Linked Open Data

L’esperienza piemontese

Vittorio Di Tomaso | CELI – Language & Information Technology

Quest' opera è distribuita con licenza Creative Commons Attribuzione - Non commerciale - Non opere derivate 3.0 Unported.

REGIONE PIEMONTE csi piemonte DATI.Piemonte.it CELI LANGUAGE & INFORMATION TECHNOLOGY

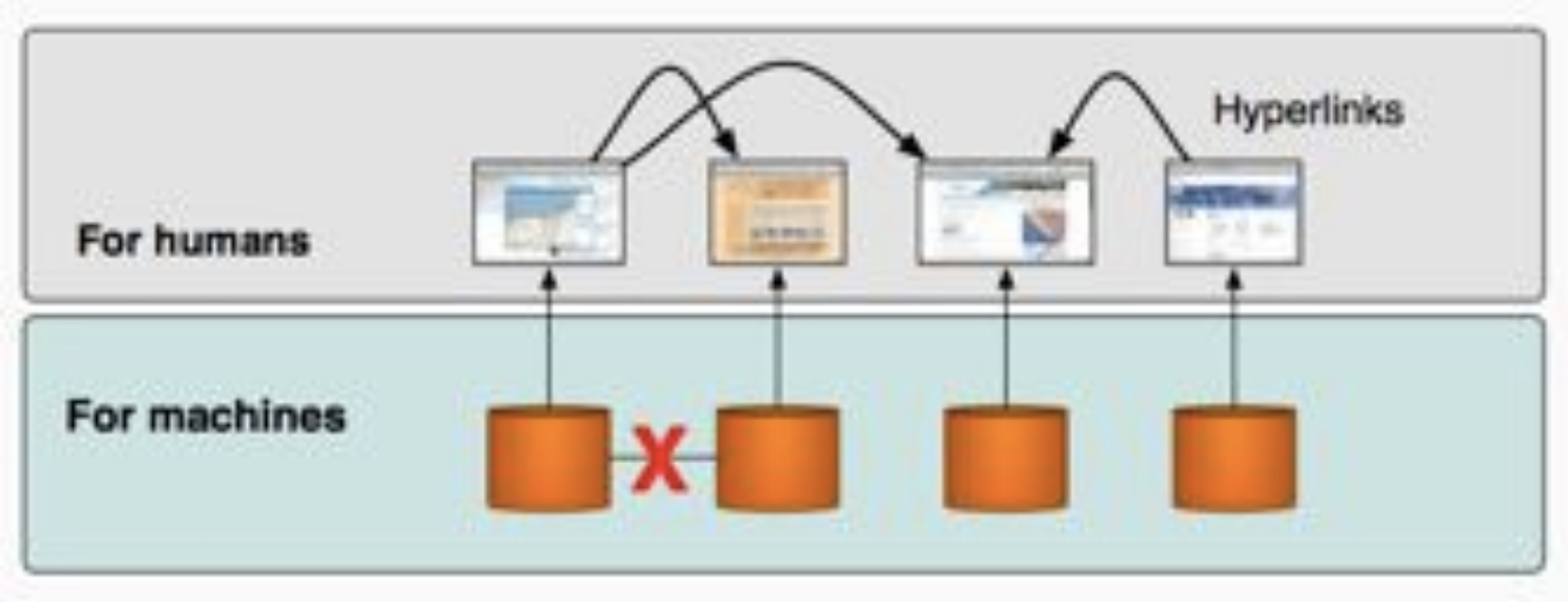
Linked Open Data (LOD)

LOD

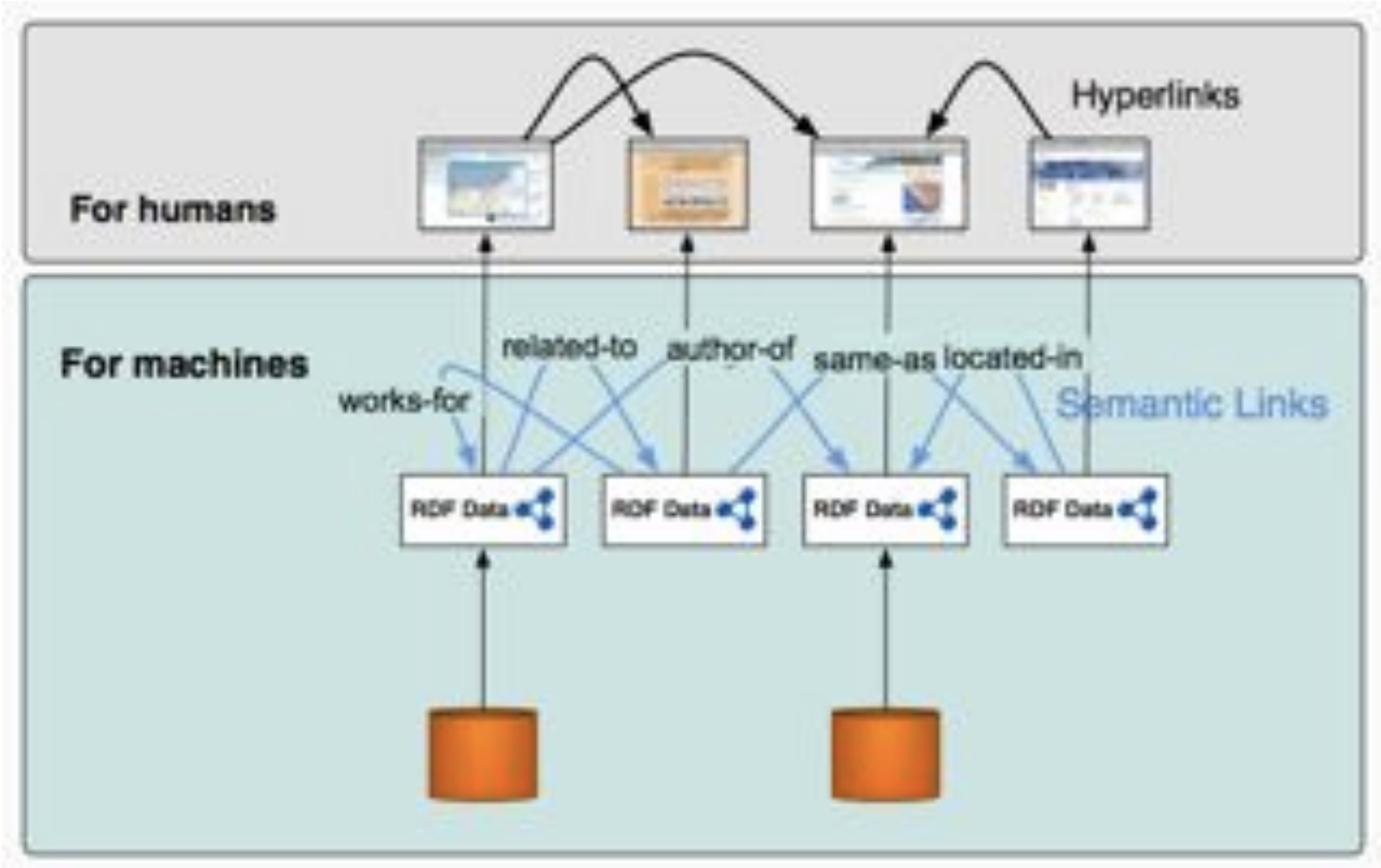
Una serie di “best practices” per pubblicare dati strutturati e interconnessi sul Web

L’idea è usare la stessa infrastruttura del Web (of documents) per costruire il Web of Data

Web of Documents



Web of Data


Hyperlinks
For humans
For machines
related-to
author-of
same-as
located-in
works-for
Semantic Links
RDF Data
RDF Data
RDF Data
RDF Data

Linked Open Data: perché?

I dati hanno meno valore quando sono isolati

LOD

Trasparenza ★
Rendere i dati legalmente aperti e fruibili sul web

Condivisione ★★
Rendere i dati tecnicamente aperti (disponibili come dati strutturati xls vs pdf)

Ri-uso ★★★
Rendere i dati tecnicamente aperti in un formato aperto (csv, vs, xls)

Linked Data ★★★★
Rendere i dati identificabili tramite URLs

Web Semantico ★★★★★
Collegare i propri dati ai dati degli altri

1: Rendere i dati disponibili

L’apertura dei dati è solo un primo (necessario) passo

Occorre che i dati aperti siano il più possibile effettivamente utilizzabili (da persone, ma anche da software)

La vertigine del CSV

```
provincia   denominazione   indirizzo   cap comune  qualifica   stelle  telefono    fax email   numCamere   numLetti    numBagni
ALESSANDRIA AFFITTACAMERE   CASCINA SAN SEBASTIANO, 71  15030   ROSIGNANO MONFERRATO    Affittacamere   nessuna             5   10
ALESSANDRIA AFFITTACAMERE   VIA ALESSANDRIA, 73 15019   STREVI  Affittacamere   nessuna             4   8   2
ALESSANDRIA AFFITTACAMERE ARCOBALENO    VIA VERONA, 102 15100   ALESSANDRIA Affittacamere   nessuna 3383127315  131257420   magrane
ALESSANDRIA AFFITTACAMERE GOGGI MARIA ROSA  VIA DANTE, 6    15045   SALE    Affittacamere   nessuna 13184141            5   11  3
ALESSANDRIA AFFITTACAMERE RA MANSENA    VIA MAZZINI GIUSEPPE, 5 15040   ALLUVIONI CAMBIO'   Affittacamere   nessuna 13184207
ALESSANDRIA AFFITTACAMERE  TENUTA PRELI STRADA DEI BOTTI, 18    15011   ACQUI TERME Affittacamere   nessuna  0144 313374    0144 31
ALESSANDRIA ALPI    FRAZIONE BORASSI, 26    15060   ROCCAFORTE LIGURE   Affittacamere   nessuna 14394167    143949900       3   8
ALESSANDRIA B&B LE ZIE  PIAZZA BORGATTA, 2  15078   ROCCA GRIMALDA  Affittacamere   nessuna 3343387659          1   2   1
ALESSANDRIA BELLAGIO    LOCALITA'  VALORIA SUPERIORE, 27    15076   OVADA   Affittacamere   nessuna 143835892   143835892       4
ALESSANDRIA BEL SOGGIORNO   VIA UMBERTO I?, 69/2    15010   CREMOLINO   Affittacamere   nessuna 0143 879012 0143 879921 info@ristor
ALESSANDRIA CA' DUBINI  VIA ROMA, 17    15020   MOMBELLO MONFERRATO Affittacamere   nessuna 142944116   142944928   info@cadubini.i
ALESSANDRIA CA' MIMIA   CASCINA MIMIA INFERIORE, 46 15070   TAGLIOLO MONFERRATO Affittacamere   nessuna 0143 89526          2   4
ALESSANDRIA CASA DEGLI ARTISTI  VIA ASTI, 8 15021   ALFIANO NATTA   Affittacamere   nessuna 0141 922455     info@casadegliartisti.i
ALESSANDRIA CASTELLO DI TRISOBBIO   CAVOUR, 1   15070   TRISOBBIO   Affittacamere   nessuna 0143 831108 0143 831742 info@ilcastello
ALESSANDRIA CENTRO TURISTICO LAGHI ISOLABELLA   LOCALITA' ISOLABELLA, 22/B  15050   ISOLA SANT'ANTONIO  Affittacamere   nessuna 013
ALESSANDRIA CENTRO TURISTICO MANEGGIO ISOLABELLA    LOCALITA' ISOLABELLA, 22/A  15050   ISOLA SANT'ANTONIO  Affittacamere   nessuna
ALESSANDRIA CENTRO TURISTICO RISTORO ISOLABELLA LOCALITA'  CASCINA ISOLABELLA, 22   15050   ISOLA SANT'ANTONIO  Affittacamere   nes
ALESSANDRIA CONCA D'ORO VIA NOVI, 52    15057   TORTONA Affittacamere   nessuna 131817238           4   6   5
ALESSANDRIA CONGREGAZIONE SUORE OBLATE DI SAN LUIGI GONZAGA PIAZZA SAN GUIDO, 14    15011   ACQUI TERME Affittacamere   nessuna 144
ALESSANDRIA DA NONNA CARLA  VIA VITTORIO EMANUELE, 11   15021   ALFIANO NATTA   Affittacamere   nessuna 3203895675      info@danonr
ALESSANDRIA GIANDUIA  AFFITTACAMERE VIALE EINAUDI LUIGI, 35 15011   ACQUI TERME Affittacamere   nessuna 0144 322442 0144 56320  alb
ALESSANDRIA IL GIRASOLE PIAZZA MARTIRI DELLA BENEDICTA, 15  15066   GAVI    Affittacamere   nessuna 143643485   143643485   spedemc
ALESSANDRIA IL GIRASOLE VIA CAV.BERRUTI, 7  15019   STREVI  Affittacamere   nessuna 0144363690/3383498106   144363690   info@bedanc
ALESSANDRIA IL NAVIGATORE   VIA ROMA, 3 15040   CUCCARO MONFERRATO  Affittacamere   nessuna 340 9238068 - 3494384553        brunolu
ALESSANDRIA IMMOBILIARE MARAL   VIA ROMA, 33    15043   FUBINE  Affittacamere   nessuna 0131778347/3337829077           5   8   5
ALESSANDRIA ISTITUTO SAN GIUSEPPE   VIA NOVI, 21    15076   OVADA   Affittacamere   nessuna 0143 80117  0143 834476     5   8   5
ALESSANDRIA LA BUCA 16  CASCINA MOGLIA, 16  15060   CAPRIATA D'ORBA Affittacamere   nessuna 14346260        golfresidence@alice.it
```

2. Rendere i dati utilizzabili

Ostacoli alla effettiva utilizzabilità:

- eterogeneità di modelli (dati poco strutturati, tabelle, CSV, relazionale, applicazioni XML,...)
- eterogeneità di formati (file di testo, file Excel,...)
- eterogeneità di interfacce (API diverse, Web services REST o di altro tipo,...)

2: Rendere i dati utilizzabili

Usare modelli e formati standard

- Attuale standard di riferimento RDF (Resource Description Framework), definito dal W3C: offre un modello uniforme per la rappresentazione dei dati e vari formati di serializzazione
- Conformarsi a RDF aumenta l'interoperabilità e l'effettiva utilizzabilità dei dati (anche da parte di applicazioni sw)

RDF

- La Regione Piemonte ha iniziato un percorso di “RDF-icazione” dei dati pubblicati
- Il valore dei dati aumenta se i dataset non sono più concepiti come silos separati, ma sono esplicitate le interconnessioni esistenti fra i dataset (4 stelle)
- Connessioni fra dataset pubblicati da una medesima P.A., ma anche connessioni con dataset esterni (5 stelle)

3: Adottare LOD

Pubblicazione dei dati secondo i principi dei Linked Data:

1. Usare URI (o IRI) per identificare le cose (non solo risorse Web e contenuti digitali, ma anche elementi del mondo reale, concetti, ecc.)
2. Usare URI HTTP che siano dereferenziabili su protocollo HTTP in una descrizione della risorsa identificata
3. Usare RDF come modello dei dati (ammettere anche interrogazioni SPARQL dei dati)
4. Collegare le risorse tra loro (specificare collegamenti fra le risorse analoghi a quelli ipertestuali usati nel “Web dei documenti”). Questi collegamenti possono collegare risorse in dataset diversi ➔ dai silos di dati ad un Web di dati!

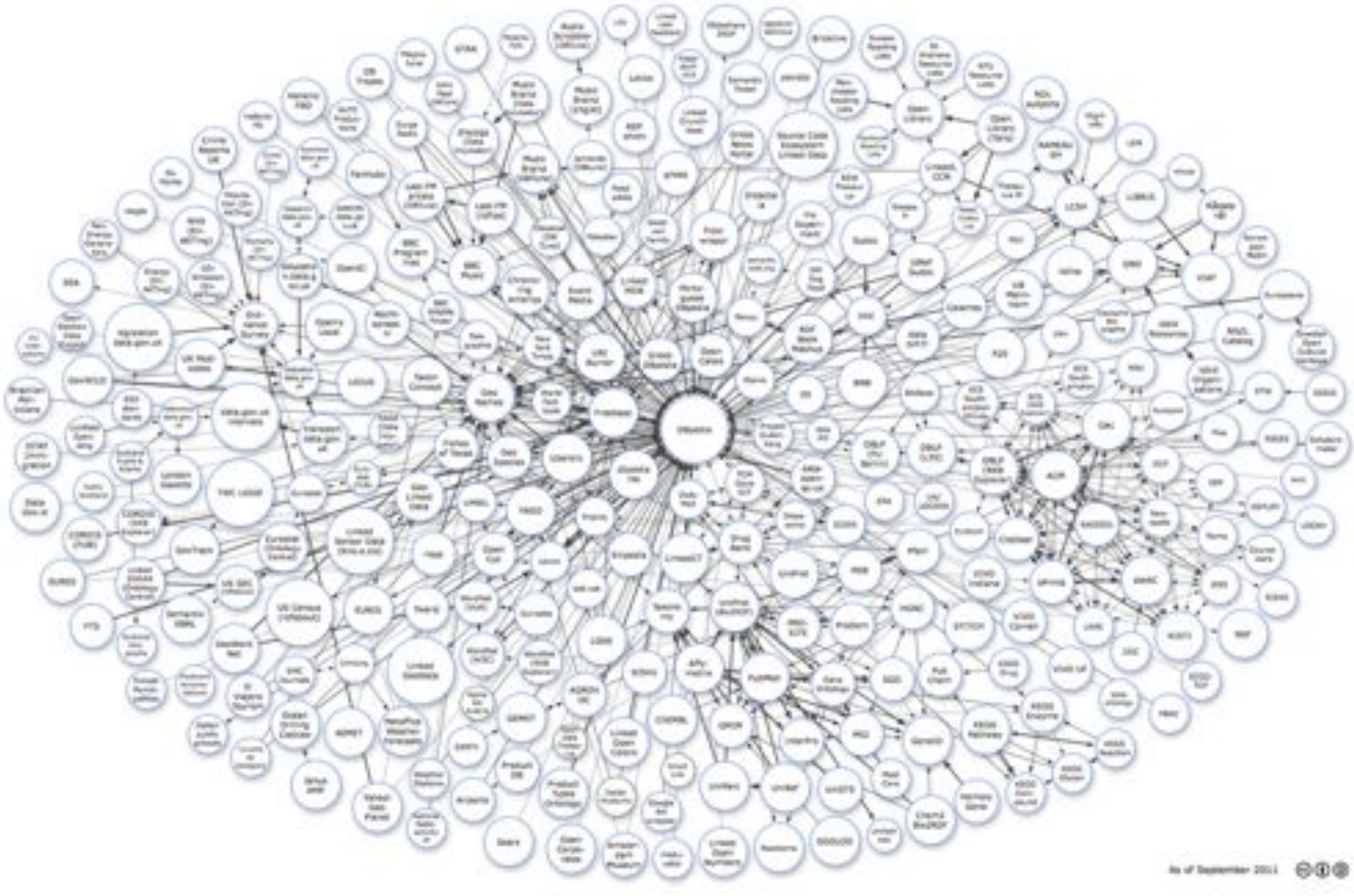
3: Adottare LOD

La pubblicazione e l'interconnessione di dati da parte delle PP.AA. secondo i principi dei **Linked (Open) Data**:

- contribuisce alla crescita del cosiddetto Web semantico o Web di dati
- inserisce i dati nel contesto della Linking Open Data Cloud (e contribuisce alla sua crescita)
- facilita l'integrazione di questi dati fra loro, ma anche con dati provenienti da sorgenti diverse
- facilita l'uso di questi dati in applicazioni sw "intelligenti" (sviluppate dalle PP.AA. stesse o da terze parti...)

aumenta il valore di questi dati per cittadini e imprese

Global Data Space



Cosa stiamo facendo: vocabolari

I dati sono descritti per mezzo di vocabolari/ ontologie che devono essere il più possibile condivisi

Cosa stiamo facendo: RDF

La Regione Piemonte sta percorrendo questa strada

Sono già stati pubblicati alcuni “authoritative files” in RDF che contengono gli URI di alcune risorse (Comuni, Province, Scuole) e i dati “anagrafici” essenziali per ciascuna di esse

Cosa stiamo facendo: linking

Sono in fase di pubblicazione altri dataset RDF per i quali sono state esplicitate le interconnessioni fra le risorse e specificati i collegamenti con noti dataset esterni (DBPedia Italiana, Freebase,...)

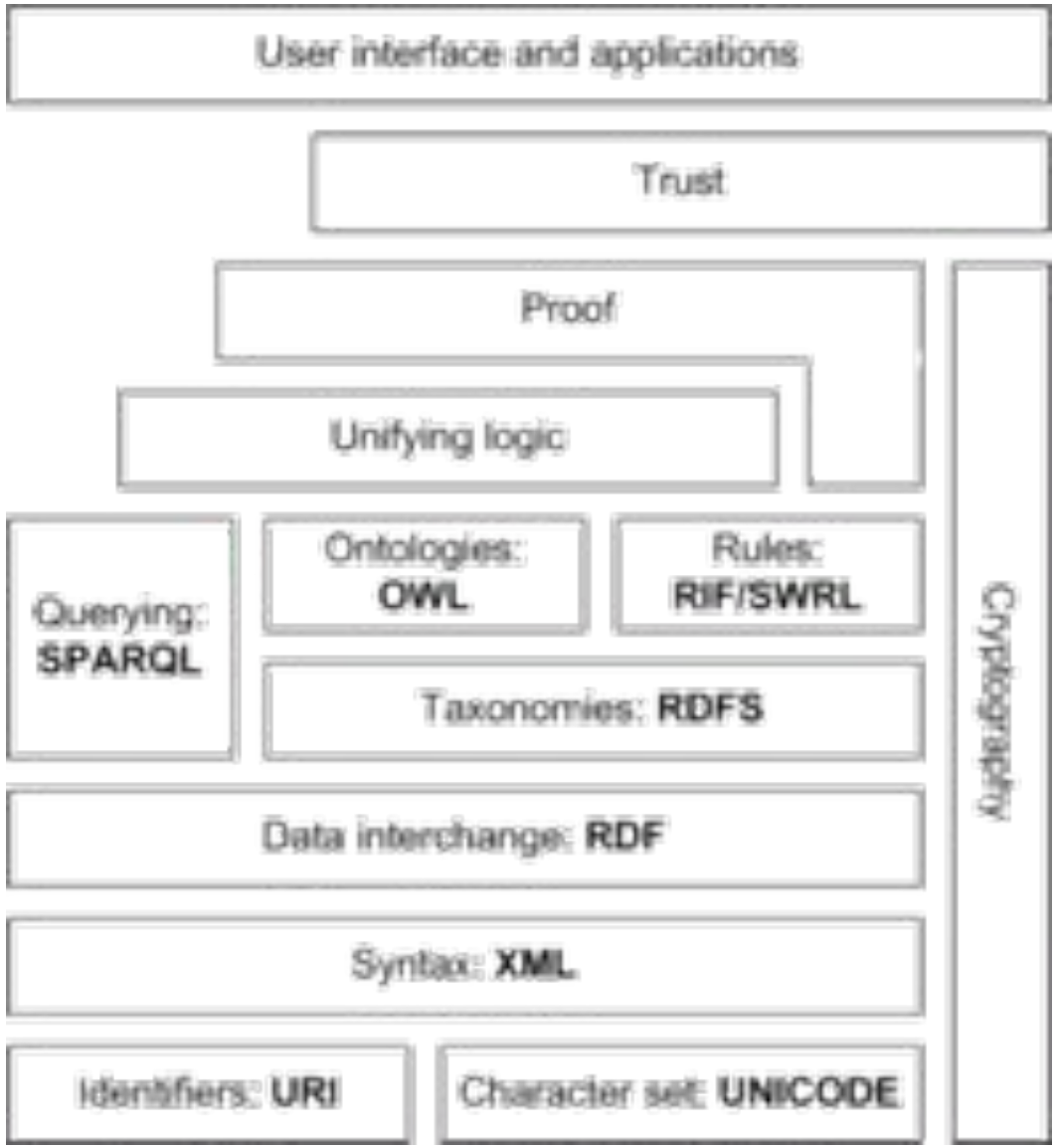
RDF-icazione: datasets

- **Scuole (Sedi)**
  - Dati realativi alle sedi delle scuole piemontesi (Indirizzi, denominazioni, tipologie, etc.)
  - Formato: Excel
- **Scuole (Dati statistici - iscritti)**
  - Rilevazioni statistiche relative al numero di iscritti delle scuole pimoentesi
  - Formato: Excel
- **Esercizi Ricettivi**
  - Dati relativi alle strutture ricettive piemontesi (Indirizzi, denominazioni, tipologie, etc.)
  - Formato: CSV
- **Strutture di Vendita**
  - Dati relativi alle medie e grandi strutture di vendita piemontesi
  - Formato: CSV
- **Centri commerciali**
  - Dati relativi ai centri commerciali piemontesi
  - Formato: CSV

RDF-icazione: metodologia

- **Import dei dati in *google refine***
  - Software open source per la manipolazione dei dati
  - Download e documentazione: http://code.google.com/p/google-refine/
  - Installazione della estensione RDF di refine (http://ab.linkeddata.deri.ie/2010/grefine-rdf-extension/)
- **Ispezione e "tuning" dei dati**
  - Controllo duplicati, uniformazione stringhe, creazione campi derivati, etc.
  - L'utilizzo della funzionalità "facetes" di refine facilita enormemente il check dei dati
- **Scelta dei vocabolari (ontologie)**
  - Riuso di vocabolari esistenti ove possibile
  - Creazione di (piccoli) vocabolari proprietari per modellare caratteristiche particolari
- **Link-ization**
  - Creazione link "interni" (per i comuni e province già identificati da URI)
  - Creazione link "esterni", verso il LOD (**http://it.DBPedia.org**, **http://freebase.com**)
- **RDF mapping**
  - Mappaggio dei campi tabellari in termini dei vocabolari scelti
  - (Non trivial) scelta identificatori per i "nodi intermedi", pattern per esprimere relazioni "complesse"
- **Esportazione dati RDF** (e caricamento du triplestore)
- **Creazione delle descrizioni VOID dei dataset**

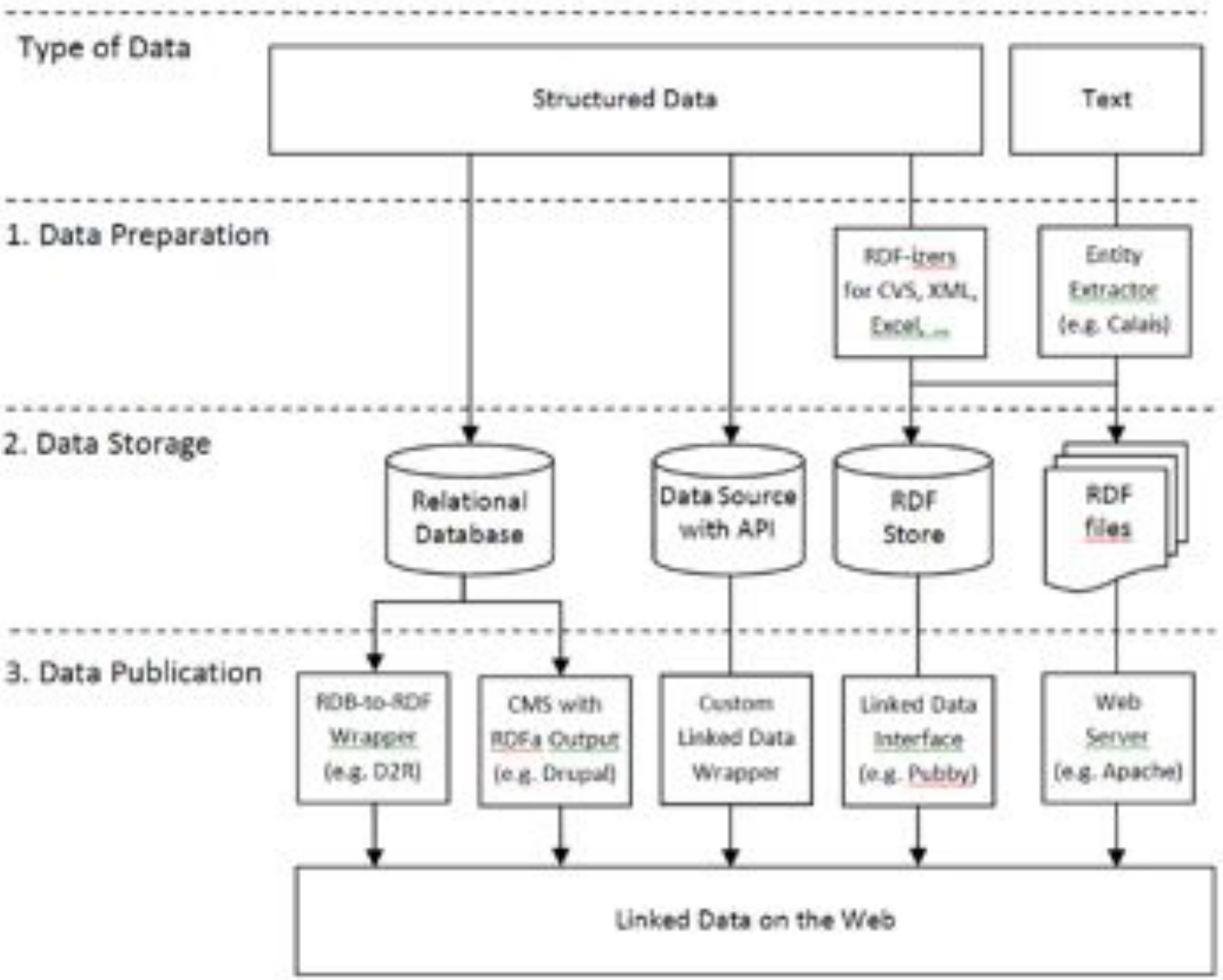
Dove siamo arrivati



1. Character set unicode: OK
2. Identifiers URI: OK
3. Uso di XML: OK
4. Uso di RDF: OK
5. Tassonomie in RDFS: OK
6. SPARQL end-point: WORKING
7. Ontologie OWL: WORKING
8. Rules, Logic, Proof, Trust: TODO
9. User interface + APIs: WORKING

Dove vogliamo arrivare

La strada da percorrere



Un motore di ricerca “semantico”

1. Analisi della query inserita dall’ utente
2. Accesso all’ ontologia e ragionamento per identificare correlazioni
3. Esecuzione della query e reperimento dei risultati, arricchiti dalle correlazioni disponibili
4.

Sophia
SEMANTIC ENGINE

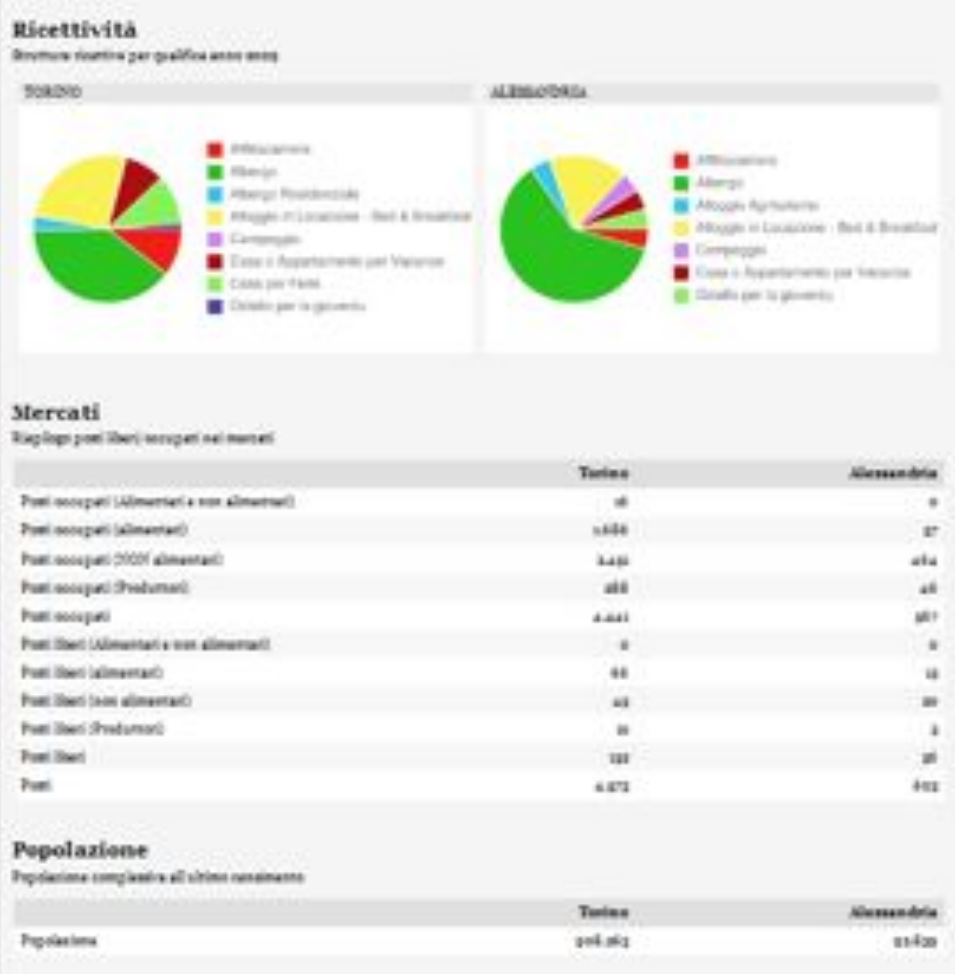
L’ accesso all’ ontologia identifica e rende disponibili:
- relazioni di iperonimia / iponimia (x è un y)
- relazione tra individui (nato a, parte di, localizzato in)
- relazioni di sinonimia
- qualunque altra relazione rappresentabile usando OWL e SKOS…

Confronto tra entità

Quando l'utente inserisce nella maschera di ricerca "Alessandria e Torino", il motore semantico identifica univocamente le due entità presenti nell'ontologia e richiede alla knowledge base le relazioni che collegato questi nodi con altri concetti rilevanti della KB stessa (strutture ricettive, aziende agricole, mercati, popolazione,).

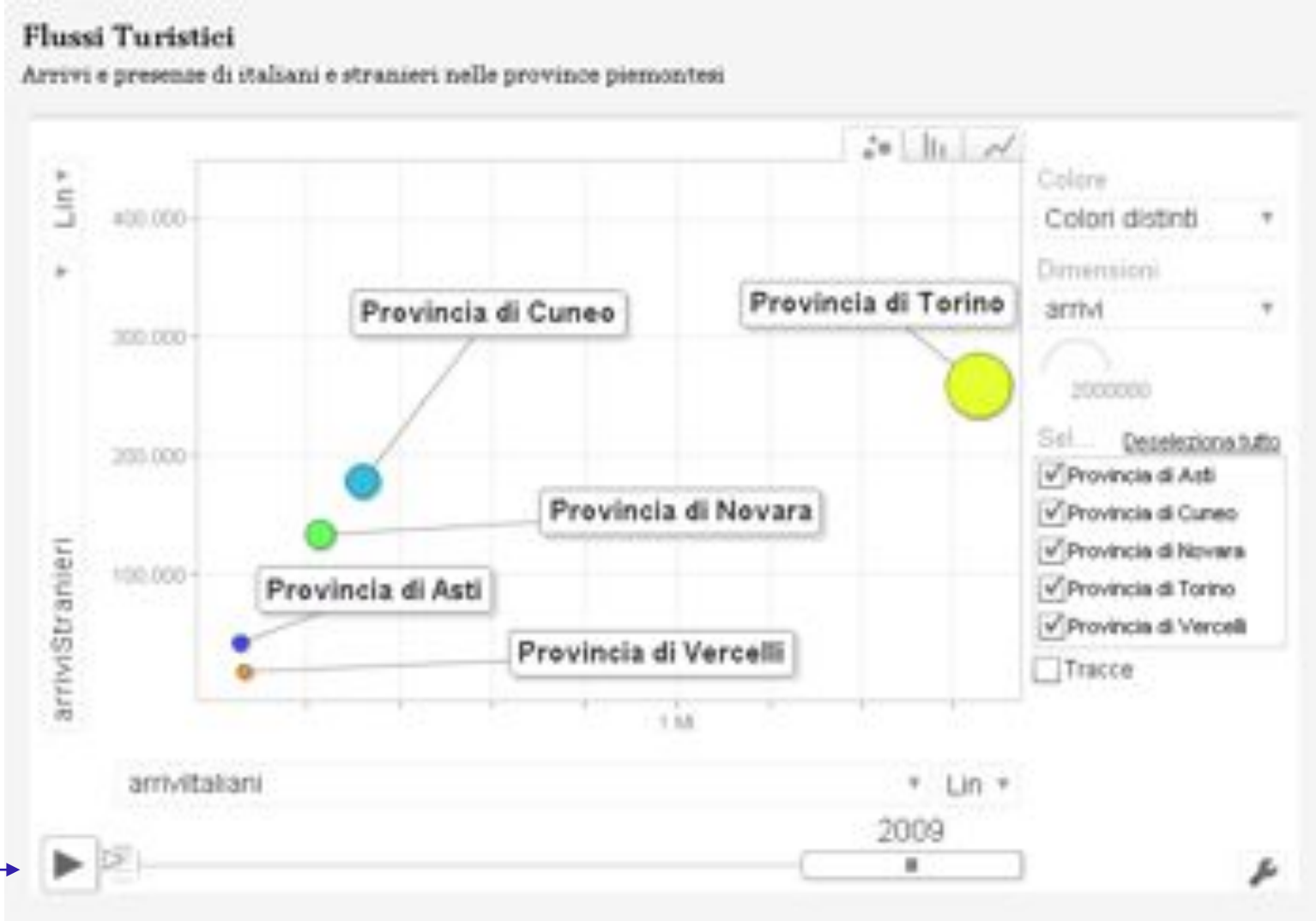
Confronto tra entità

Le informazioni ottenute vengono messe a confronto in parallelo, attraverso specifiche visualizzazioni che fanno uso di grafici, mappe e tabelle (a seconda del tipo di dato disponibile



Evoluzione temporale

Possibilità di confrontare l' evoluzione temporale tra entità diverse dimensioni selezionate (in questo caso i flussi turistici



Per avviare l' animazione

Ricerche in linguaggio naturale

Quando l' utente inserisce nella maschera di ricerca una query complessa, il sistema estrae entità e relazioni e cerca la risposta nella knowledge base
In questo caso: l' entità è il vino Albugnano, e la relazione che si vuole esplorare è "produzione tipica". Il risultato sarebbe diverso se la relazione riconosciuta fosse "mangiare"

Ricerche in linguaggio naturale

L'albugnano è prodotto nei comuni Castelnuovo don bosco, Albugnano, Passerano marmorito, Pino d'Asti

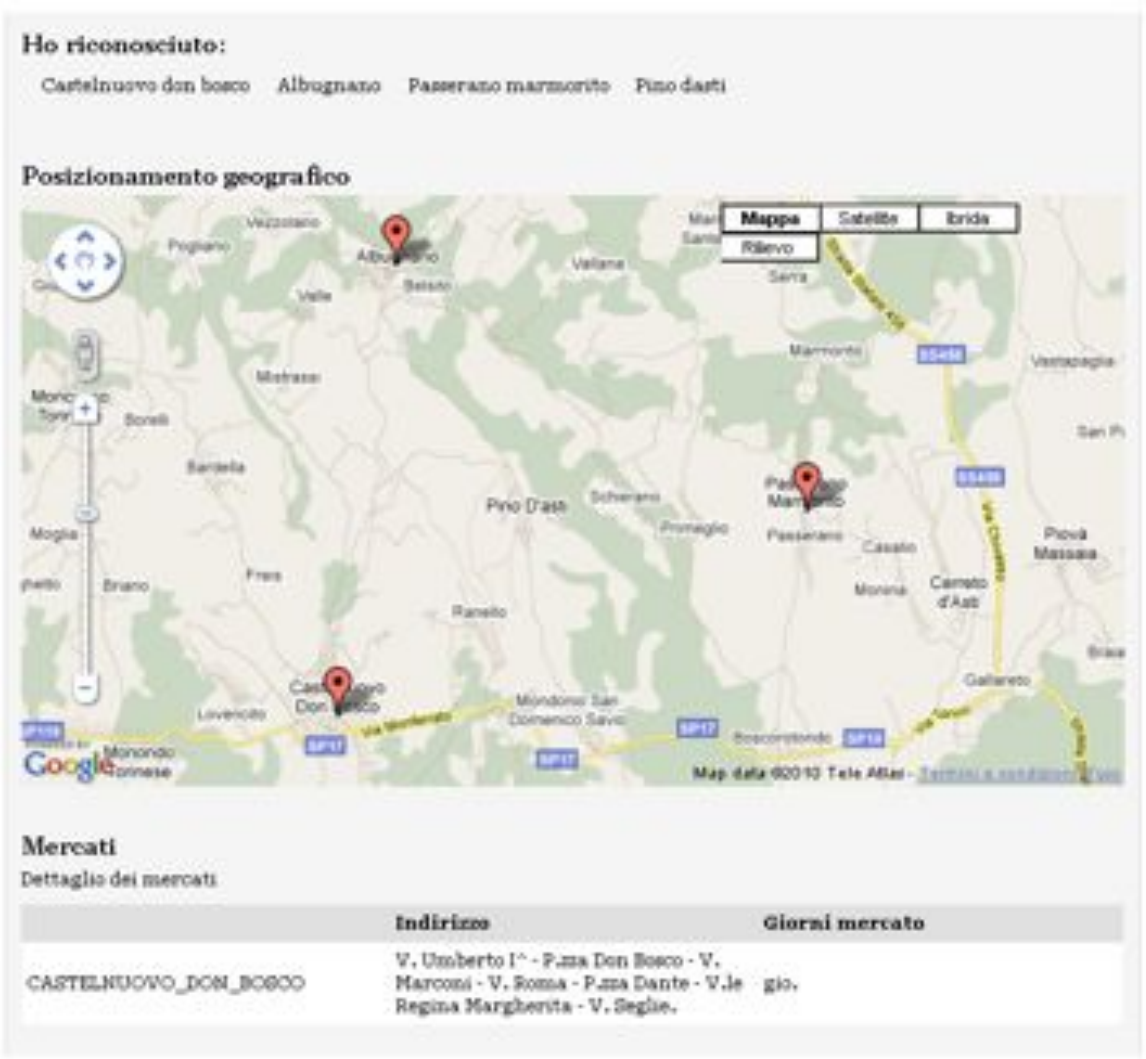


Foto di Giuliana Bonello

Su cosa stiamo ragionando

Quali dati pubblicare secondo principi Linked Data

- Tutti? O Soltanto alcuni?
- Per quelli più “stabili” ha sicuramente senso specificare collegamenti fra dataset interni e con dataset esterni.
- Per i dataset più “volatili” o ad alta frequenza di aggiornamento, è opportuno costruire collegamenti fra dataset o è più opportuno limitarsi alla loro “RDF-icazione”?

Su cosa stiamo ragionando

- In quali circostanze è opportuno “RDF-icare” i dati producendone un’effettiva duplicazione?
- In quali altre circostanze è invece opportuno avvalersi di tool che producono una RDF-icazione dinamica (in tempo reale) dei dati senza effettiva duplicazione?

Su cosa stiamo ragionando

Verso quali dataset esterni costruire i collegamenti?

- Un dataset esterno (es., DBpedia, Freebase, ...) è fuori dal controllo della P.A.
- Stabilire “ufficialmente” un legame fra un dataset pubblicato dalla PA e un dataset esterno è una responsabilità …
- A corollario di quanto sopra: come specificare l’identità in casi come questi (quando sia sensato/sicuro farlo)?
- (tecnicamente: meglio “owl:sameAs” o “skos:exactMatch”?)

Su cosa stiamo ragionando

Che tipo di URI creare?

- URI “parlanti” più human-firendly, ma a rischio di divenire fuorvianti se le risorse identificate o la loro caratterizzazione concettuale dovessero variare significativamente?
- URI “opache”, meno human-friendly, ma anche meno a rischio di divenire fuorvianti?

Su cosa stiamo ragionando

- Quali vocabolari/ontologie utilizzare?
- Sicuramente è utile riutilizzare vocabolari ben noti (SKOS, vCard, DCMI Terms, FOAF, SIOC, DBpedia ontology,...).

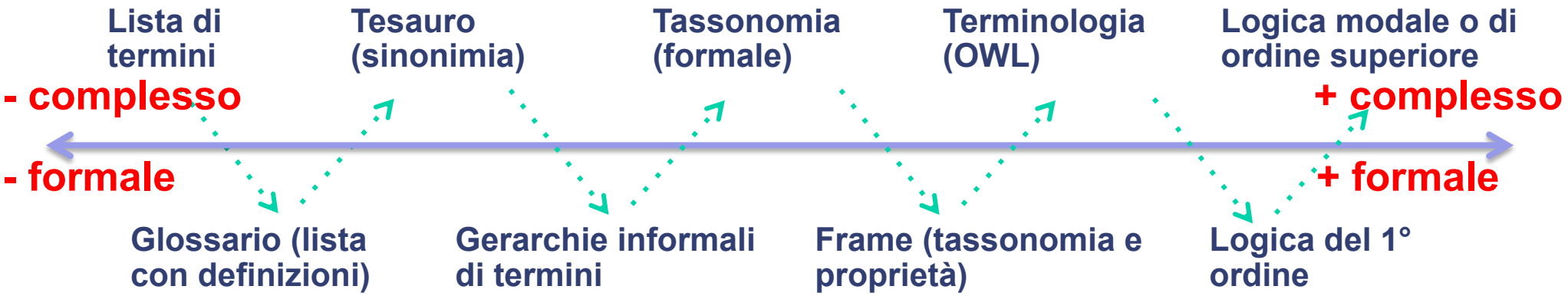
Su cosa stiamo ragionando

- Quando è opportuno creare propri vocabolari/ ontologie?
- Quali linguaggi scegliere per specificarli (sempre restando negli standard W3C...): RDFS? OWL?

Su cosa stiamo ragionando

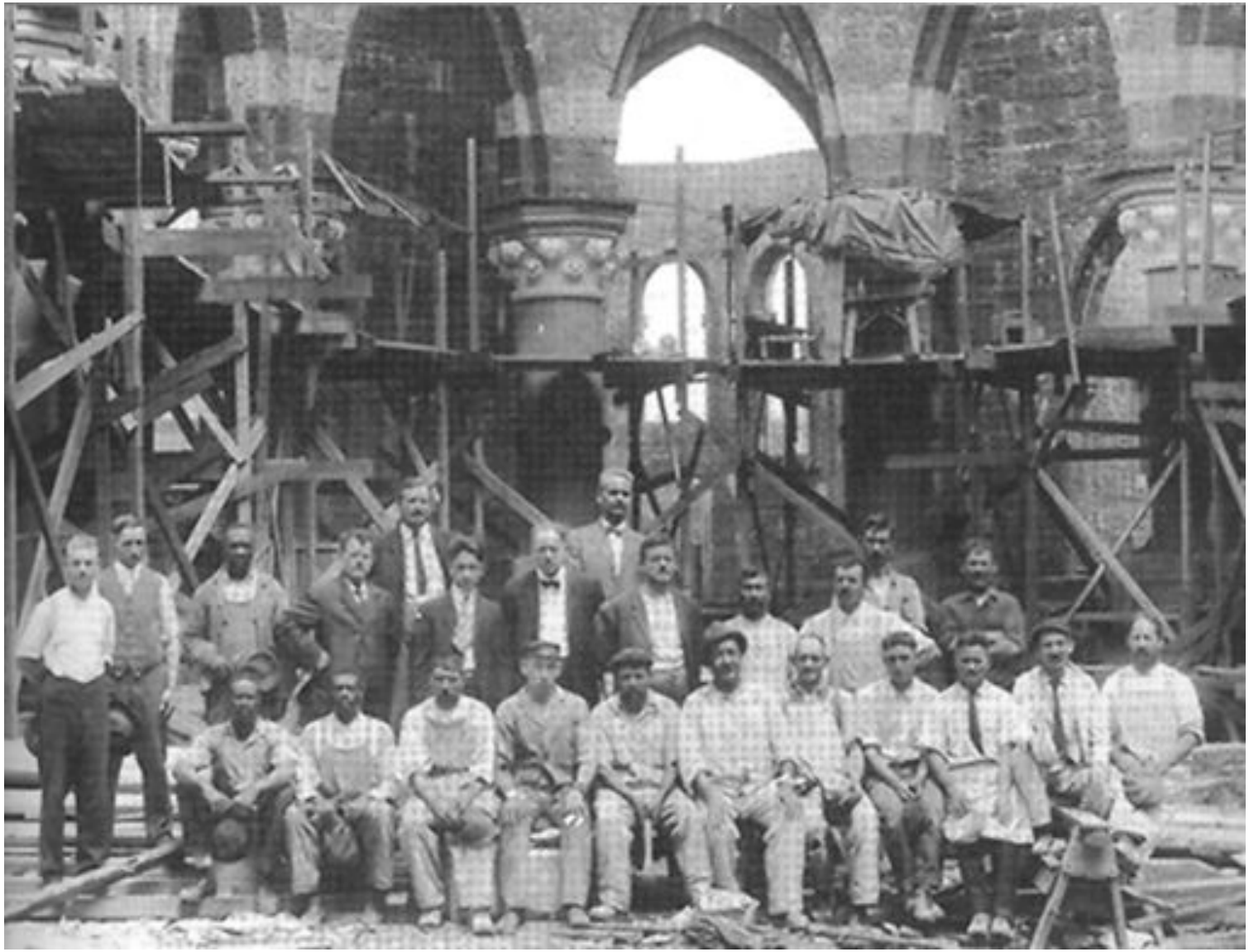
- Quale livello di assiomatizzazione per le ontologie create?
- Meglio un'unica ontologia (stile DBpedia ontology) o più ontologie fra loro collegate (modularizzazione)

Ontologie



1. Formalizzare un modello dell'universo di discorso di riferimento
2. Definire lo spazio delle ricerche possibili parte degli utenti

Siamo soltanto all'inizio…



Grazie!
www.DATIPIEMONTE.it

Vittorio Di Tomaso | ditomaso@celi.it

Contributi di Diego Magro (Università di Torino), Christian Morbidoni (SpazioDati)