadato esterno parla di una risorsa terza, a cui dà un significato.
- Un metadato riflessivo parla del documento ***e pure*** di se stesso.

Autoriale vs. redazionali (o editoriali)

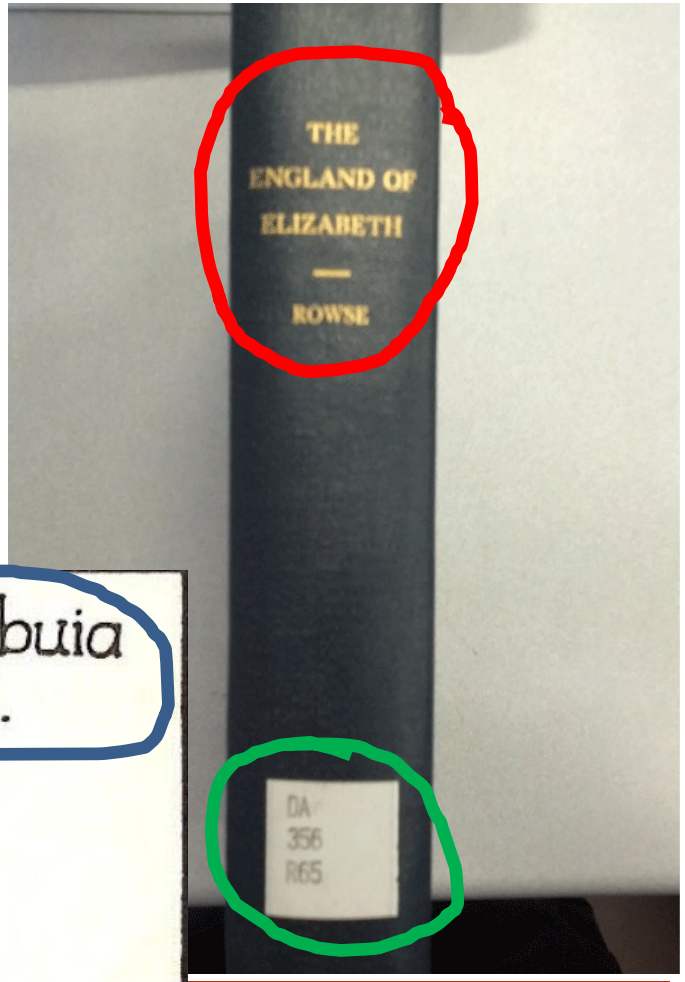
- I metadati autoriali che vengono forniti dall'autore del documento.
- I metadati redazionali (*editorial* in inglese) vengono forniti da membri della catena di produzione del documento, redattori o editor del documento.
- *Ad esempio "I Promessi Sposi" è un titolo autoriale, mentre "Divina Commedia" è un titolo redazionale.*

Oggettivi vs. soggettivi (o interpretativi)

- I metadati su cui non c'è discussione o dubbio, sono detti oggettivi.
- I metadati che rappresentano un'interpret
  può essere dissenso o critica, sono detti s

Dissettivi o anti-dissettivi

- Un metadato dissettivo si applica a ogni p
  Vitali è l'autore del documento, è autore d
- Un metadato anti-dissettivo si applica al d
  sue parti (in un documento che parla di ca
  che ogni frase parla di calcio)



Usi e problemi dei metadati

Usi

- Classificare, catalogare, organizzare documenti
- Fare ricerche su concetti discussi nei documenti invece che su parole presenti nei documenti
- Fare analisi e statistiche su basi documentali massicce

Problemi

- Ambiguità dei valori usati nei metadati
- Ambiguità degli scopi dei metadati
- Varietà eccessiva di termini nei metadati.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Vocabolario controllato

Anche: *linguaggio di indicizzazione*

Alcuni metadati (ad esempio l'identificazione dell'autore di una risorsa) richiedono valori da un insieme aperto (tutti i nomi di persone al mondo)

Altri metadati richiedono che i valori siano compresi in un insieme di valori precisi:

- Dotati di significato e di applicabilità
- Non ridondanti
- Non ambigui
- Completi rispetto al dominio dei valori possibili

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Tassonomia (1)

Termine inventato da Carlo Linneo nel XVIII secolo per la classificazione degli esseri viventi.

La ***tassonomia*** crea una gerarchia tra i termini di un vocabolario controllato, in grado di esplicitare relazioni di specificità o generalità tra i termini.

Ad esempio:

- **Romeo** è un *soriano*
- Un *soriano* è un *gatto*,
- Un *gatto* è un *felino*,
- Un *felino* è un *mammifero*,
- Un *mammifero* è un *animale*.

La tassonomia non cambia il metadato, né i valori possibili (che sono sempre appartenenti al vocabolario controllato)

Fornisce un ordine e una organizzazione ai termini del vocabolario controllato.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Tassonomia (2)

Alcune (molte) tassonomie introducono termini non istanziabili - cioè non usabili come valori di metadati, unicamente come raccordo tra i valori possibili.

- non esiste nessun animale che sia un mammifero senza essere anche un felino, o un canide, o un primate, ecc.
- In OOP le chiameremmo *classi astratte.*

La tassonomia è un'operazione linguistica, non scientifica:

- fa parte del modello della realtà, non della realtà.
- Serve agli umani per comprendere e usare la realtà, ma non ha nessuna necessaria attinenza con la realtà vera.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

I tesauri (o thesauri)

In breve, un ***tesauro*** è una tassonomia a cui si aggiungono relazioni di pari livello tra termini.

Definizione di tesauro (ISO 2788-1986) «il thesaurus è il vocabolario di un "linguaggio di indicizzazione" controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni "a priori" fra i concetti»

Il tesauro permette di

- trovare un punto di incontro tra lessico dell'autore e lessico dell'utente,
- Proporre una relazione biunivoca tra termine e concetto, così da ottenere *univocità semantica*:
- un termine per ogni concetto, un concetto per ogni termine.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

I tesauri (2)

L'univocità semantica elimina i problemi connessi con l'uso del linguaggio naturale

- ridondanze, ambiguità, polisemie, omonimie, omografie
- queste caratteristiche garantiscono ricchezza ed espressività al linguaggio naturale, ma rendono difficile l'organizzazione funzionale dei motori di ricerca.

I tesauri generalizzano la gerarchia tra termini della tassonomia in un generico insieme di relazioni tra termini, alcuni gerarchici, altri no.

- Relazione gerarchica
- Relazione preferenziale o sinonimica
- Relazione associativa

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazioni tra termini nei tesauri

Relazione gerarchica

- Relazione di subordinazione all'interno di uno stesso albero gerarchico.
- Es.: **matematica/geometria, felini/gatti, veicoli/automobili**

Relazione preferenziale o sinonimica

- Identifica tra più termini per lo stesso concetto quello preferito. Identifica classi di equivalenza (sinonimi)
- Es.: **regola/norma, week-end/finesettimana, mal di testa/cefalea**

Relazione associativa

- Relazione residuale, identifica tra due termini una relazione né di equivalenza, né di subordinazione, ma comunque esistente ed innegabile.
- Es.: **barca/nave, ecologia/inquinamento, ecc.**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione preferenziale (1)

Identifica un gruppo di equivalenza tra termini, tra i quali si sceglie il termine preferito. Gli altri vengono detti termini non preferiti o sinonimi.

La relazione tra termine non preferito (**NPT**) e termine preferito (**PT**) si chiama **USE**. La relazione inversa si chiama **UF** (Use For)

Thesaurus **Tesauro**

USE Tesauro **UF Thesaurus**

Rientrano in questa categoria:

- Sinonimia vera — **regola e norma**
- Varianti ortografiche — **psicoanalisi e psicanalisi**
- Sigle e acronimi — **CNR e Centro Nazionale delle Ricerche**
- Preferenza linguistica
  - T. straniero e italiano — **week-end e finesettimana**
  - T. attuale e antico — **bicicletta e velocipede**
  - T. comune e scientifico — **mal di testa e cefalea**
  - T. di origini diverse — **poliglotta e multilingue, antologia e florilegio**
  - T. comuni e marche — **penna a sfera e biro, fotocopiatrice e xerox**
  - Varianti recenti — **telefonino, telefono cellulare, smartphone**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione preferenziale (2)

Oltre alla sinonimia propria, con relazione preferenziale si possono mettere in relazione anche termini non strettamente sinonimici (*sinonimia convenzionale*), in cui i termini sono considerati sinonimi solo all'interno del contesto dei documenti gestiti da tesauro.

Possiamo distinguere:

- Quasi-sinonimia **punizione, ammenda, sanzione, pena**
- *Upward posting* **TIR e camion**

  (si parla di upward posting *per termini in relazione gerarchica di cui non interessa gestire la specificità.*
- antinomia **guerra e pace, amore e odio, malattia e salute**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione gerarchica

Descrive un albero di termini, tra i quali esiste un rapporto di subordinazione o sovraordinazione. I termini subordinati vengono anche detti iponimi, quelli sovraordinati vengono anche detti iperonimi.

La relazione tra termine e termine inferiore è NT (narrower term), tra termine e termine superiore è BT (broader term)

Geometria

NT1 Geometria euclidea

NT1 Geometria non euclidea

NT2 Geometria iperbolica

NT2 Geometria ellittica

Geometria ellittica

BT Geometria non euclidea

BT Geometria

BT Matematica

Rientrano in questa categoria:

- Relazione generica o genere/specie *(is_a)*
- Relazione partitiva o parte/tutto *(has_a)*
- Relazione esemplificativa o classe/istanza *(instance_of)*

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione gerarchica ***is_a***
(genere/specie)

Detta anche relazione genere/specie o relazione **is-a** (**è-un**).
Sigla specifica: **BTG** e **NTG**.

E' il legame che esiste tra una categoria e i suoi membri. Perché sia corretta, è necessario che tutte le istanze del termine subordinato siano sotto-concetti del termine sovraordinato.

Ad esempio, felino/gatto è una coppia di termini in relazione *generica*, mentre animale domestico/gatto non lo è, perché esistono gatti selvatici.

Questa differenza assoluta, però, può non essere vera nell'ambito dei documenti che vengono trattati (se per design non si parla di animali selvatici, la relazione *is-a* vale anche per animale domestico/gatto.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione gerarchica ***has_a***
(parte/tutto)

Detta anche relazione parte/tutto o relazione **has-a** (**ha-un**). Sigla specifica **BTP** e **NTP**.

E' il legame che esiste tra un concetto complesso e i suoi componenti. Perché sia corretta, è necessario che tutte le istanze del termine subordinato implichino il termine sovraordinato. Ovvero non possono esistere due esempi dello stesso termine all'interno di gerarchie diverse.

In generale questo è possibile solo in quattro casi:

- Organi del corpo (**sistema circolatorio - vene**)
- Nomi geografici (**Italia - Emilia-Romagna - Bologna**)
- Discipline (**scienze - biologia - botanica**)
- Strutture sociali (**divisione - reggimento**)

mentre è inappropriato in altre situazioni apparentemente simili:

- **Automobile – motore** *(i motori vengono usati anche non nelle automobili)*

Altrimenti è possibile solo per organizzazioni specifiche interne al tesauro.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione gerarchica ***part_of***
(gerarchia composizionale)

E' il legame che esiste tra un concetto complesso e le parti di cui è composto. Si differenzia dal precedente perché non è vero che le istanze del termine subordinato implichino il termine sovraordinato, e quindi lo stesso termine può comparire in gerarchie diverse (quindi tecnicamente non è un albero ma un grafo diretto aciclico).

Il modello di tesauro secondo ISO 2788-1986 non comprende la gerarchia composizionale, perché non forma necessariamente una gerarchia:

- **Automobile**
  - **motore**
  - **pneumatici**
- **Aereo**
  - **motore**
  - **ali**

Il modello di tesauro secondo ISO 2788-1986 non comprende la gerarchia composizionale, proprio perché non forma necessariamente una struttura ad albero.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Relazione gerarchica ***instance_of*** *(classe/istanza)*

Detta anche relazione classe/istanza o specie/esempio *(instance_of)*.

E' il legame che esiste tra una classe ed un suo individuo (classe di uno), tra una categoria e un nome proprio.

La relazione si indica con **NT1** (Narrower term individual). Ad esempio:

Pontefici

- **NT1 Giovanni XXIII**
- **NT1 Paolo VI**
- **NT1 Giovanni Paolo I**
- **NT1 Giovanni Paolo II**
- **NT1 Benedetto XVI**
- **NT1 Francesco I**

Attori

- **NT1 Brad Pitt**
- **NT1 Tom Cruise**
- **NT1 Matt Damon**
- **NT1 John Malkovich**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Monogerarchie e poligerarchie

Le relazioni gerarchiche possono assumere strutture complesse nel momento in cui assumiamo una classe specifica come derivato da più classi generiche.

E' importante allora mettere ben in chiaro se si adottano gerarchie multiple o semplici. Ad esempio:

- **Organo**
  - **BTG Strumenti a fiato**
  - **BTG Strumenti a tastiera**
    - **BTG Strumenti**
- **Strumenti a fiato**
  - **BTG Strumenti**
  - **NTG Organo**
  - **NTG Flauto**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Classificazioni a faccette (1)

Un termine introdotto da S. R. Ranganathan negli anni '30 per indicare la possibilità di descrivere un oggetto complesso attraverso un insieme di affermazioni appartenenti ad uno schema fisso di proprietà, ciascuna delle quali in grado di usare valori da un apposito tesauro.

Ogni risorsa viene descritta dunque dalla tupla di tutti i valori specificati nell'ordine definito dallo schema designato.

Attenzione: lo schema deve anche essere in grado di arrivare ad identificare, e non solo descrivere, una specifica risorsa individuale. Cioè data una tupla completa, debbo trovare zero o una risorsa, non di più.

Questo è solitamente realizzato identificando una (o più) proprietà dette chiave.

Classificazione a faccette (2)

Ogni volta che prevediamo una molteplicità di fattori descrittivi, indipendenti gli uni dagli altri, con cui classifichiamo una risorsa, abbiamo una classificazione a faccette

Ad esempio *Dublin Core*:

- Tipo Documento: slide
- Destinatari: studenti di Tecnologie Web
- Titolo: Lezione di Tecnologie Web
- Autore: Fabio Vitali
- **URL: http://vitali.web.cs.unibo.it/twiki/ … /Metadati.pptx**
- Formato: MS Powerpoint
- Data di creazione: 27/4/2022

chiave

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Ontologie

Il culmine della progressione che abbiamo visto finora.

Il principio fondamentale è che il valore di una proprietà non deve necessariamente essere un termine da un vocabolario controllato, ma può essere un riferimento ad un risorsa, a sua volta descritta da una serie di proprietà.

- Non è la stringa “Fabio Vitali” ad essere l'autore di queste slide, ma quella persona il cui nome è la stringa “Fabio Vitali”.

Un'ontologia allora è una composizione di classi, in relazione con le altre attraverso il riferimento esplicito (diretto o indiretto) espresso nelle proprietà di uno schema di classificazione a faccette.

Quando i valori di una proprietà sono termini, usiamo un tesauro, altrimenti relazioni ad altre classi dell'ontologia.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dai metadati alle ontologie (1/4)

Una collezione di metadati:

- **Fabio Vitali, Bologna University, 5 maggio 2023, Informatica, Tecnologie Web, Università di Bologna, PowerPoint, dispense, metadati, Metadata and ontologies, raw metadata collections vs. ontologies, using metadata.**

Un grande casino:

- Che cos'è un Fabio Vitali?
- Perché due volte Università di Bologna?
- Perché alcune parole sono in Italiano e altre in inglese?
- Le dispense sono stanze? Cibi? O un tipo di documento?
- Perché Informatica e Metadata appaiono varie volte?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dai metadati alle ontologie (2/4)

Vocabolario controllato

- Se restringiamo i metadati a specifici insiemi di termini, riduciamo un po' di ambiguità e di polisemie (***dispense***) e di varietà (***Bologna University, Università di Bologna***)

Tassonomie/Thesauri

- Se organizziamo i termini in un vocabolario controllato possiamo almeno dedurre i significati di qualche concetto
  - **Productivity Tools**
    - **Computer Applications**
      - » **Microsoft PowerPoint**
  - **Documenti testuali**
    - **Documenti didattici**
      - » **dispense**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dai metadati alle ontologie (3/4)

Classificazione a faccette

- Associare etichette descrittive ad ogni termine ci permette di fare alcune deduzioni forti
  - **Author: Fabio Vitali**
  - **Title: Metadati**
  - **Subject: Metadata and ontologies, raw metadata collections vs. ontologies, using metadata**
  - **Format: Microsoft PowerPoint**
  - **Date: 5 maggio 2023**
- Se poi restringiamo le etichette al set stabilito da un formato (ad esempio Dublin Core) possiamo garantire che le etichette saranno presenti sempre e sistematicamente in tutti i documenti di una collezione.

Dai metadati alle ontologie (4/4)

Ontologie

- Se poi specifichiamo che alcuni valori di metadati sono non strighe, ma riferimenti a concetti complessi, potremmo evitare ulteriori ambiguità e associare metadati e proprietà a quei valori.
- Ad esempio, Fabio Vitali non è una stringa, ma il nome di un'istanza di una classe detta Persona.
  - Se ho il riferimento corretto alla Persona, posso anche scrivere o riconoscere scritture diverse, come F. Vitali, Vitali Fabio o "il docente di Tecnologie Web".
  - Anche se esistono più Fabio Vitali nello stesso contesto, possiamo lo stesso identificare con esattezza la persona.
  - Volendo possiamo utilizzare meccanismi diversi per identificare la stessa persona, come il suo codice universitario o il suo codice fiscale, e comunque ci riferiamo sempre alla stessa persona.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Usare le ontologie

```
{
    "@context": {
      "@vocab": "http://www.fabiovitali.it/"
    },
    "@type": "document",
    "author": {
      "@type": "person",
      "name" : "Fabio Vitali",
      "affiliation": {
         "@type": "organization",
         "name":  "Università di Bologna"
      }
    },
    "title": "Metadati",
    "coverage": {
      "@type": "lesson",
      "date": "5 maggio 2023",
      "context": {
         "@type": "course",
         "authority": {
           "@type": "organization",
           "name":  "Università di Bologna"
         },
         "title": ["Tecnologie Web"]
      }
    },
    "subject": [
      "Metadata and ontologies",
      "Raw metadata collections vs. ontologies",
      "Using metadata"
    ]
}
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Oltre le ontologie?

Le strutture concettuali viste in precedenza (vocabolario controllato, tassonomia, tesauro, classificazione a faccette, ontologia), indipendentemente dalla loro caratterizzazione:

- Richiedono personale qualificato per generare e gestire la strutture (vocabolario, concetti e relazioni)
- Richiedono contemporaneamente competenza di dominio e competenza di classificazione.
- Può categorizzare solo sui vocaboli e le relazioni previsti.
- Bisogna che gli utenti siano d'accordo sulla concettualizzazione
- Ogni concettualizzazione prematura conduce ad un modello incompleto e difficilmente estendibile.
- Analogamente, è difficile progettare una concettualizzazione in continua evoluzione.
- Richiedono dunque una progettazione completa e dettagliata prima di iniziare a valutare e descrivere le singole risorse descritte.

Complessivamente, sono un approccio costoso, ingessato, non democratico, centralizzato e riduzionistico. Inoltre scala male su dimensioni veramente grandi (ad esempio il World Wide Web).

Le folksonomie (1)

Un'idea per risolvere questi problemi viene dalle folksonomie (*tassonomie generate dal popolo: folk).* A volte anche: *tag cloud.*

Attraverso le folksonomie,

- Gli utenti finali stessi generano (molteplici) termini descrittivi delle risorse
- Non c'è vocabolario controllato, non c'è modello concettuale
- Ogni risorsa viene associata ad una categoria totalmente identificata dal termine usato, in proporzione al numero di utenti che usano quel termine per descriverla
- La prevalenza statistica di alcuni termini su altri rende la risorsa più identificata da quel termine che da altri.
- Non è possibile fare inferenze o deduzioni sui termini (sono stringhe opache e non ulteriormente analizzabili).

Le folksonomie (2)

Rispetto alle critiche sui modelli visti in precedenza:

- Il personale qualificato viene sostituito dalla massa degli utenti finali
- La competenza di dominio è automaticamente presente,
- Della competenza di classificazione si fa a meno.
- Non c'è modello concettuale, il vocabolario è totalmente libero.
- Gli utenti sono automaticamente d'accordo sulla concettualizzazione
- Il modello (implicit) è sempre incompleto e sempre estendibile (anzi esteso).
- Non c'è progettazione completa e dettagliata di alcun modello concettuale, né prima né dopo.

Complessivamente, sono un approccio gratuito, flessibile, democratico, decentralizzato e olistico. Inoltre scala benissimo su dimensioni veramente grandi.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Vocabolari importanti nel SW

Vocabolari nel SW

Un vocabolario semantico è un insieme di termini che, autonomamente o in collaborazione con altri vocabolari complementari, può essere usati per descrivere in maniera sufficientemente ricca e completa un dominio concettuale.

I vocabolari sono dunque specifici di un dominio e debbono poter essere integrandoli usandoli insieme ad altri vocabolari di un dominio contiguo.

I vocabolari del semantic web tipicamente descrivono:

- Entità (singoli individui meritevoli di menzione)
- Classi (raggruppamenti di entità dotati di caratteristiche comuni)
- Proprietà:
  - proprietà semplici (ad esempio <doc X> <ha titolo> "Promessi sposi"
  - relazioni con altre entità (ad esempio: <doc X> <è stato scritto da> <Persona Y>
- Eventualmente ulteriori vincoli di numero, classe e obbligatorietà.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Serializzazioni del SW

Lo stesso modello concettuale, composto di semplici affermazioni che utilizzano termini del/dei vocabolario/i scelti, può essere rappresentato in molti modi diversi, secondo molte sintassi diverse (*serializzazioni di RDF*).

Alcuni esempi:

- ***RDF-XML***: il primo vocabolario RF, in XML, un po' verboso e ormai poco usato:

```
<rdf:Description rdf:about="https://www.dbpedia.org/MonaLisa">
  <dc:creator rdf:resource="https://www.dbpedia.org/DaVinci" />
</rdf:Description>
```

- ***Turtle (Trig)***: una serializzazione molto minimale, composta da brevi frasi RDF:

```
:MonaLisa dc:creator :LeoDaVinci.
```

- ***JSON-LD***: un'estensione di JSON meno verbosa di RDF-XML basata su JSON:

```
{
  "@context": {
    "dc": "http://purl.org/dc/terms/"
  },
  "@id": "https://www.dbpedia.org/MonaLisa",
  "dc:creator": "https://www.dbpedia.org/DaVinci"
}
```

- ***RDFa***: un'estensione di HTML5 che mescola testo e RDF, basato su nuovi attributi:

```
<p about="https://www.dbpedia.org/MonaLisa">Il quadro La Gioconda <span
property="http://purl.org/dc/terms/creator">fu dipinto da <span
href="https://www.dbpedia.org/DaVinci">Leonardo</span></span>.</p>
```

Alcuni vocabolari famosi

Organizzazioni classiche (non digitali)

- Classificazione Dewey
- Classificazione Library of Congress
- Modello PMEST (Ranghanatan)
- Marc 21

Ontologie

- Dublin Core
- FRBR
- FOAF
- SKOS

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Classificazione Dewey

Usata in moltissime biblioteche, inventata da Melvil Dewey nel 1876, è organizzata in una struttura a lunghezza variabile di blocchi numerici gerarchici dove ogni cifra rappresenta un concetto. 10 numeri indicano al massimo dieci insiemi.

Un'opera allora è classificata secondo una sequenza di numeri, ciascuno dei quali identifica una specifica categoria del pensiero.

Ad esempio:

- **345.914 - diritto penale europeo** (3: scienze sociali, 34: diritto, 345: diritto penale - 9: storia, geografia e biografia, 91: geografia, 914: europa
- **6368 - Allevamento dei gatti** (6: tecnologia, 63: agricoltura, 636: allevamento animale, 6368: allevamento dei gatti).

Quindi tutta la conoscenza è organizzata in 10 macro classi, ciascuna delle quali è organizzata in 10 sottoclassi, che a loro volta sono organizzate in 10 sottoclassi e via così.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Problemi di Dewey (1)

Dewey forza tutta la conoscenza in classi di 10 elementi ciascuna.

Questa operazione è fatta a priori, non può essere riorganizzata, e la struttura dipende fortemente dall'autore che l'ha organizzata.

Computer science non esisteva nell'ottocento. Hanno dovuto riciclare una categoria poco usata, la 0, per indicare computer science.

Che succede se adesso emerge una nuova disciplina mai vista prima?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Problemi con Dewey (2)

- 200: Religione (generale)
- 210: Teologia naturale
- 220: Bibbia
- 230: Teologia cristiana
- 240: Morale cristiana
- 250: Organizzazione cristiana e ordini religiosi
- 260: Teologia sociale cristiana
- 270: Storia della chiesa cristiana
- 280: Sette cristiane e denominazioni protestanti
- 290: Altre religioni
  - 291: Religioni comparative
  - 292: Religioni classiche
  - 293: Religioni germaniche
  - 294: Religioni di origine indiana
  - 295: Zoroastrismo
  - 296: Ebraismo
  - 297: Islam, Baba e Bahai
  - 298: *Non assegnato*
  - 299: Altre religioni

Dewey era bianco, cristiano, americano, dell'800. Forse questo spiega?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PMEST

S.R. Ranganathan, l'inventore della classificazione a faccette, aveva anche identificato le 5 faccette primarie (PMEST) per descrivere gli oggetti classificabili:

- Personality: ciò di cui si parla
- Matter: il materiale fisico con cui si agisce sulla personality
- Energy: l'azione che si compie
- Space: il luogo fisico in cui si compie l'azione
- Time: il tempo in cui si compie l'azione.

Ogni faccetta può essere usata più volte, con valori in gerarchia generica (*is_a*). Ad esempio, i libri sulla ricerca sulla cura della tubercolosi ai polmoni coi raggi X nell'India degli anni cinquanta verrebbero resi come:

- Personality: Medicina/Polmoni
- Matter: Tubercolosi
- Energy: Cura
- Space: India
- Time: Anni 50

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Marc 21

- **Ma**chine **R**eadable **C**ataloging, inventato negli anni '60 per la catalogazione informatica delle risorse librarie. Una stringa lunga 24 byte ciascuno dei quali contiene codici di catalogazione, descrizione e controllo. La versione attualmente usata è la 21esima (Marc 21).
- C'è una corrispondenza (numerica) tra le etichette della scheda (autore, titolo, anno di pubblicazione, …) e il campo corrispondente al valore dell'etichetta.
- C'è anche una suddivisione in sottocampi, per aggiungere informazioni sull'entità dell'etichetta rappresentata. Ad esempio, l'etichetta 100 corrispondente all'autore, 100$a, 100$b, …, ciascuna corrispondente a una qualche informazione specifica dell'autore.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dublin Core

Dublin Core

Modello di metainformazioni ideato per assegnare etichette ragionevoli alle risorse della rete.

E' sostanzialmente una classificazione a faccette considerata come un'ontologia attraverso la creazione (artificiosa) di una classe documento a cui tutte le proprietà della classificazione fanno riferimento. Dublin Core è indipendente da qualunque sintassi, ma si usa bene con RDF (esiste anche una sintassi per HTML).

Si chiama Dublin Core perché è considerato il nucleo (core) delle meta-informazioni interessanti per qualunque risorsa, e perché è nato da un'iniziativa di bibliotecari, archivisti, fornitori di contenuto e esperti di markup svoltasi nel 1995 a Dublino (Ohio, non Irlanda).

Dublin Core versione 1 ha introdotto esattamente quindici categorie di meta-informazioni utili per la catalogazione di risorse di rete.

La versione 2 ha aggiunto un meccanismo di sottoclassi (detti qualificatori) delle categorie, ed ha introdotto un elenco iniziale di qualificatori.

DC – Categorie e Classificazione

Le quindici categorie descrivono metainformazioni di tre tipi:

Contenuto	***Proprietà intellettuale***	***Istanza***
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

I qualificatori permettono di specificare ulteriormente informazioni di queste categorie, secondo questi criteri:

- ***Raffinamento dello schema***: fornisce alcuni significati più precisi sui termini. Ad esempio, “Date” ha come qualificatori: “created”, “valid”, “available”, “issued”, “modified”).
- ***Supporto per codifiche specifiche***: permette di usare i valori di particolari codifiche all'interno del Dublin Core. Ad esempio, “Subject” ha come qualificatori: “LCSH” (Library of Congress Subject Headings), “MeSH” (Medical Subject Headings), “DDC” (Dewey Decimal Classification), ecc.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dublin Core – Entità (1/4)

Titolo (Title) - Nome dato alla risorsa. In particolare, un Titolo sarà un termine con il quale la risorsa è formalmente conosciuta.

Autore (Creator) - Entità che ha la responsabilità principale della produzione del contenuto della risorsa. Esempi di Autore possono essere una persona, un'organizzazione o un servizio responsabili del contenuto intellettuale della risorsa.

Soggetto (Subject) – Topic principale della risorsa. In particolare un Soggetto può essere espresso da parole o frasi chiave, o da codici di classificazione che descrivono l'argomento della risorsa. Solitamente questi termini vengono scelti tra i valori di un vocabolario controllato o di uno schema di classificazione formale.

Descrizione (Description) – Spiegazione del contenuto della risorsa. Testo descrittivo libero che può includere un riassunto analitico, un indice, o una rappresentazione grafica del contenuto.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dublin Core – Entità (2/4)

Editore (Publisher) - Entità responsabile della pubblicazione della risorsa. Esempi di Editore possono essere una persona, un'organizzazione o un servizio che si occupa di rendere disponibile la risorsa nella sua forma attuale.

Autore di contributo subordinato (Contributor) - Entità responsabile della produzione di un contributo al contenuto della risorsa. Esempi di Autore secondario includono una persona, un'organizzazione o un servizio che contribuiscono alla produzione della risorsa.

Data (Date) - Data associata ad un evento del ciclo di vita della risorsa. Normalmente la data è associata al momento di creazione o di disponibilità della risorsa e viene indicata attraverso una stringa di 8 caratteri nella forma YYYY-MM-DD, come definita nel profilo dello standard ISO 860190. In questo schema l'elemento data 1994-11-05 corrisponde al 5 novembre 1994.

Tipo (Type) - Natura o genere del contenuto della risorsa. L'elemento "Tipo" include termini che descrivono categorie generali, funzioni, generi, o livelli di aggregazione per contenuto presi generalmente da un vocabolario controllato.

Dublin Core – Entità (3/4)

Formato (Format) - Manifestazione fisica o digitale della risorsa. Normalmente l'elemento "Formato" può includere il tipo di supporto o le dimensioni, ossia grandezza e durata, della risorsa. Format può essere usato per determinare il software o l'hardware necessari alla visualizzazione o all'elaborazione della risorsa.

Identificatore (Identifier) - Riferimento univoco alla risorsa nell'ambito di un dato contesto. Solitamente le risorse vengono identificate per mezzo di una sequenza di caratteri alfabetici o numerici secondo un sistema di identificazione formalmente definito. Esempi di tali sistemi di identificazione includono l'Uniform Resource Identifier (URI) (incluso l'Uniform Resource Locator (URL)), il Digital Object Identifier (DOI) e l'International Standard Book Number (ISBN).

Relazione (Relation) - Riferimento ad una risorsa correlata.

Fonte (Source) - Riferimento a una risorsa dalla quale è derivata la risorsa in oggetto. La risorsa in questione potrebbe derivare, in tutto o in parte, da un'altra risorsa fonte.

Dublin Core – Entità (4/4)

Lingua (Language) – Lingua del contenuto intellettuale della risorsa. Per i valori dell'elemento Lingua si utilizza un codice di linguaggio, seguito opzionalmente da un codice di paese, entrambi su due caratteri. Ad esempio "it" per l'italiano o "en-uk" per l'inglese usato nel Regno Unito.

Copertura (Coverage) - Estensione o scopo del contenuto della risorsa. Normalmente Copertura include la localizzazione spaziale (il nome o le coordinate geografiche di un luogo), il periodo temporale (l'indicazione di un periodo, una data o una serie di date) o una giurisdizione (ad esempio il nome di un'entità amministrativa).

Gestione dei diritti (Rights Management) - Informazione sui diritti esercitati sulla risorsa. Normalmente un elemento "Diritti" contiene un'indicazione sulla gestione dei diritti sulla risorsa, o un riferimento al servizio che fornisce questa informazione. Questo campo comprende gli Intellectual Property Rights (IPR), il copyright, e vari diritti di proprietà. Se l'elemento Rights è assente, non si può fare alcuna ipotesi sui diritti della risorsa.

DC – Esempio in JSON-LD

Questo è un esempio di classificazione di una pubblicazione in Dublin Core:

```
{
  "@context": {
    "dc": "http://purl.org/metadata/dublin_core#"
  },
  "@id": "http://www.dlib.org",
  "dc:Date": "2023-05-05",
  "dc:Description": "Bla Bla",
  "dc:Format": "text/html",
  "dc:Language": "en",
  "dc:Publisher": "Corp. For National Research Initiatives",
  "dc:Subject": [
     "Research; statistical methods",
     "Education, research, related topics",
     "Library use Studies"
  ]
}
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DC – Esempio in Turtle

Questo è un esempio di classificazione di una pubblicazione in Dublin Core:

```
@prefix dc:    <http://purl.org/metadata/dublin_core#> .

<http://www.dlib.org>
  dc:Date          "2023-05-05" ;
  dc:Description   "Bla Bla" ;
  dc:Format        "text/html" ;
  dc:Language      "en" ;
  dc:Publisher     "Corp. For National Research Initiatives";
  dc:Subject       "Library use Studies" ,
                   "Education, research, related topics" ,
                   "Research; statistical methods" .
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DC – Esempio in XML-RDF

Questo è un esempio di classificazione di una pubblicazione in Dublin Core:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/metadata/dublin_core#">
  <rdf:Description rdf:about="http://www.dlib.org">
    <dc:Title>D-Lib Program</dc:Title>
    <dc:Description>Bla Bla</dc:Description>
    <dc:Publisher>
       Corp. For National Research Initiatives
    </dc:Publisher>
    <dc:Date>2023-05-05</dc:Date>
    <dc:Subject><rdf:Bag>
      <rdf:_1>Research; statistical methods</rdf:_1>
      <rdf:_2>Education, research, related topics</rdf:_2>
      <rdf:_3>Library use Studies</rdf:_3>
    </rdf:Bag></dc:Subject>
    <dc:Type>World Wide Web Home Page</dc:Type>
    <dc:Format>text/html</dc:Format>
    <dc:Language>en</dc:Language>
  </rdf:Description>
</rdf:RDF>
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

FRBR

Verso FRBR: il problema

- Quando due documenti sono lo stesso documento?
  - Quando sono copie fisiche diverse dello stesso documento (due libri identici)
  - Quando sono modi diversi in cui vengono presentate le stesse parole (un file PDF e il libro stampato)
  - Quando sono insiemi di parole diverse che hanno lo stesso nome e lo stesso scopo (due versioni diverse della stessa legge)

- Quando due documenti sono lo stesso documento?
  - La gerarchia FRBR mi permette di dare una buona organizzazione concettuale al termine complesso "documento"
    - Work, Expression, Manifestation, Item

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

FRBR

Functional Requirements for Bibliographic Records

Sviluppato dall'International Federation of Library Associations negli anni 1990, associazione nata con lo scopo di definire concetti di catalogazione.

Perché questa esigenza? La pressione sui costi di catalogazione dà il via all'utilizzo di modelli FRBR in vista di un riesame dei concetti classici di catalogazione.

Scopo: Elaborare un modello concettuale che permetta di identificare i requisiti minimi delle descrizioni che interessano l'utente consultatore.

Questo modello definisce teoricamente le finalità dell'archiviazione (allocazione e struttura), in relazione alle tipologie dei media e alle molteplici necessità degli utenti. Ogni utente deve poter essere in grado di **reperire** informazioni in base alla propria lingua ed **esigenze**, **identificare** documenti, **riutilizzare** i dati ottenuti.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

La gerarchia IFLA FRBR (1)

Work: una creazione intellettuale distinta.

Expression: la forma specifica in cui un'opera viene realizzato

- Concretamente, tutte le varianti e le versioni di un testo che incorporano modifiche, correzioni, forme varianti da una versione precedente
- Tutte le forme derivate (traduzioni, rappresentazioni, film) che esprimono in una forma contenutisticamente diversa lo stesso testo

Manifestation: la rappresentazione di un espressione sulla base dei requisiti di un medium

- Concretamente, la forma rappresentata di un'espressione, come espressa da un formato dati o una rappresentazione fisica (ad esempio in uno stampato o in un libro)
- PDF vs. XML vs. Word - forma teatrale vs. MPG vs. AVI

Item: un esemplare individuale di una manifestazione

Ogni copia di un libro; ogni file identico byte per byte.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

La gerarchia IFLA FRBR (2)

Work:

- La tragedia "Amleto" di William Shakespeare
- La legge italiana n. 3 (5 Gennaio 2008)

Caratteristica tipica: l'***identità***

Expression:

- Il primo quarto di "Amleto" (1601);
- Il primo *folio* of "Amleto" (1623);
- La versione in film di "Amleto" di K. Brannagh (1996)
- La versione originale della legge italiana n. 3/2008;
- La versione modificata della legge italiana n. 3/2008 come risulta al 30/3/2010

Caratteristica tipica: il ***contenuto***

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

La gerarchia IFLA FRBR (3)

Manifestation:

- Una versione stampata del primo folio di "Amleto" (e.g.: Penguin Books, 1994)
- Una versione su computer di "Amleto" (e.g., Progetto Gutenberg)
- La versione in XML NIR della legge italiana n. 3/2008 alla data del 31/3/2010
- La versione stampata dal PDF della legge italiana n. 3/2008 alla data del 31/3/2010

Caratteristica tipica: la ***forma*** (o il ***flusso di byte***)

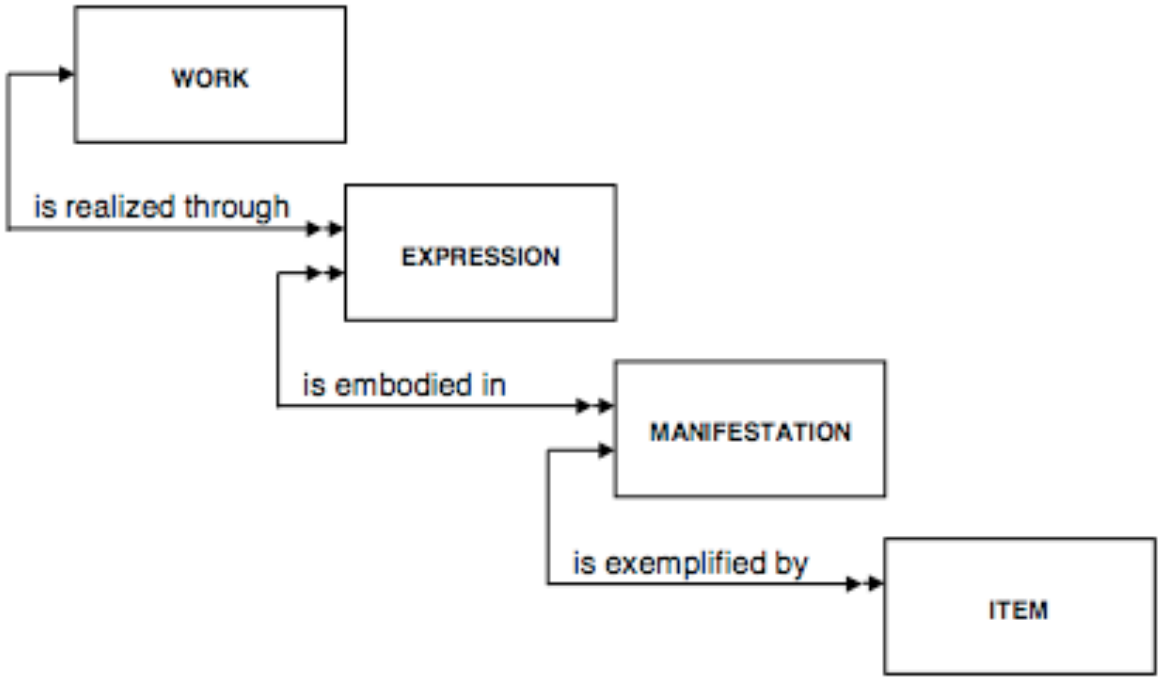
Item:

- La mia copia dell'Amleto nell'edizione dei Penguin Books;
- La copia dell'Amleto sul sito del Progetto Gutemberg
- La copia della versione in XML NIR della legge italiana n. 3/2008 alla data del 31/3/2010 che sta sul sito NIR
- Lo stesso file sul mio computer.

Caratteristica tipica: la ***locazione***

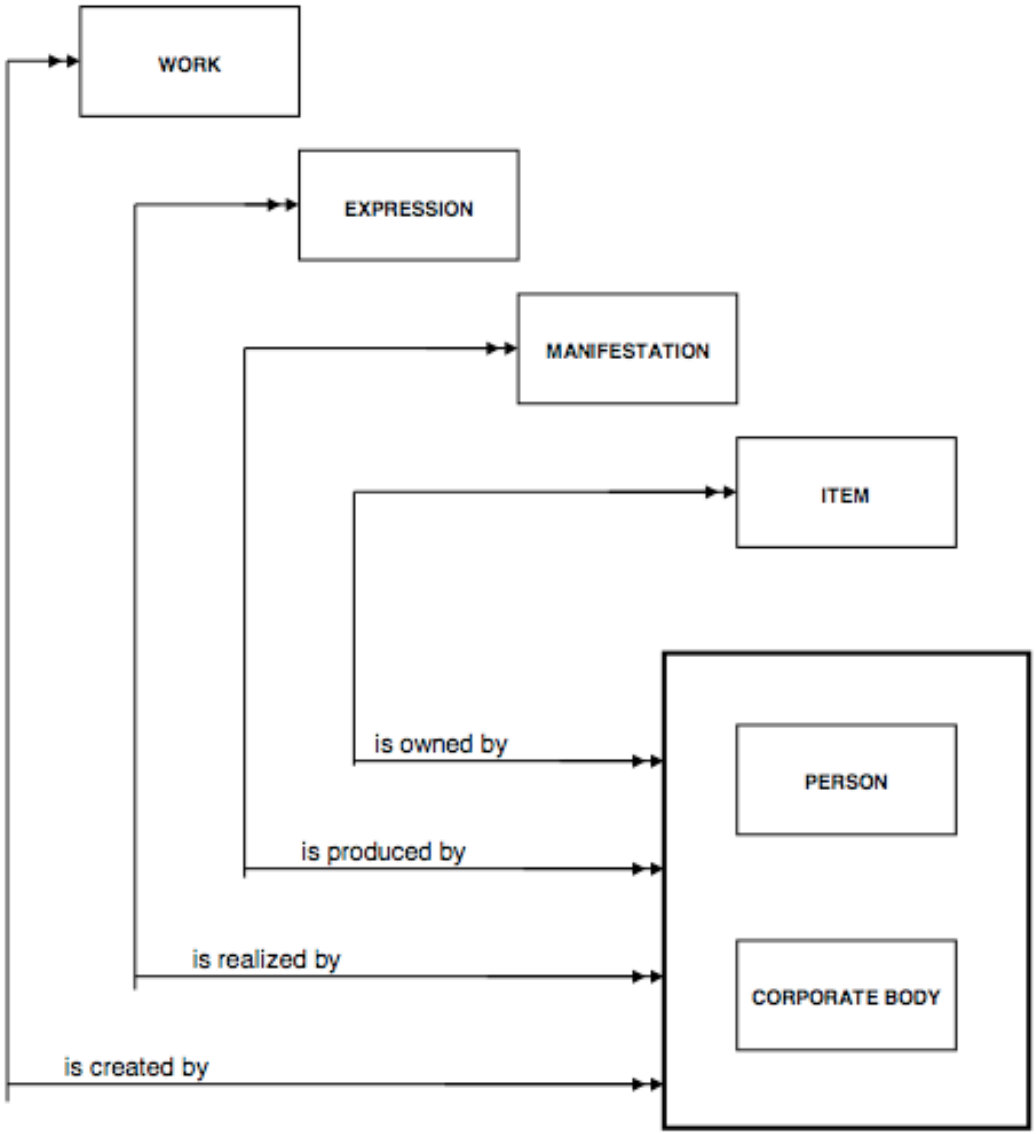
FRBR: Relazioni primarie 1/2

Il seguente schema rappresenta le relazioni primarie delle entità (classi) FRBR



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

FRBR: Relazioni primarie 2/2



- Schema generale delle assegnazioni in FRBR.
- Nei rettangoli sono rappresentate le classi e le proprietà sono le frecce.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dublin Core e FRBR

Molte entità del modello Dublin Core sono esplicitamente associabili agli attributi FRBR.

- Dc:Creator -> frbr:Creator (di manifestazione)
- Dc:Title -> frbr:Title (di manifestazione)
- … e molti altri

Il modello FRBR è molto più esteso. Analizza la storia della risorsa, il periodo di vita, e la diffusione di un opera in generale.

Dublin Core mira invece a descrivere un documento digitale (quello che in FRBR è visto come “manifestation”) attraverso uno standard diffuso in modo tale che i documenti possano essere confrontati tra loro attraverso una struttura comune.

Possibilità di fondere e integrare i due modelli.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

FOAF

Friend Of A Friend (FOAF)

Il progetto Friend Of A Friend è un iniziativa a base volontaria ma sposata dal W3C per definire una **descrizione machine-readable** di persone, gruppi, aziende e altri tipi di concetti connessi.

Il vocabolario FOAF definisce in termini RDF/OWL alcune classi e loro proprietà.

Classi: agent (sottoclassi: persona, gruppo, organizzazione), document (sottoclasse: image, PersonalProfile), OnlineAccount (con varie sottoclassi), project.

```
{
    "@context" : {
      "foaf" : "http://xmlns.com/foaf/0.1/"
    },
    "@type" : "foaf:Person",
    "foaf:name" : "Fabio Vitali",
    "foaf:homepage" : "http://www.cs.unibo.it/~fabio/",
    "foaf:img" : "http://www.cs.unibo.it/~fabio/fabio.jpg"
}
```

Classi FOAF

foaf:agent è la classe che racchiude l'entità persona, organizzazione e gruppo. Le caratteristiche di agent vengono ereditate da foaf:person e dalle altre sottoclassi, che rappresentano le classi principali del modello.

foaf:document rappresenta anch'essa una classe fondamentale, in quanto è la classe che rappresenta tutti i documenti (es. testi, immagini, documenti cartacei..etc). La classe document non distingue un documento cartaceo da uno digitale, o una copia da un originale. Da la possibilità di distinguere le diverse instance di un documento attraverso la proprietà foaf:sha1.

foaf:OnLineAccount è la classe che descrive qualsiasi tipo di servizio online. Dalla pagina web, all'account personale, al filmato online etc.

foaf:project, è la classe progetto. Può assumere un aspetto formale o informale, individuale o collettivo.

Le classi generiche danno definizione generiche dell'oggetto che si sta descrivendo. L'aggiunta di proprietà a queste classi (dove possibile) ne delinea la natura.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

SKOS

Simple Knowledge Organisation System (SKOS)

SKOS è un modello ***per esprimere le strutture di base e i contenuti*** di schemi concettuali come tesauri, schemi di classificazione, tassonomia, terminologie, glossari e altri tipi di vocabolari controllati.

Uno schema concettuale è definito come un insieme di concetti eventualmente contenenti affermazioni sulle relazioni semantiche tra questi concetti. Ciascuno di questi concetti è descritto con un insieme di affermazioni RDF espresse secondo uno schema di classi e proprietà RDF-S.

SKOS definisce come classe principalmente *concept*, che ha come proprietà importanti le *label* (come vengono indicati i termini) e le relazioni semantiche *broader*, *narrower*, *related*. La relazione di sinonimia viene svolta da una label "nascosta".

SKOS usa Dublin Core per riferirsi ai documenti, e FOAF per riferirsi ad agenti (autori, organizzazioni, individui, ecc.).

Esempio: due concetti (JSON-LD)

```
{
  "@context" : {
    "s" : "http://www.w3.org/2004/02/skos/core#",
    "c" : "http://www.site.com/myconcepts/"
  },
  "@graph" : [ {
      "@id" : "c:animals",
      "@type" : "s:Concept",
      "s:narrower" : "c:mammals",
      "s:prefLabel" : "animals"
    }, {
      "@id" : "c:mammals",
      "@type" : "s:Concept",
      "s:broader" : "c:animals",
      "s:prefLabel" : "mammals"
    }
  ]
}
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Esempio: due concetti (Turtle)

```
@prefix s:     <http://www.w3.org/2004/02/skos/core#> .
@prefix c:     <http://www.site.com/myconcepts/> .

c:mammals  a          s:Concept ;
        s:broader     c:animals ;
        s:prefLabel   "mammals" .

c:animals  a          s:Concept ;
        s:narrower    c:mammals ;
        s:prefLabel   "animals" .
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Esempio: due concetti (RDF-XML)

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:c="http://www.site.com/myconcepts/"
     xmlns:s="http://www.w3.org/2004/02/skos/core#" >

  <rdf:Description rdf:about="c:mammals">
    <rdf:type rdf:resource="s:Concept"/>
    <s:prefLabel>mammals</s:prefLabel>
    <s:broader rdf:resource="c:animals" />
  </rdf:Description>

  <rdf:Description rdf:about="c:animals">
    <rdf:type rdf:resource="s:Concept"/>
    <s:prefLabel>animals</s:prefLabel>
    <s:narrower rdf:resource="c:mammals" />
  </rdf:Description>

</rdf:RDF>
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Conclusioni

Oggi abbiamo parlato di

- La necessità del Semantic Web
- PICS
- L'organizzazione delle informazioni
- L'indicizzazione di soggetti
- Digitalizzazione di archivi
- Modelli ontologici per la rappresentazione documentale
- Modelli per la rappresentazione concettuale

I modelli appena osservati possono essere utilizzati e combinati tra loro purché vengano rispettate le regole di mapping degli modelli stessi.

Riferimenti

- International Standard ISO-2788, Documentation -- *Guidelines for the development of monolingual thesauri*, Second edition -- 1986-11-15
- Chris Taylor, *An Introduction to Metadata*, http://www.library.uq.edu.au/iad/ctmeta4.html
- Serafina Spinelli, *Introduzione all'indicizzazione,* http://mail.biocfarm.unibo.it/~spinelli/indicizzazione/
- Serafina Spinelli, *Introduzione ai thesauri*, http://mail.biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm

Bibliografia

- MARC21, http://www.loc.gov/marc/bibliographic/ecbdhome.html/
- PREMIS Data Dictionary for Preservation Metadata, http://www.loc.gov/standards/premis/
- Functional requirements for bibliographic records : final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. - München : Saur, 1998
- DMCI, Dublin Core Metadata Initiative, 05/11/2007, http://www.dublincore.org/
- FOAF – The Friend of a Friend project, http://www.foaf-project.org/
- SKOS - Simple Knowledge Organisation System, http://www.w3.org/2004/02/skos/