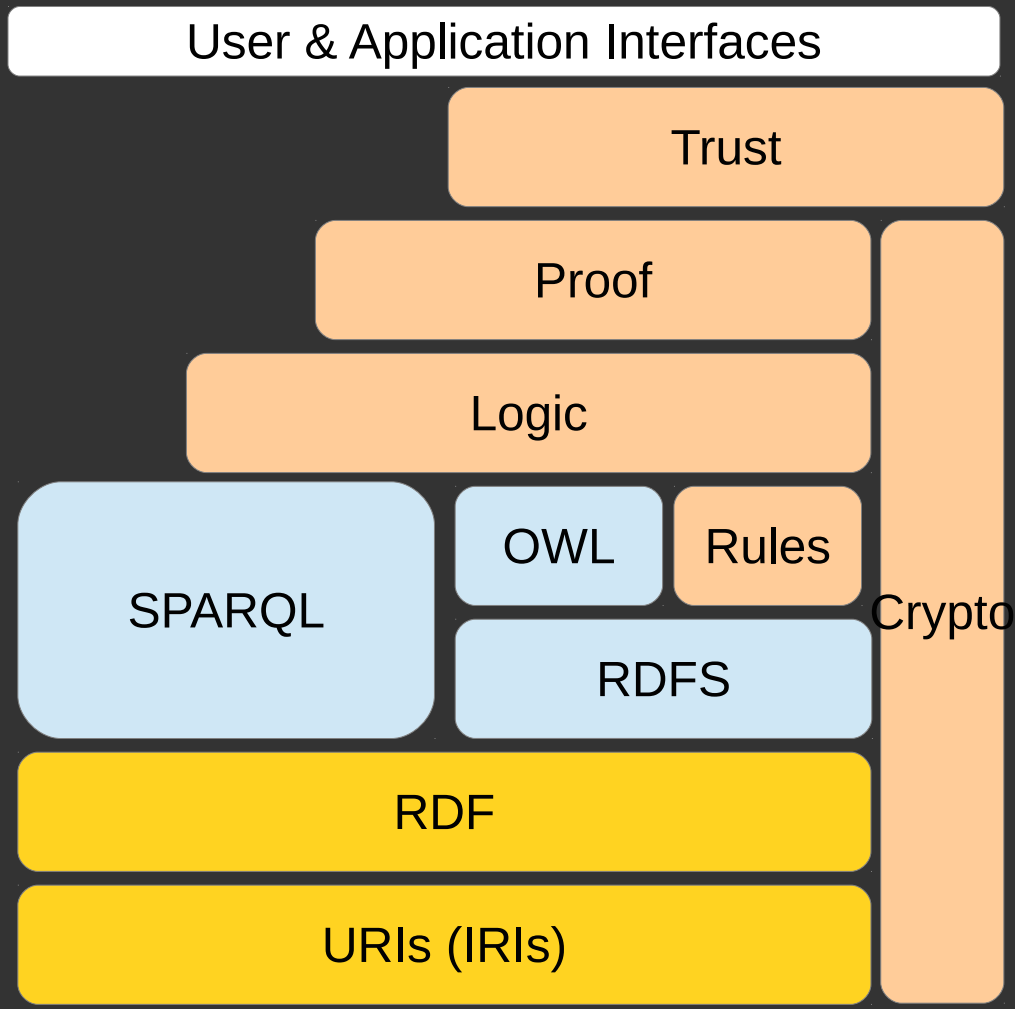
Εργαστήριο Σημασιολογικού Ιστού

Ενότητα 5: Resource Description Framework (RDF)

Μ.Στεφανιδάκης

26-3-2020

Τα επίπεδα του Σημασιολογικού Ιστού



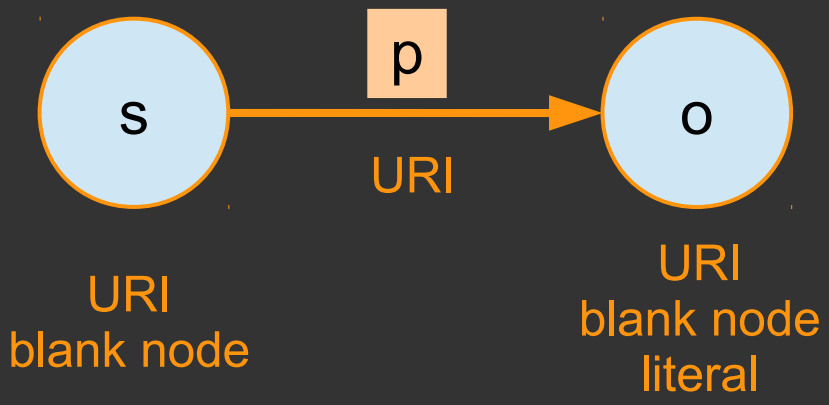
RDF: Το κύριο πρότυπο του Σημασιολογικού Ιστού, χρησιμοποιεί αναγνωριστικά URIs

Resource Description Framework (RDF)

- Βασικό πρότυπο του Σημασιολογικού Ιστού
- Αν και λέμε συχνά “η RDF” (υπονοώντας “η γλώσσα RDF”)...
- ...στην πραγματικότητα είναι ένα μοντέλο οργάνωσης γνώσης
  - Που επιτρέπει να κάνουμε δηλώσεις (statements)
  - σε μορφή τριάδων (triples) (που ανήκουν σε έναν γράφο (graph))
  - σχετικά με οντότητες (resources)
  - οι οποίες συμβολίζονται με URIs
- Από τα πρώτα πρότυπα του σημασιολογικού ιστού (2004), με μια πιο πρόσφατη επανέκδοση (RDF 1.1, 2014)

Μοντέλο δεδομένων κατά το πρότυπο RDF

- Η RDF προσδιορίζει ένα μοντέλο (abstract syntax) βασισμένο στις τριάδες, ακριβώς όπως τις έχουμε δει ως τώρα
  - (υποκείμενο s, κατηγόρημα p , αντικείμενο o)
  - ως μέρος γράφου με δύο κόμβους (s,o) και μία κατευθυνόμενη ακμή (από το s προς το o)
    - οι κόμβοι μπορούν να είναι URIs (IRIs), ανώνυμοι κόμβοι (blank nodes) ή σταθερές τιμές (literals)
- Το p δηλώνει μια ιδιότητα (property), μια διμερή δηλαδή σχέση (binary relation) μεταξύ s και o



URIs

- Τα URIs δρουν ως σφαιρικά αναγνωριστικά οντοτήτων
  - Ένα URI δεν πρέπει ποτέ να αναφέρεται σε περισσότερες από μία οντότητα
  - Ένα URI, άπαξ και δημιουργηθεί, δεν πρέπει ποτέ να αλλάξει οντότητα, στην οποία αναφέρεται
- Αν και δεν είναι υποχρεωτικό, ένα URI καλό θα ήταν να οδηγεί σε κάποιο έγγραφο στο web, με πληροφορία σχετική με την οντότητα του URI

Blank nodes και literals

- Οι ανώνυμοι κόμβοι (blank nodes) αναγνωρίζουν οντότητες χωρίς ρητό όνομα
  - απλά λένε ότι κάτι (ανώνυμο) έχει τις περιγραφόμενες σχέσεις
- Οι σταθερές τιμές literals εξ'ορισμού δεν αλλάζουν
  - Η RDF όμως τους προσδίδει κάτι πολύ σημαντικό: τύπο δεδομένων (datatype)

Τύποι δεδομένων (Datatypes)

- Συμβολίζονται με ένα URI, συνήθως της μορφής:
- http://www.w3.org/2001/XMLSchema#xxx
  - xxx είναι ο εκάστοτε τύπος δεδομένων
  - βασίζεται στο πρότυπο XML Schema
  - συντομογραφικά: xsd:xxx
- Η RDF περιγράφει μια σειρά συμβατών τύπων δεδομένων
  - xsd:string, xsd:boolean, xsd:integer, xsd:double, xsd:float,..
  - xsd:date, xsd:time, xsd:dateTime,..
  - κ.ο.κ..
- Η RDF χρησιμοποιεί επίσης το URI
  - http://www.w3.org/1999/02/22-rdf-syntax-ns#langString
  - για κείμενο με ένδειξη γλώσσας (π.χ. en, el, el-GR ..)

Literals και Datatypes

- Τι προσδίδει η σύνδεση ενός literal με έναν τύπο δεδομένων;
  - Προσδιορίζει τη μέθοδο χειρισμού της τιμής του literal
    - Πώς το κείμενο του literal (lexical form) θα μετατραπεί στην κατάλληλη τιμή
    - Η μετατροπή προσδιορίζεται από τον τύπο δεδομένων
- Παράδειγμα: ο τύπος xsd:boolean
  - Διαθέτει δύο τιμές (value space): `{true,false}`
  - Δέχεται τα εξής strings (lexical space): {“true”, “false”, “1”, “0”}
  - Μετατρέπει ως εξής (Lexical-to-value mapping): $< \text{“}true\text{”} \rightarrow true >, < \text{“}false\text{”} \rightarrow false >, < \text{“}1\text{”} \rightarrow true >, < \text{“}0\text{”} \rightarrow false >$

Πηγές RDF και συλλογές γράφων RDF

- RDF Source
  - Πηγή πληροφορίας RDF, περιέχει συλλογές γράφων RDF σε δεδομένη χρονική στιγμή
  - Οι γράφοι (και οι τριάδες) που περιέχει μπορούν να αλλάξουν με την πάροδο του χρόνου
- RDF Dataset
  - Μια συλλογή γράφων RDF, όπου
  - όλοι οι γράφοι εκτός από έναν αναγνωρίζονται με ένα URI (ή blank node) και ονομάζονται επώνυμοι γράφοι (named graphs)
  - Ο μοναδικός γράφος χωρίς σύνδεση με κάποιο URI είναι ο γράφος default
- Οι επώνυμοι γράφοι χρησιμοποιούνται για την τοποθέτηση των τριάδων σε υποσύνολα
  - Η χρήση προσδιορίζεται από την εκάστοτε εφαρμογή

Χώροι ονομάτων RDF

- Η RDF (και το συνοδευτικό RFD Schema που θα δούμε σε επόμενα) χρησιμοποιούν δικά τους (built-in) λεξιλόγια (vocabularies)
  - Για την “οντολογική” περιγραφή των διαφόρων οντοτήτων
  - Και για μια σειρά πρόσθετων βοηθητικών εννοιών (utilities)
- Οι χώροι ονομάτων για τα λεξιλόγια αυτά είναι
  - http://www.w3.org/1999/02/22-rdf-syntax-ns# (συντομογραφικό πρόθεμα rdf)
  - http://www.w3.org/2000/01/rdf-schema# (συντομογραφικό πρόθεμα rdfs)
- Παράδειγμα: `rdfs:label`
  - Χρησιμοποιείται για να συνδέσει μια ετικέτα αναγνώσιμη από τον άνθρωπο σε μια οντότητα
  - (http://ex.com/A, rdfs:label, ”Semantic Web”@en)

Αποθήκευση δεδομένων RDF

- Η RDF εκτός από το αφηρημένο μοντέλο οργάνωσης, περιγράφει και διάφορα μορφότυπα αποθήκευσης των τριάδων σε αρχεία κειμένου
- Το απλούστερο από τα μορφότυπα αυτά ονομάζεται N-Triples
  - Ξεκίνησε ως “η γλώσσα των παραδειγμάτων” της RDF
  - Αλλά πολύ γρήγορα χρησιμοποιήθηκε για μαζική ανταλλαγή δεδομένων RDF
  - Πολύ απλή επεξεργασία, δεν χρειάζονται εξειδικευμένες βιβλιοθήκες
  - Σήμερα υποστηρίζει Unicode χαρακτήρες (κωδικοποίηση utf-8)
    - Αρχικά, μόνο ASCII χαρακτήρες: όλοι οι άλλοι χρειάζονταν ειδική κωδικοποίηση

N-Triples: βασική σύνταξη

- Κάθε γραμμή του αρχείου κειμένου περιέχει ακριβώς μία τριάδα
  - Στη μορφή `s p o .` (κενά/tab μετά από κάθε ένα s,p,o, στη συνέχεια ακολουθεί τελεία και newline)
    - Στη συνιστώμενη κανονική μορφή: ακριβώς ένα κενό
- Τα URIs γράφονται μεταξύ < και >
  - `<http://ex.com/A>`
  - σε πλήρη μορφή, χωρίς συντομογραφικά προθέματα
- Τα literals γράφονται μεταξύ ” και ”
  - `"Semantic Web"`
  - Προαιρετικά ακολουθεί ο τύπος δεδομένων ή η γλώσσα, π.χ.

```
"Semantic Web"@en
"1.663E-4"^^<http://www.w3.org/2001/XMLSchema#double>
```

N-Triples: βασική σύνταξη (2)

- Οι ανώνυμοι κόμβοι (blank nodes) έχουν πρόθεμα _:
  - _:b1234
  - Μετά το _: ακολουθεί η ετικέτα του ανώνυμου κόμβου
    - Η τελεία δεν μπορεί να είναι στην αρχή ή το τέλος της ετικέτας
    - Το - δεν μπορεί να είναι στην αρχή της ετικέτας

Δοκιμάστε κι εσείς

- Ξεκινήστε από τα πιο πρόσφατα δεδομένα του ωρολογίου προγράμματος (σε μορφή csv τριάδων)
  - Όπως έχει διαμορφωθεί μετά την εισαγωγή των URIs σε θέση κατηγορήματος και αντικειμένου
- Κατασκευάστε νέο πρόγραμμα Python για να μετατρέψετε τα δεδομένα σας σε έγκυρο αρχείο N-Triples
  - Ακολουθήστε τις οδηγίες σύνταξης του προτύπου N-Triples στις προηγούμενες διαφάνειες
  - Θα χρησιμοποιήσετε τύπο δεδομένων xsd:time (hh:mm:ss) για την ώρα
    - http://www.w3.org/2001/XMLSchema#time
    - Χρησιμοποιήστε ώρα έναρξης και λήξης