Εργαστήριο Σημασιολογικού Ιστού

Ενότητα 10: Συνδεδεμένα Δεδομένα (Linked Data)

Μ.Στεφανιδάκης

16-5-2016

Συνδεδεμένα δεδομένα και Σημασιολογικός Ιστός

- Ο όρος “Συνδεδεμένα Δεδομένα” (Linked Data) είναι μεταγενέστερος του Σημασιολογικού Ιστού
  - Περιγράφει ένα σύνολο καλών πρακτικών για την οργάνωση των σημασιολογικών δεδομένων
- Τα Συνδεδεμένα Δεδομένα χρησιμοποιούν όλα τα πρότυπα και τις τεχνολογίες του Σημασιολογικού Ιστού
  - Κομβικό σημείο το μοντέλο οργάνωσης δεδομένων RDF
- Σε τέτοιο βαθμό, ώστε οι δύο όροι να συγχέονται!
  - Ποια είναι λοιπόν η διαφορά;

Διασυνδεδεμένα Δεδομένα (Linked Data)

Οι αρχικές “4 Εντολές” (Tim Berners-Lee, 2006):

1. Χρησιμοποιήστε URIs για να αναγνωρίσετε οντότητες
   - Εξασφαλίζοντας τη μοναδικότητα των αναγνωριστικών
2. Χρησιμοποιήστε HTTP URIs (δηλαδή URLs)
   - Για να μπορούν οι άλλοι να προσπελάσουν την πληροφορία σας
3. Στην προσπέλαση, απαντήστε με χρήσιμη πληροφορία σε πρότυπη μορφή (RDF*,SPARQL)
   - Δώστε δυνατότητα στις μηχανές να ρωτήσουν και να καταλάβουν την απάντηση
4. Διασυνδέστε τα αναγνωριστικά σας URI με άλλα, τρίτων
   - Έτσι ώστε αυτός που ρωτάει να βρεί και άλλες συμπληρωματικές πηγές πληροφόρησης

Η σχέση με τον Σημασιολογικό Ιστό

Τι ξέρουμε μέχρι τώρα:

1. Χρησιμοποιήστε URIs για να αναγνωρίσετε οντότητες
   - Ναι!
2. Χρησιμοποιήστε HTTP URIs (δηλαδή URLs)
   - Όχι υποχρεωτικά
3. Στην προσπέλαση, απαντήστε με χρήσιμη πληροφορία σε πρότυπη μορφή (RDF*,SPARQL)
   - Εν μέρει (αλλά, τι σημαίνει προσπέλαση;)
4. Διασυνδέστε τα αναγνωριστικά σας URI με άλλα, τρίτων
   - Παίρνει αξία με το 2

Διηγούμενοι μια ιστορία..

- Ή αλλιώς: χρησιμοποιώντας HTTP URIs
- Τα Συνδεδεμένα Δεδομένα επιβάλλουν τη χρήση διευθύνσεων στο Web ως URIs
  - Όταν προσπελάσετε ένα τέτοιο URI θα πάρετε πίσω χρήσιμη πληροφορία σχετική με την οντότητα που αντιπροσωπεύει το URI
  - Ο κάτοχος του URI (της διεύθυνσης στο Web δηλαδή) μπορεί να “διηγηθεί μια ιστορία” για την οντότητα του URI
    - Για κατανάλωση από τον άνθρωπο (π.χ. HTML, εικόνα PNG κλπ) ή από τη μηχανή (RDF)
- Προσοχή: Μέσω της προσπέλασης δεν λαμβάνετε την οντότητα, αλλά μια πληροφορία για αυτήν
  - Ο μηχανισμός θυμίζει έμμεση προσπέλαση μέσω δεικτών στη C (!)
  - Έτσι ονομάζεται παραστατικά URI dereferencing

Αποσαφήνιση: Μηχανισμός URI Dereferencing

ο “Οδοντωτός” (η φυσική οντότητα, όχι η φωτό!)

(σε μορφή αναγνώσιμη από τον άνθρωπο και τη μηχανή)

data document

πληροφορία για

αναφέρεται σε

προσπέλαση

<http://fr.dbpedia.org/resource/Ligne_de_Diakofto_à_Kalavryta>

(URI)

URI dereferencing: ένας νέος τρόπος προσπέλασης

- Μέχρι τώρα γνωρίζαμε έναν μόνο τρόπο για να προσπελάσουμε σημασιολογικά δεδομένα:
  - Ερωτήματα σε SPARQL endpoints
  - Η απάντηση, ανάλογα με το ερώτημα, είναι πίνακας (SELECT) ή γράφος (CONSTRUCT, DESCRIBE)
- Όμως πολύ περισσότερα sites υποστηρίζουν URI dereferencing
  - Συνήθως χωρίς να παρέχουν SPARQL endpoint!
- Συνεπώς, είναι σημαντικό οι εφαρμογές μας να μπορούν να εκμεταλλευτούν αυτόν τον τρόπο προσπέλασης
  - Η απάντηση είναι πάντα γράφος (τριάδες), σε διάφορες μορφές (μέσω διαπραγμάτευσης περιεχομένου)

Διαπραγμάτευση περιεχομένου

Μορφή	Ζητήστε
HTML	(default, για τον άνθρωπο)
RDF/XML	application/rdf+xml (υποστηρίζεται πάντα)
N3	text/rdf+n3
Turtle	application/x-turtle
JSON	application/rdf+json
JSON-LD	application/ld+json

Δοκιμάστε κι εσείς!

- Δοκιμάστε μέσω curl:
  - τη διεύθυνση `http://dbpedia.org/resource/Lodovico_Giustini`
  - την ίδια διεύθυνση με εναλλακτικό περιεχόμενο
    - σε μορφή application/rdf+xml
    - σε μορφή text/rdf+n3

```
curl -H 'Accept: text/rdf+n3'
    'http://dbpedia.org/resource/Lodovico_Giustini'
```

- Μια στιγμή! Δεν βλέπω τίποτα...
  - Προσθέστε το -L
  - Τι βλέπετε τώρα; Γιατί όμως συμβαίνει αυτό;

Οι άγγελοι, η καρφίτσα και το httpRange-14

- Όπως στον μεσαίωνα, όπου οι θεολόγοι διαφωνούσαν διαρκώς για το πόσοι άγγελοι χωράνε στη μύτη μιας καρφίτσας
- Παρόμοιο και το θέμα με το κρυπτικό όνομα httpRange-14
  - “Τι πρέπει να επιστρέφεται ως κωδικός κατά την προσπέλαση ενός URI;”
    - Οι ιστοσελίδες, για παράδειγμα, επιστρέφουν `200 OK`
    - Οι ιστοσελίδες όμως είναι πληροφοριακές πηγές (information resources)
    - Και ό,τι επιστρέφεται είναι μια αναπαράστασή τους (representation)
    - Με αυστηρή ερμηνεία, δεν πρέπει να επιστρέφεται `200 OK` για τα URIs μας...
    - ..γιατί αντιπροσωπεύουν μη πληροφοριακές πηγές (non-information resources: ανθρώπους, ιδέες, φυσικά αντικείμενα κ.ο.κ)
  - Τι επιστρέφεται λοιπόν κατά το URI dereferencing;
    - 303 See Other
    - Μην το πάρετε όμως πολύ σοβαρά!

HTTP Response: Status Codes

Οι πιο σημαντικοί κωδικοί:

200	OK
301	Moved Permanently
302	Found
303	See Other
304	Not Modified
400	Bad Request
403	Forbidden
404	Not Found
415	Unsupported Media Type
500	Internal Server Error
501	Not Implemented

303: Εκτεταμένη χρήση στα Συνδεδεμένα Δεδομένα

Ανακατευθύνσεις (Redirects)

Οι κωδικοί απόκρισης 3xx δηλώνουν ανακατεύθυνση

- Ο web client πρέπει να προσπελάσει κάτι άλλο!
- curl: πώς θα ακολουθήσει την ανακατεύθυνση

```
curl -L 'http://www.example.com/'
```

Δοκιμάστε κι εσείς!

- Για να δείτε την ανακατεύθυνση (όχι να την ακολουθήσετε!)

```
curl -s -i -H 'Accept: text/rdf+n3'
    'http://dbpedia.org/resource/Lodovico_Giustini'
```

- Ποιος ο κωδικός της απόκρισης HTTP (πρώτη γραμμή);
  - Πού σας ανακατευθύνει η επικεφαλίδα Location;
  - Δοκιμάστε να ζητήσετε και τις άλλες μορφές περιεχομένου
- Η ανακατεύθυνση 303 συνηθίζεται σε URIs σαν το προηγούμενο
  - `http://ex.com/path/to/resource/...`
  - “Slash (/) namespaces”

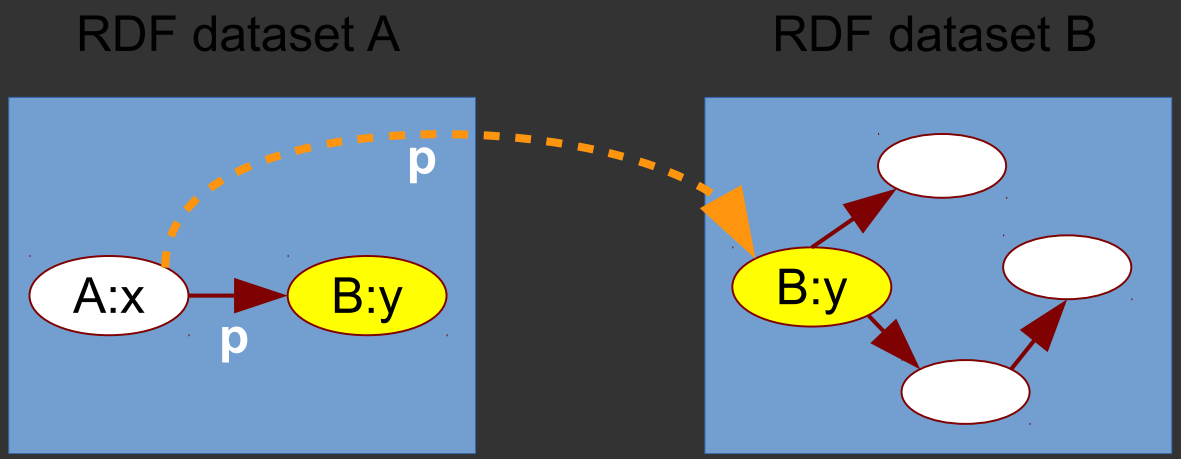
Hash (#) Namespaces

- Δοκιμάστε το εξής:

  ```
  curl -s -i  -H 'Accept: application/rdf+xml'
     'http://www.w3.org/2000/01/rdf-schema#label'
  ```

- Ποια η απόκριση; Υπάρχει ανακατεύθυνση;
- Χώροι ονομάτων με # χρησιμοποιούνται κυρίως για λεξιλόγια RDF
  - Εδώ η οντότητα είναι το `http://www.w3.org/2000/01/rdf-schema#label`
  - Σύφωνα με το HTTP η αίτηση θα γίνει στο `http://www.w3.org/2000/01/rdf-schema`
  - Η απόκριση είναι έγγραφο που περιέχει πληροφορία και για το ζητούμενο URI
  - Πού βρίσκεται το έγγραφο καθεαυτό; δείτε την επικεφαλίδα Content-Location

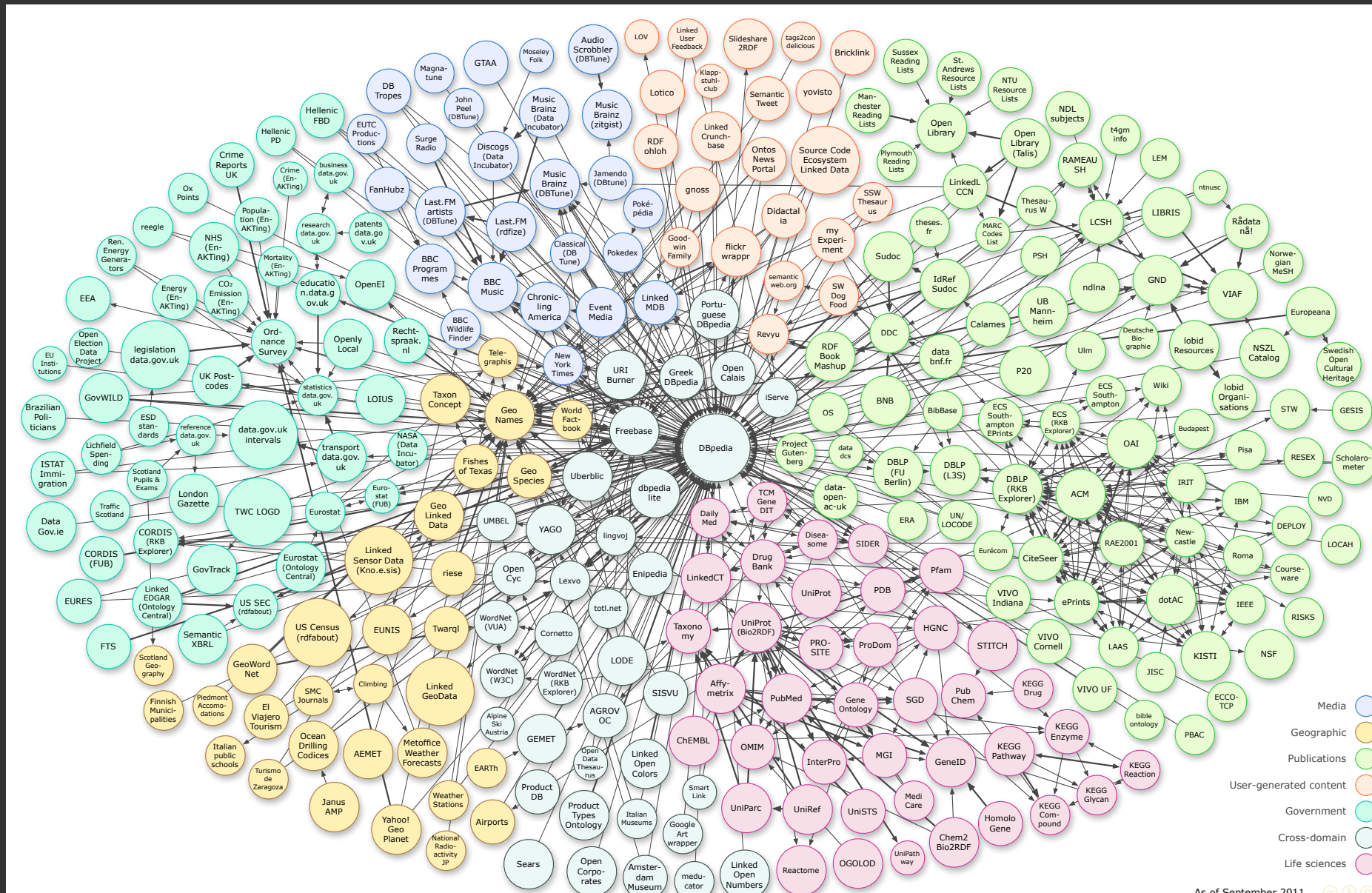
“Διασυνδέστε τα αναγνωριστικά σας URI με άλλα, τρίτων”



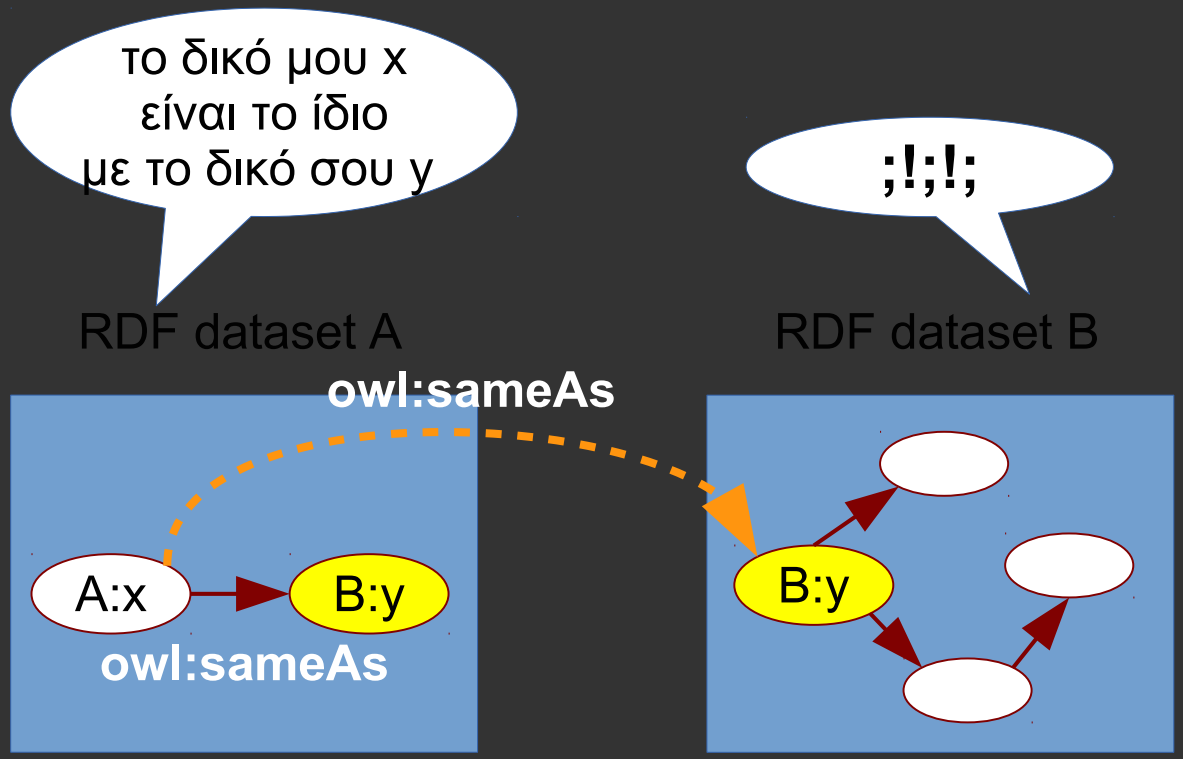
Όταν στο site (dataset) A μάθετε για το B:y, μπορείτε να μετακινηθείτε στο site B, και με προσπέλαση του B:y να μάθετε για νέες τριάδες

LOD cloud

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/



owl:sameAs



Από ένα ισχυρότερο του RDFS οντολογικό λεξιλόγιο (Web Ontology Language, OWL) το `owl:sameAs` εκφράζει ισχυρή ισοδυναμία οντοτήτων, χωρίς όμως να σημαίνει ότι είναι αποδεκτό από τον στοχευόμενο...

Η οικογένεια Bach

- Επισκεφτείτε τη διεύθυνση `http://d-nb.info/gnd/11850553X`
  - Δείτε επίσης την παρεχόμενη μορφή RDF
- Η εργασία σας:
  - Χρησιμοποιήστε Python και rdflib (βλ. παράδειγμα 1 και παράδειγμα 2)
  - Φορτώστε σε γράφο το παραπάνω URI
    - Βάλτε την parse() μέσα σε try...except για να προστατευτείτε απο 404 Not Found
  - Διαλέξτε το όνομα (ιδιότητα `gndo:preferredNameForThePerson`) της οντότητας
  - Ανακτήστε όλα τα URIs των οντοτήτων που συνδέεται οικογενειακά (ιδιότητα `gndo:familialRelationship`) και προσπελάστε τις κι αυτές με τον ίδιο τρόπο
    - Φροντίστε να μην επισκεφτείτε δύο φορές την ίδια οντότητα!
  - Τυπώστε στο τέλος τα ονόματα που ανακτήσατε
  - Μπορείτε να βρείτε τα επαγγέλματα του καθενός;

```
PREFIX gndo: <http://d-nb.info/standards/elementset/gnd#>
```