semantic data

alessandro bollini
20.11.2009

introduzione

obiettivi

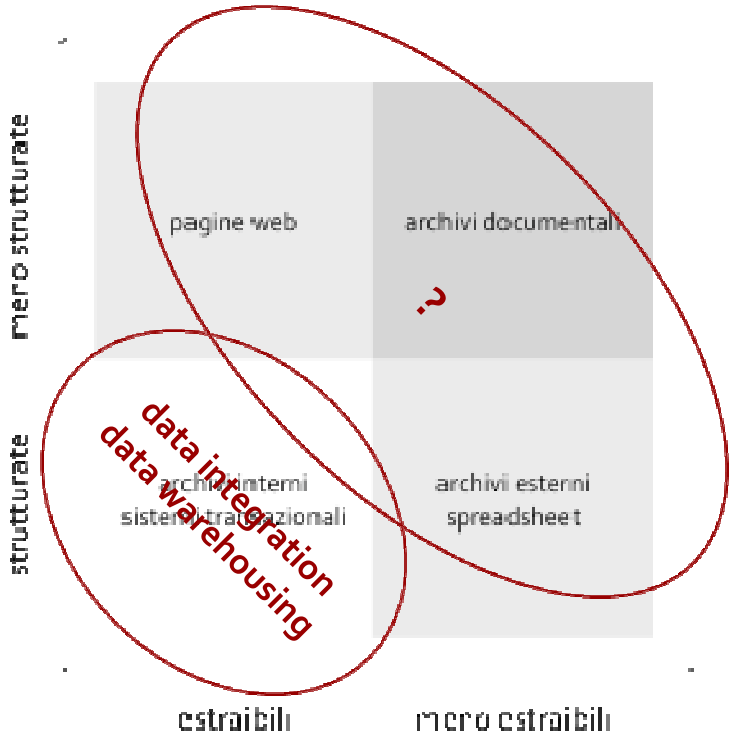
- introduzione alle tecnologie semantiche
- applicate al trattamento di dati distribuiti e parzialmente strutturati

struttura

- contesto
- strategia
- tecnologie
- casi di studio

informazione e conoscenza

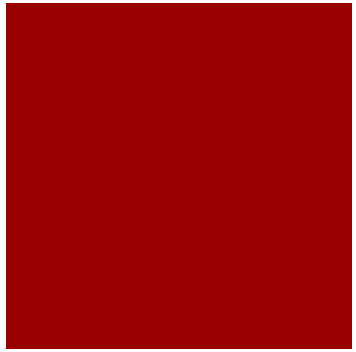
- la rete ed i sistemi informativi istituzionali contengono un'enorme quantità di informazione
  - che potrebbe essere trasformata in conoscenza utile ad attività esplorative o decisionali
- la quantità di informazione disponibile richiede un trattamento automatizzato
  - strutturazione
    - schema regolare
    - popolamento
  - estraibilità
    - protocollo di trasferimento
    - formato convertibile



tecnologie semantiche

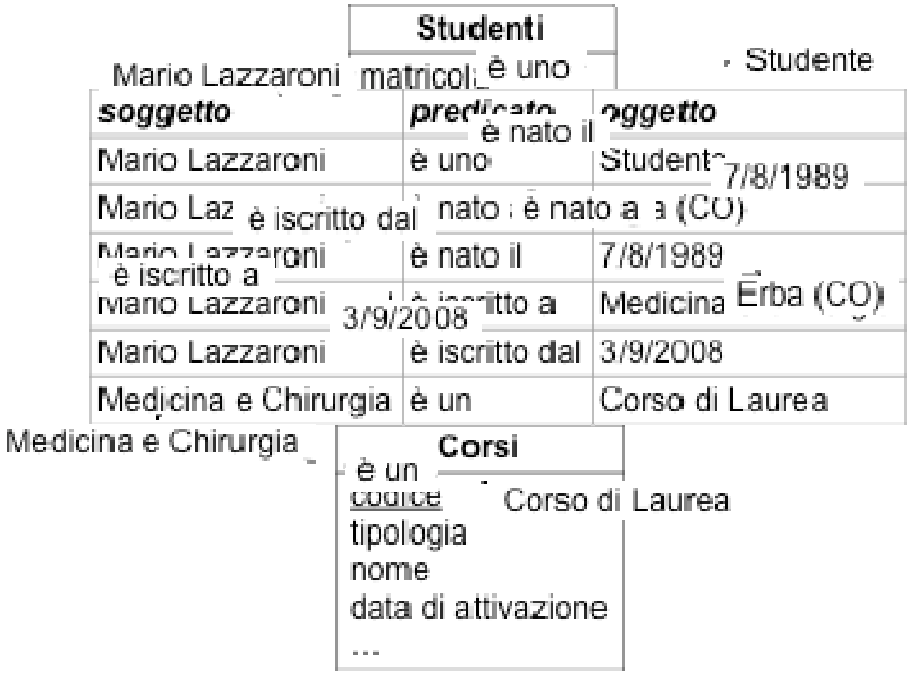
- una strategia per
  - la **decentralizzazione** delle attività di pubblicazione di informazioni anche solo parzialmente strutturate
  - l'automazione delle attività di **integrazione** di informazioni provenienti da fonti distribuite ed eterogenee
- tramite standard aperti per
  - la **rappresentazione** dell'informazione
  - la descrizione della struttura **semantica**
  - l'**accesso** alle fonti di informazione distribuite

rappresentazione dell'informazione



rappresentazione dell'informazione
modello reticolare

- Mario Lazzaroni, nato a Erba (CO), il 7/8/1989, è iscritto al Corso di Laurea in Medicina e Chirurgia dal 3/9/2008.



soggetto	*predicato*	*oggetto*
Mario Lazzaroni	è uno	Studente
Mario Lazzaroni	è nato il	7/8/1989
Mario Lazzaroni	è nato a	Erba (CO)
Mario Lazzaroni	è iscritto a	Medicina
Mario Lazzaroni	è iscritto dal	3/9/2008
Medicina e Chirurgia	è un	Corso di Laurea

rappresentazione dell'informazione estensione della struttura

- Mario Lazzaroni, nato a Erba (CO), il 7/8/1989, è iscritto alla Scuola di Specialità in Cardiologia dal 7/9/2015 e presta servizio presso l'Unità Operativa Coronarica dalla stessa data.

rappresentazione dell'informazione identificazione degli elementi

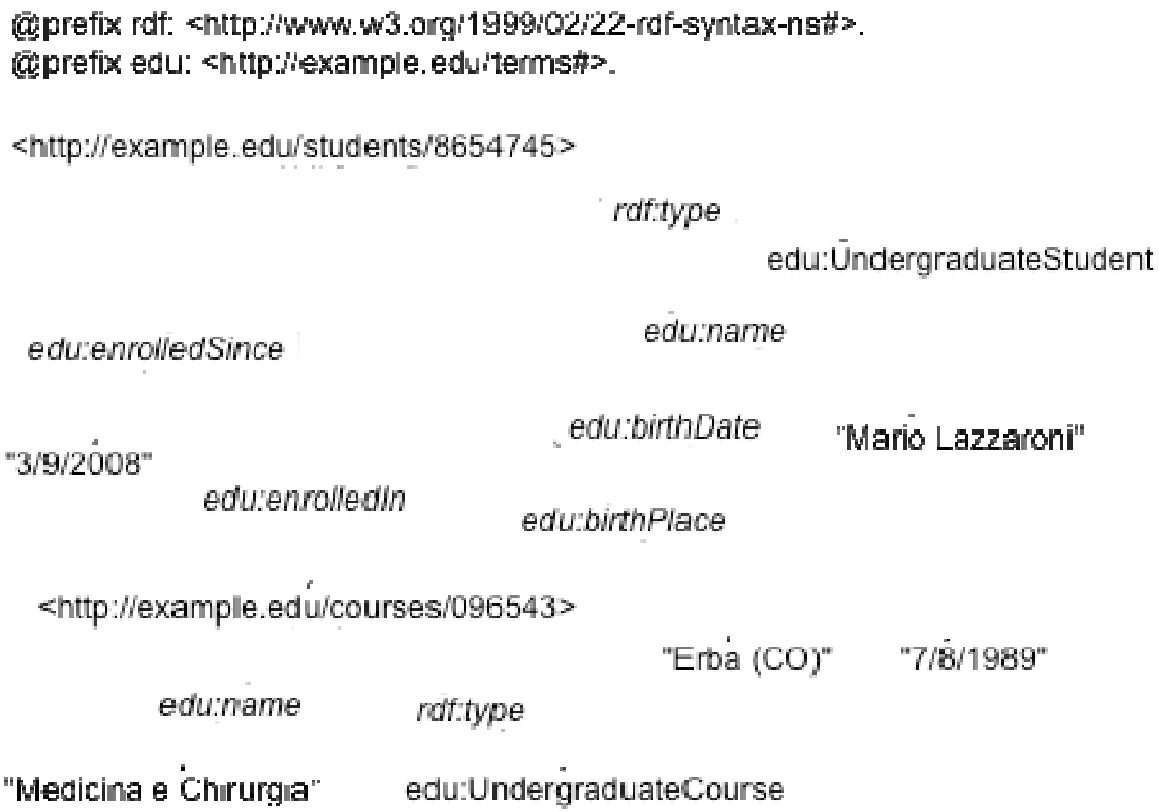
- per evitare ambiguità gli elementi della struttura devono essere identificati univocamente
- la struttura contiene elementi di diversa natura
  - entità (tipi e istanze)
  - valori letterali
  - relazioni tra entità e valori
- ad entità e relazioni viene assegnato un identificatore
  - secondo lo schema URI
  - ampiamente diffuso e globale

soggetto	*predicato*	*oggetto*
Mario Lazzaroni	è uno	Studente
Mario Lazzaroni	è nato a	Erba (CO)
Mario Lazzaroni	è nato il	7/8/1989
Mario Lazzaroni	è iscritto a	Medicina e Chirurgia
Mario Lazzaroni	è iscritto dal	3/9/2008
Medicina e Chirurgia	è un	Corso di Laurea

soggetto	*predicato*	*oggetto*
<URI>	<URI>	<URI>
<URI>	<URI>	"valore"

rappresentazione dell'informazione resource description framework (RDF)

- rappresentazione reticolare
  - nodi
    - entità
    - valori letterali
  - archi orientati
    - relazioni
      - tra soggetti (origine)
        - entità
      - e oggetti (destinazione)
        - entità
        - valori letterali
  - etichette
    - identificatore entità (URI)
    - tipo di relazione (URI)
    - rappresentazione testuale

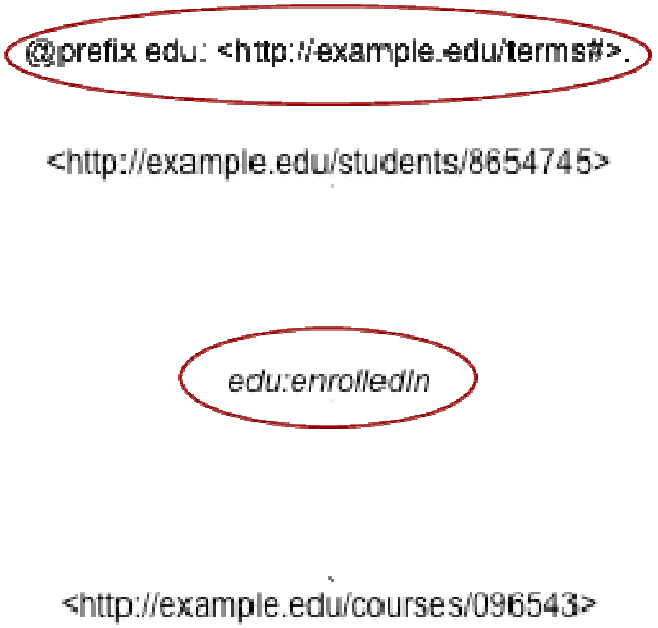


rappresentazione dell'informazione uniformità e flessibilità

- tutte le informazioni possono essere rappresentate tramite un'unica struttura reticolare
  - non è necessario intervenire sullo schema per trattare nuove fonti
- entità e relazioni sono identificate univocamente e globalmente
  - dati provenienti da fonti diverse possono coesistere senza ambiguità
- non viene imposto uno schema predefinito ai dati
  - dati strutturati e semi-strutturati vengono trattati uniformemente
  - i soggetti possono essere descritti nello stesso contesto secondo schemi diversi e potenzialmente conflittuali

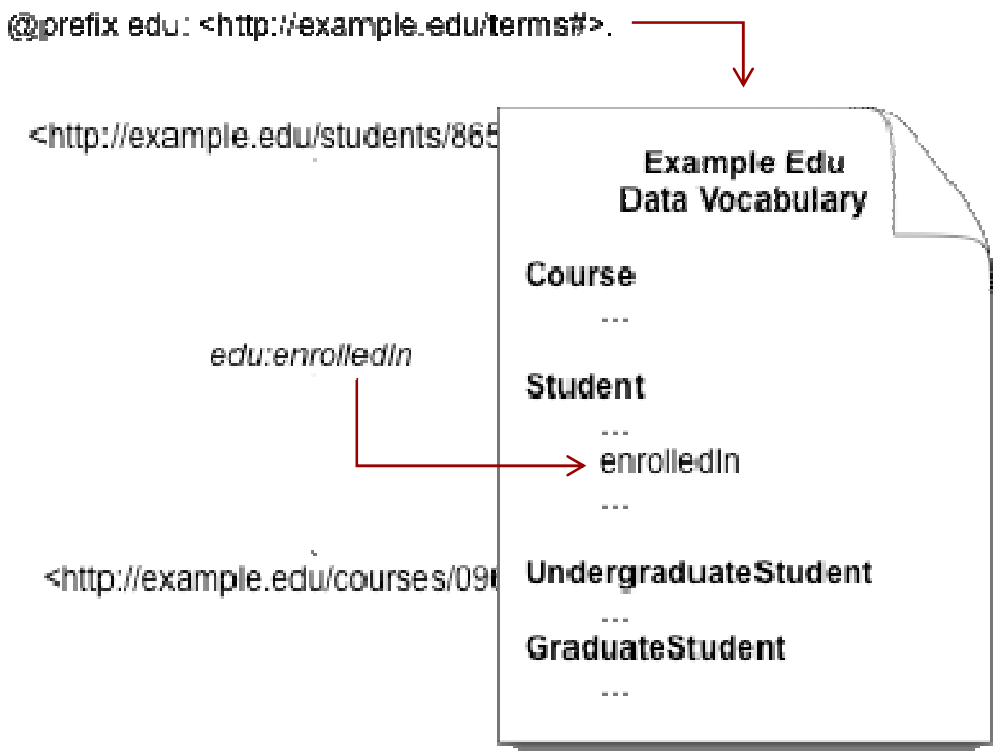
rappresentazione dell'informazione
descrizione della semantica

- il tipo delle entità e delle relazioni può essere identificato tramite termini descritti in uno schema
- i termini dello schema devono essere utilizzati sistematicamente secondo un significato convenzionale
  - come viene descritto il significato?
  - come possono utenti ed applicazioni accedere alla descrizione?



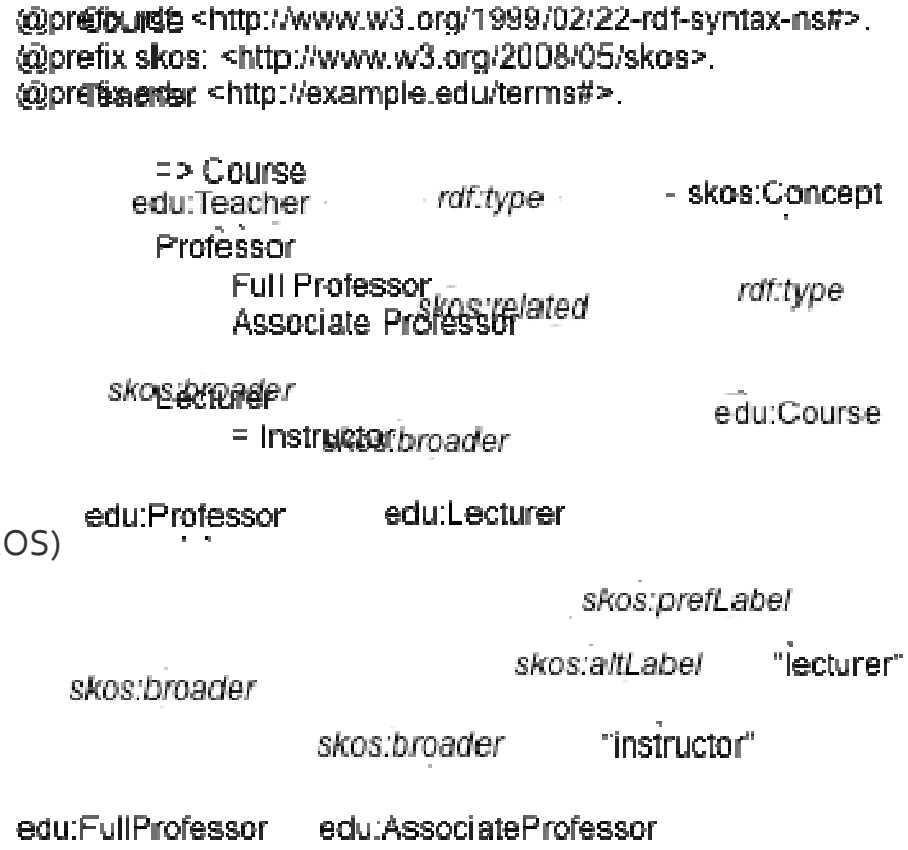
rappresentazione dell'informazione
vocabolari

- vocabolario dei termini utilizzati per identificare entità e tipi di relazione
  - descrizioni testuali
  - pubblicate online
- termine -> descrizione
  - accesso tramite URI dereferenziabili (URL)
- gli utenti possono leggere ed utilizzare le definizioni (esplorazione, verifica, ...)



rappresentazione dell'informazione tassonomie

- formalizza ed estende il concetto del vocabolario online
  - relazioni gerarchiche
  - relazioni associative
  - etichette e annotazioni
- ma utilizza il modello reticolare per descrivere termini e relazioni
  - Simple Knowledge Organization System (SKOS)
- applicazioni intelligenti possono leggere ed utilizzare la tassonomia (ricerche estese, visualizzazione, ...)



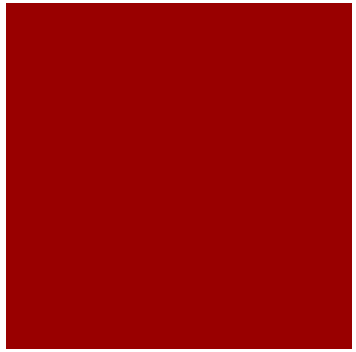
rappresentazione dell'informazione ontologie

- estende il concetto di tassonomia
  - classi e relazioni in termini di logica descrittiva
  - regole di inferenza
- utilizza il modello reticolare per descrivere termini e relazioni
  - Web Ontology Language (OWL)
  - Semantic Web Rule Language (SWRL)
  - Rule Interchange Format (RIF)
- applicazioni intelligenti possono leggere ed utilizzare l'ontologia (inferenza, validazione, …)

rappresentazione dell'informazione uniformità e flessibilità (II)

- l'informazione rappresentata in forma reticolare può essere associata alla descrizione semantica dei termini utilizzati nella descrizione
  - l'assenza di uno schema predefinito non ne preclude la possibilità
- la semantica dei termini può essere descritta a diversi livelli di complessità e precisione
  - lo schema può essere sviluppato al livello più conveniente ed eventualmente fatto evolvere verso livelli superiori
- la semantica dei termini può essere descritta nell'ambito della stessa rappresentazione utilizzata per i dati
  - è possibile sviluppare applicazioni che operano contemporaneamente su dati e schemi con gli stessi strumenti (validatori, motori di inferenza, …)

accesso distribuito

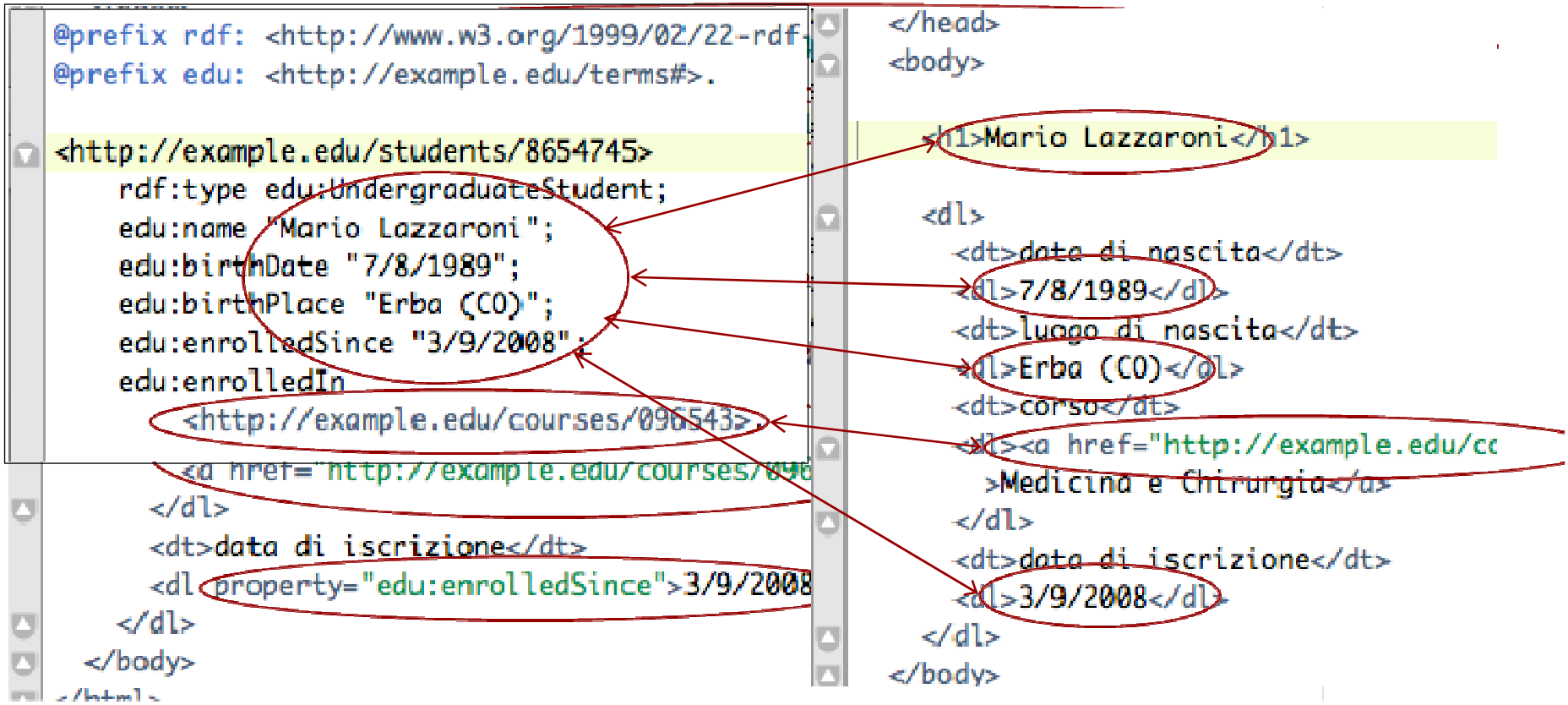


accesso distribuito
serializzazioni testuali (XML/Turtle)

Example.ttl

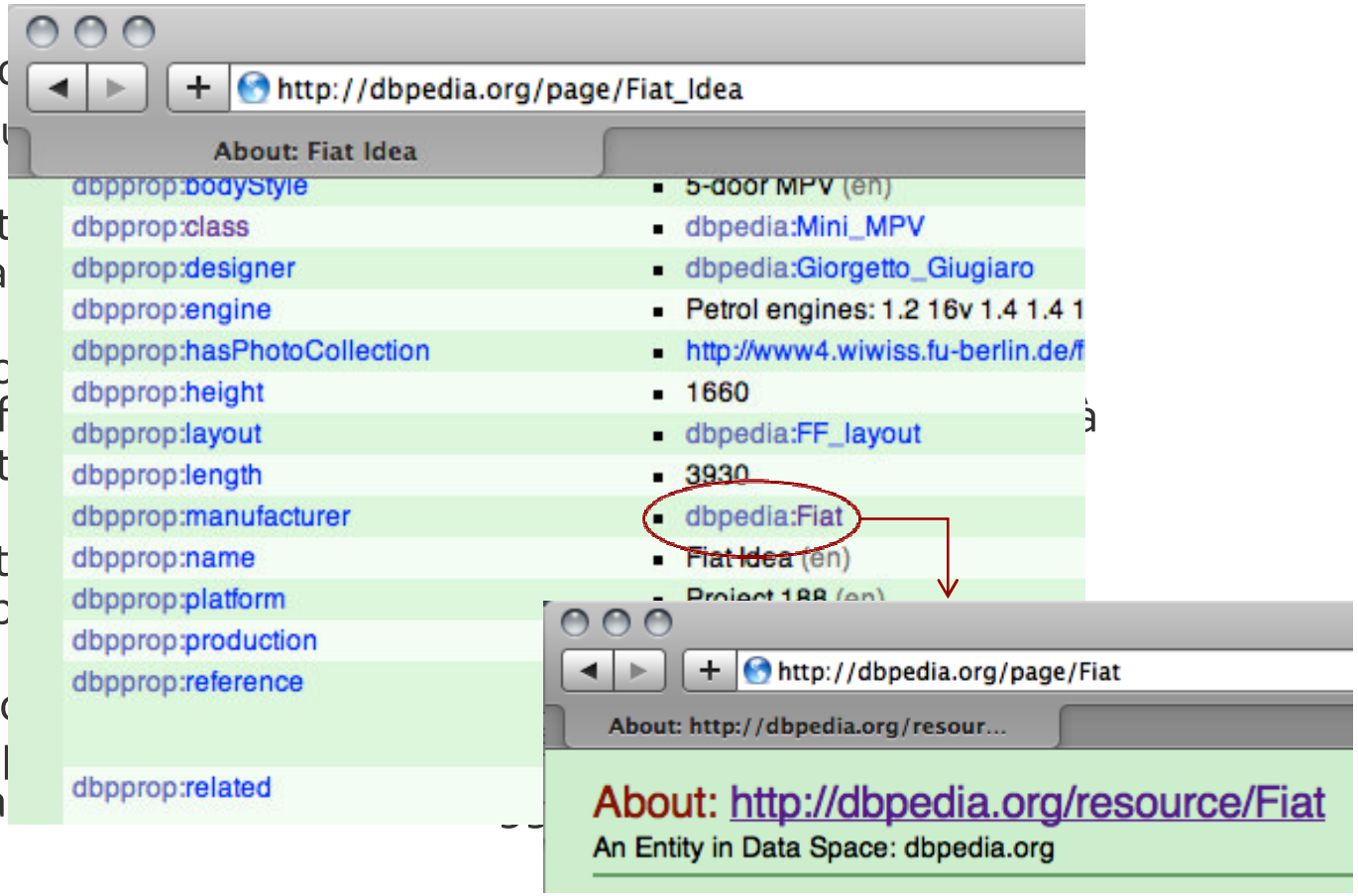
```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix edu: <http://example.edu/term#>.

<http://example.edu/student/8654745> rdf:type edu:UndergraduateStudent;
    edu:name "Mario Lazzaroni";
    edu:birthDate "7/8/1989";
    edu:birthPlace "Erba (CO)";
    edu:enrolledSince "3/9/2008";
    edu:enrolledIn <http://example.edu/course/096543>.

<http://example.edu/course/09654> rdf:type edu:UndergraduateCourse;
    edu:name "Medicina e Chirurgia".
    </edu:UndergraduateCourse
</rdf:RDF>  "Medicina e Chirurgia"  edu:UndergraduateCourse
```

accesso distribuito
serializzazioni embedded (RDFa)



accesso distribuito linked data

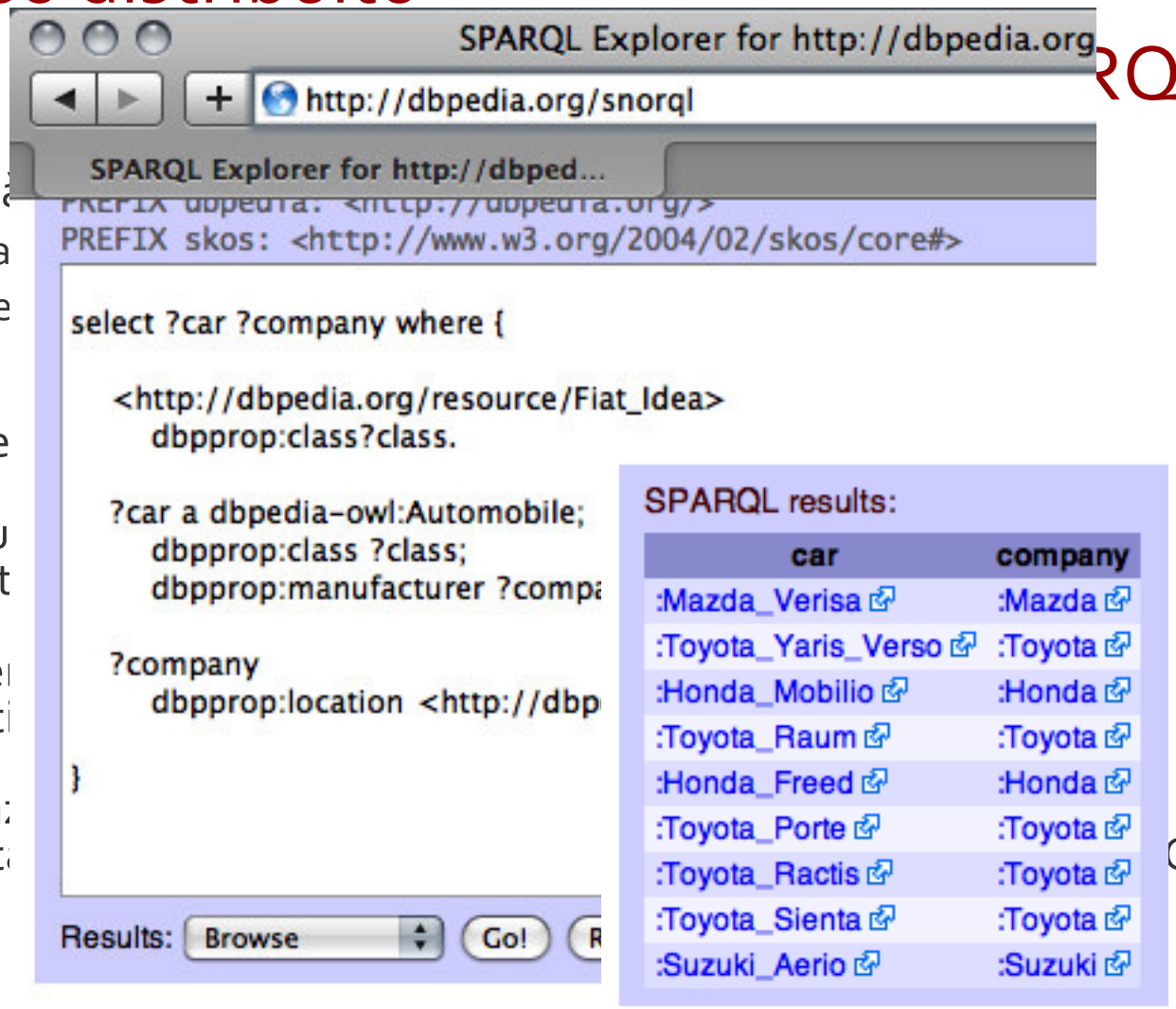
- modalità d
  - basata su
- identificat
  punto di a
- contenutc
  sotto-graf
  identificat
- riferiment
  utilizzati p
- negoziazi
  utilizzata
  (serializza



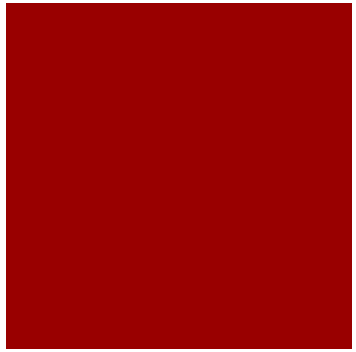
accesso distribuito

query RQL)

- modalità
  - basata
  - inviate
- query > descrive
- contenu data-set
- riferime utilizzati
- negozia utilizzat ON)



uk central office of information



uk central office of information
requisiti e contesto

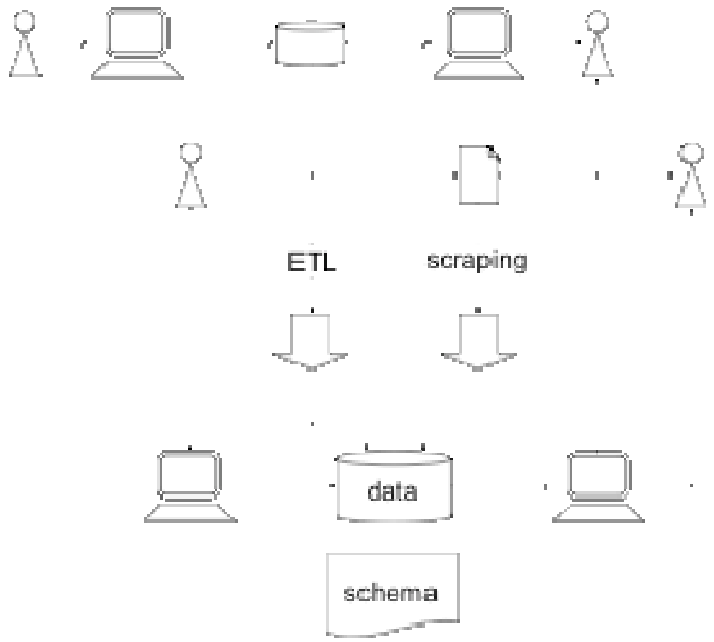
- un'applicazione di ricerca e consultazione per l'offerta globale di lavoro nel settore pubblico
- devono essere consolidati ed integrati dati provenienti da un vasto insieme di fonti fortemente differenziate per
  - dipendenza istituzionale
  - collocazione geografica
  - dimensioni e modalità tecniche di gestione e pubblicazione

uk central office of information approccio centralizzato

- tutte le fonti locali inseriscono direttamente le proprie offerte in un archivio centrale che alimenta l'applicazione di consultazione
- la soluzione deve essere accettata o fatta accettare...
  - devono essere sostituiti o affiancati tutti i sistemi locali
  - modificati tutti i flussi di lavoro
  - riaddestrato il personale

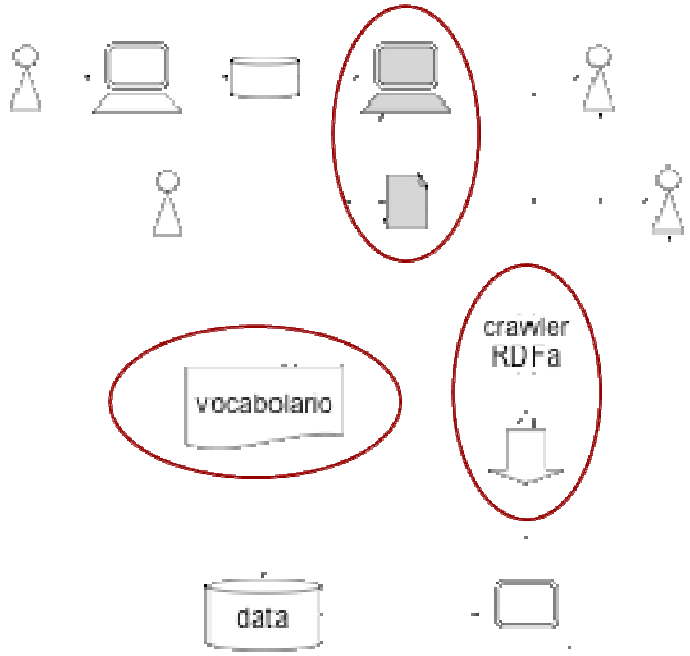
uk central office of information approccio decentralizzato

- le fonti vengono gestite localmente e successivamente consolidate in un archivio centrale che alimenta l'applicazione di consultazione
- è necessario ottenere la collaborazione delle fonti locali ed avere risorse centrali adeguate…
  - devono essere analizzati tutti gli schemi locali e definito un formato di interscambio
  - implementate e mantenute le relative procedure ETL
  - fornite applicazioni ad-hoc o sistemi di scraping per i sistemi locali sprovvisti di DB



uk central office of information approccio semantico

- le fonti locali vengono consolidate in un archivio semantico che alimenta l'applicazione di consultazione
  - viene definito un vocabolario comune per la descrizione delle offerte di lavoro
  - annotate le offerte pubblicate come pagine HTML, utilizzando il vocabolario comune e RDFa
  - estratte le descrizioni RDF dalle pagine HTML tramite un crawler che utilizza la stessa interfaccia destinata agli utenti finali
- è ancora necessario ottenere la collaborazione delle fonti locali ma ad un diverso livello di impegno…



uk central office of information annotazioni RDFa

uk central office of information discussione

- nessun intervento sull'infrastruttura
  - complessità tecnica, tempi e costi di intervento sono ridotti
  - diventa possibile recuperare i dati superando le resistenze politiche
- distribuzione delle attività
  - coinvolgimento di risorse aggiuntive e maggiore focalizzazione
- gradualità dell'intervento
  - la fattibilità viene verificata in tempi brevi e con investimenti limitati
- disaccoppiamento dei dati dalle applicazioni di consultazione
  - l'informazione può essere riutilizzata in contesti e con obiettivi diversi

conclusioni

- le tecnologie semantiche offrono un approccio pragmatico ed incrementale per la pubblicazione e l'integrazione di fonti informative ad oggi non economicamente utilizzabili
- standard e tecnologia vanno rapidamente maturando ed assumeranno un ruolo strategico su un orizzonte di 2/3 anni
- opportuno iniziare un percorso formativo su queste tematiche