



# semantic data

alessandro bollini

20.11.2009

# introduzione



## obiettivi

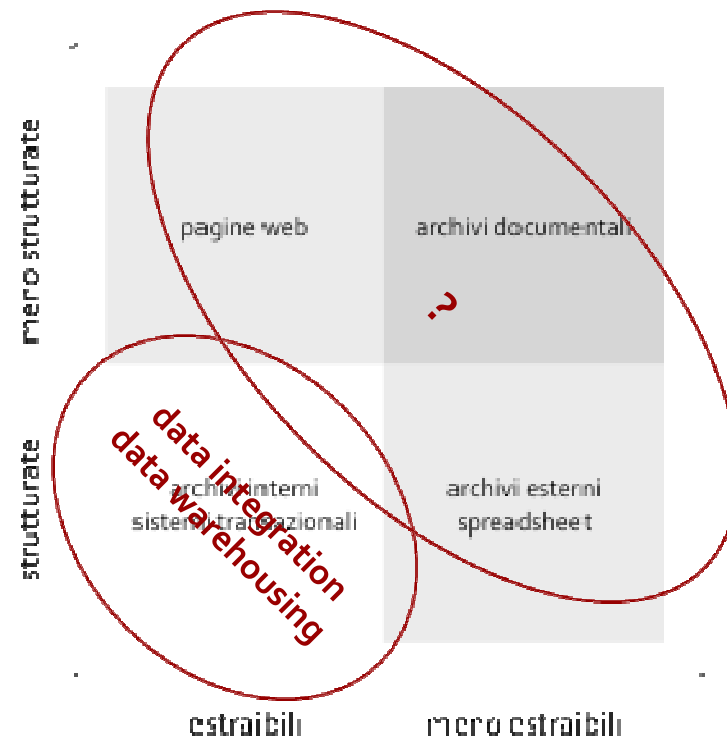
- introduzione alle tecnologie semantiche
- applicate al trattamento di dati distribuiti e parzialmente strutturati

## 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



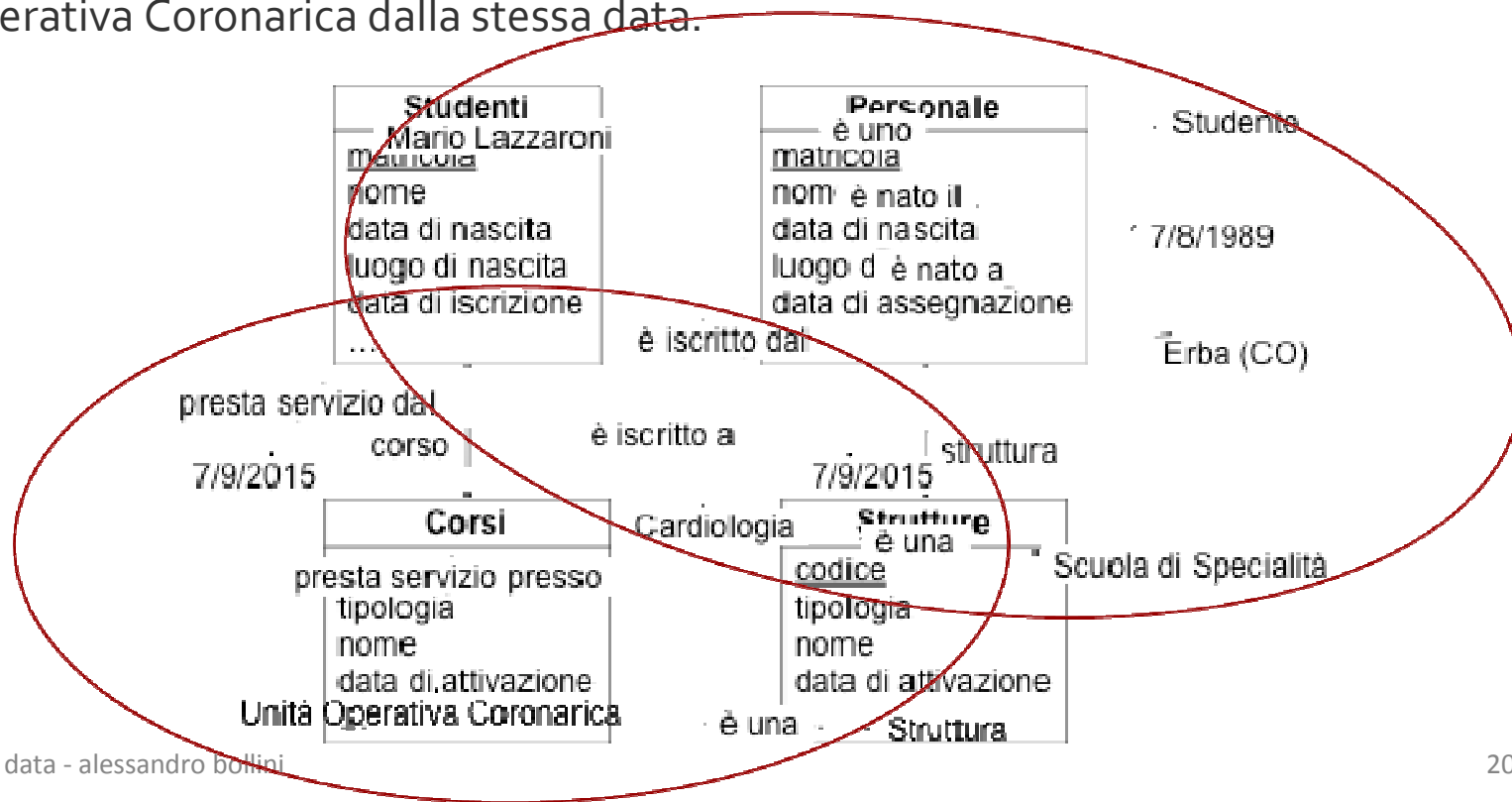
# rappresentazione dell'informazione modello reticolare

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

Studenti		
Mario Lazzaroni	matricola	è uno Studente
soggetto	predicato	oggetto
Mario Lazzaroni	è nato il	Studente 7/8/1989
Mario Laz	è iscritto dal	nato : è nato a a (CO)
Mario Lazzaroni	è nato il	7/8/1989
Mario Lazzaroni	è iscritto a	Medicina Erba (CO)
Mario Lazzaroni	3/9/2008	è iscritto dal
Mario Lazzaroni	è iscritto dal	3/9/2008
Medicina e Chirurgia	è un	Corso di Laurea
Corsi		
Medicina e Chirurgia	è un	Corso di Laurea
	codice	
	tipologia	
	nome	
	data di attivazione	
	...	

## 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

<i>soggetto</i>	<i>predicato</i>	<i>oggetto</i>
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

<i>soggetto</i>	<i>predicato</i>	<i>oggetto</i>
<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

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

@prefix edu: <http://example.edu/terms#>.

<http://example.edu/students/8654745>

rdf:type

edu:UndergraduateStudent

edu:enrolledSince

edu:name

"3/9/2008"

edu:birthDate

"Mario Lazzaroni"

edu:enrolledIn

edu:birthPlace

<http://example.edu/courses/096543>

"Erbà (CO)"

"7/8/1989"

edu:name

rdf:type

"Medicina e Chirurgia"

edu:UndergraduateCourse

# 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?

@prefix edu: <http://example.edu/terms#>.

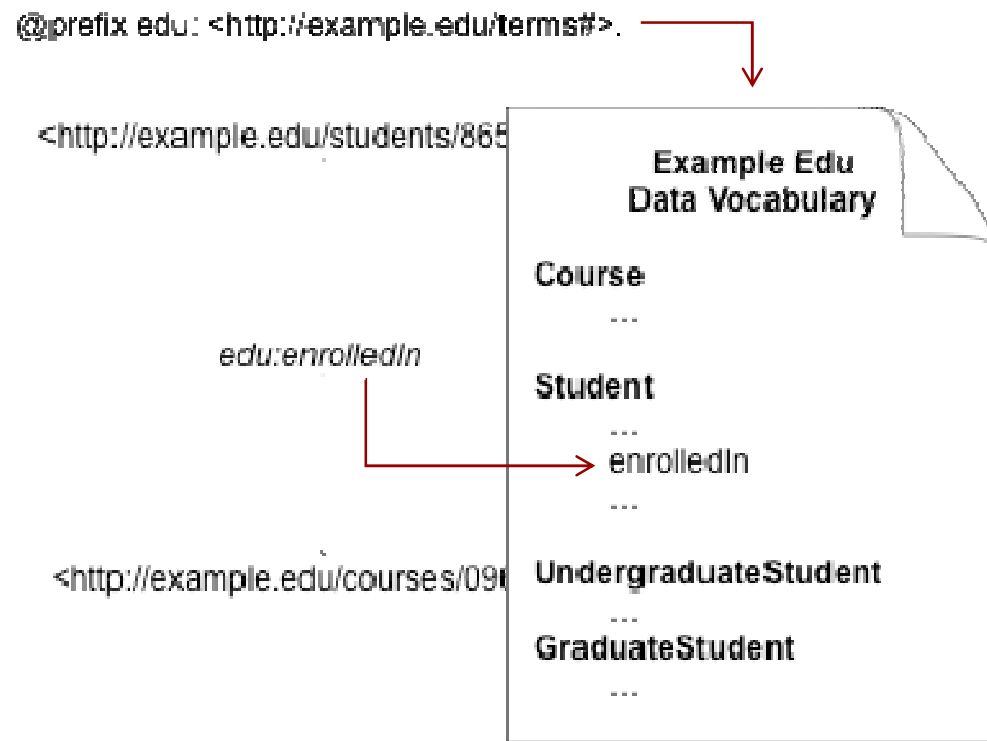
<http://example.edu/students/8654745>

*edu:enrolledIn*

<http://example.edu/courses/096543>

# 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, ...)

```

@prefix rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix skos: <http://www.w3.org/2008/05/skos>.
@prefix edu: <http://example.edu/terms#>.

```

```

=> Course
edu:Teacher rdfs:type skos:Concept
Professor
Full Professor skos:related rdfs:type
Associate Professor
Lecturer skos:broader edu:Course
= Instructor skos:broader
edu:Professor edu:Lecturer
skos:prefLabel
skos:broader skos:altLabel "lecturer"
skos:broader "instructor"
edu:FullProfessor edu:AssociateProfessor

```

# 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, ...)

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix edu: <http://example.edu/terms#>.

Class: Teacher
Class: Student
Class: Professor
Class: GraduateStudent

SubClassOf:
  Teacher
  Student.
  supervisor only Professor
  supervisor exactly 1
  edu:GraduateStudent

ObjectProperty: supervisor
  rdf:type
    edu:GraduateStudent(?Student).
    edu:supervisor(?Student. ?Supervisor)
    edu:appointedTo(?Supervisor. ?Structure).
    edu:Department(?Structure)
    => edu:appointedTo(?Student. ?Structure)
  owl:unProperty

  owl:allValuesFrom
  rdf:type
    owl:Restriction

edu:Professor

```

## 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:09654>
</rdf:RDF> "Medicina e Chirurgia" edu:UndergraduateCourse
```

## accesso distribuito serializzazioni embedded (RDFa)

The image shows a side-by-side comparison of RDFa source code and its HTML rendering. Red circles and arrows indicate the mapping between the two.

**RDFa Source Code (Left):**

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-
@prefix edu: <http://example.edu/terms#>.

<http://example.edu/students/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/courses/096543>
    <a href="http://example.edu/courses/096543">
      </dl>
      <dt>data di iscrizione</dt>
      <dl property="edu:enrolledSince">3/9/2008
    </dl>
  </body>
</html>

```

**HTML Rendering (Right):**

```

</head>
<body>
  <h1>Mario Lazzaroni</h1>
  <dl>
    <dt>data di nascita</dt>
    <dl>7/8/1989</dl>
    <dt>luogo di nascita</dt>
    <dl>Erba (CO)</dl>
    <dt>corso</dt>
    <dl><a href="http://example.edu/cc
      >Medicina e Chirurgia</a>
    </dl>
    <dt>data di iscrizione</dt>
    <dl>3/9/2008</dl>
  </dl>
</body>

```

**Mapping (Red Circles and Arrows):**

- `edu:name "Mario Lazzaroni"` maps to `<h1>Mario Lazzaroni</h1>`.
- `edu:birthDate "7/8/1989"` maps to `<dl>7/8/1989</dl>`.
- `edu:birthPlace "Erba (CO)"` maps to `<dl>Erba (CO)</dl>`.
- `edu:enrolledSince "3/9/2008"` maps to `<dl>3/9/2008</dl>`.
- `edu:enrolledIn <http://example.edu/courses/096543>` maps to `<dl><a href="http://example.edu/cc >Medicina e Chirurgia</a></dl>`.

# accesso distribuito linked data

- modalità di accesso
  - basata su URI
- identificazione del punto di accesso
- contenuto del sotto-grafo identificato
- riferimenti utilizzati per
- negoziazione utilizzata per (serializza

The image shows two browser windows illustrating linked data access. The top window displays the 'About: Fiat Idea' page with a list of properties and their values. The 'manufacturer' property is highlighted with a red circle, and an arrow points to a smaller window below it. This smaller window shows the resource URI 'http://dbpedia.org/resource/Fiat' and identifies it as an entity in the Data Space dbpedia.org.

Property	Value
dbpprop:bodyStyle	5-door MPV (en)
dbpprop:class	dbpedia:Mini_MPV
dbpprop:designer	dbpedia:Giorgetto_Giugiaro
dbpprop:engine	Petrol engines: 1.2 16v 1.4 1.4 1.6
dbpprop:hasPhotoCollection	<a href="http://www4.wiwiss.fu-berlin.de/f">http://www4.wiwiss.fu-berlin.de/f</a>
dbpprop:height	1660
dbpprop:layout	dbpedia:FF_layout
dbpprop:length	3930
dbpprop:manufacturer	dbpedia:Fiat
dbpprop:name	Fiat Idea (en)
dbpprop:platform	Project 188 (en)
dbpprop:production	
dbpprop:reference	
dbpprop:related	

# accesso distribuito query

- modalità
  - basata
  - inviate
- query > describe
- contenuto data-set
- riferimenti utilizzati
- negoziazione utilizzata

SPARQL Explorer for http://dbpedia.org

http://dbpedia.org/snorql

SPARQL Explorer for http://dbped...

PREFIX dbpedia: <http://dbpedia.org/>  
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

```
select ?car ?company where {
  <http://dbpedia.org/resource/Fiat_Idea>
    dbpprop:class?class.

  ?car a dbpedia-owl:Automobile;
    dbpprop:class ?class;
    dbpprop:manufacturer ?compa

  ?company
    dbpprop:location <http://dbp

}
```

Results:

RQL)

## SPARQL results:

car	company
:Mazda_Verisa <a href="#">↗</a>	:Mazda <a href="#">↗</a>
:Toyota_Yaris_Verso <a href="#">↗</a>	:Toyota <a href="#">↗</a>
:Honda_Mobilio <a href="#">↗</a>	:Honda <a href="#">↗</a>
:Toyota_Raum <a href="#">↗</a>	:Toyota <a href="#">↗</a>
:Honda_Freed <a href="#">↗</a>	:Honda <a href="#">↗</a>
:Toyota_Porte <a href="#">↗</a>	:Toyota <a href="#">↗</a>
:Toyota_Ractis <a href="#">↗</a>	:Toyota <a href="#">↗</a>
:Toyota_Sienta <a href="#">↗</a>	:Toyota <a href="#">↗</a>
:Suzuki_Aerio <a href="#">↗</a>	:Suzuki <a href="#">↗</a>

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

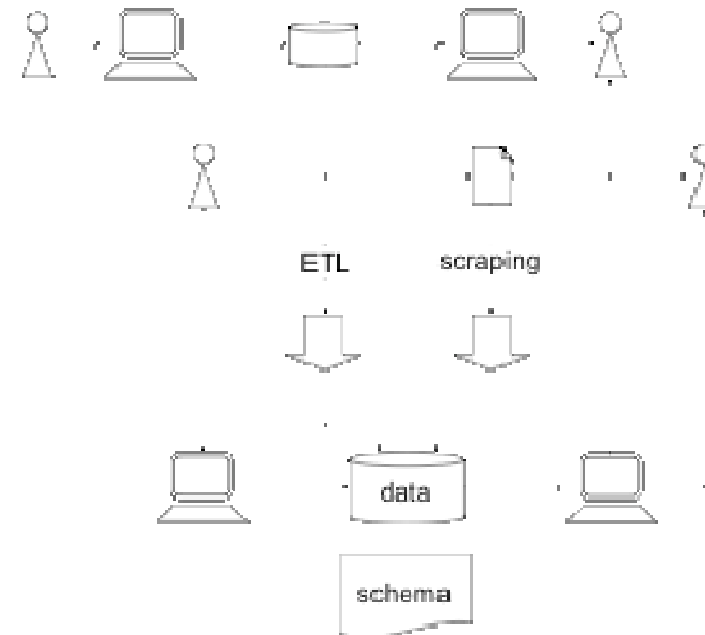
# uk central office of information approccio centralizzato

- 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 annotazioni RDFa

```

@prefix dc: <http://purl.org/dc/terms/> .
@prefix arg: <http://purl.oclc.org/argot/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://www.civilservice.gov.uk/jobs/careers-detail.aspx?JobId=6896>
  a foaf:Document ;
  dc:publisher <http://www.civilservice.gov.uk/> ;
  dc:type arg:Vacancy ;
  dc:source <http://www.civilservice.gov.uk/jobs/careers-detail.aspx?JobId=6896> ;
  dc:title "Supreme Court Support Staff" ;
  dc:identifier "SGB/DL/09/09" ;
  arg:salaryFrom "12764" ;
  arg:salaryTo "14586" ;
  arg:salaryPeriod "Per Annum" ;
  dc:coverage "City of Edinburgh, Scotland" ;
  dc:valid "2009-09-10" ;

```

NDPBs	Location	Appointment Terms
	City of Edinburgh, Scotland	Monthly

# 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