Εργαστήριο Σημασιολογικού Ιστού

Ενότητα 4: Χρησιμοποιώντας Ενιαία Αναγνωριστικά – URIs και IRIs

Μ.Στεφανιδάκης

8-3-2018

Η έννοια της οντότητας

- Στον Σημασιολογικό Ιστό οι τριάδες μπορούν να εκληφθούν ως σύνολο δηλώσεων (statements)
  - Π.χ., *Για τα Α, Β και Γ ισχύουν τα Χ και Υ*
  - Τι μπορούν να είναι τα Α, Β, Γ, Χ και Υ;
  - Οτιδήποτε! Πόροι (ιστοσελίδες και άλλα αρχεία), πράγματα, άνθρωποι, έννοιες, συναισθήματα...
- Στον Σημασιολογικό Ιστό κάνουμε δηλώσεις σχετικά με οντότητες (entities)
  - Απαιτούνται αναγνωριστικά ονόματα (identifiers)
  - Που αναφέρονται σε κάθε τέτοια οντότητα
  - Τα ονόματα αυτά θα χρησιμοποιηθούν στις τριάδες (δηλώσεις)

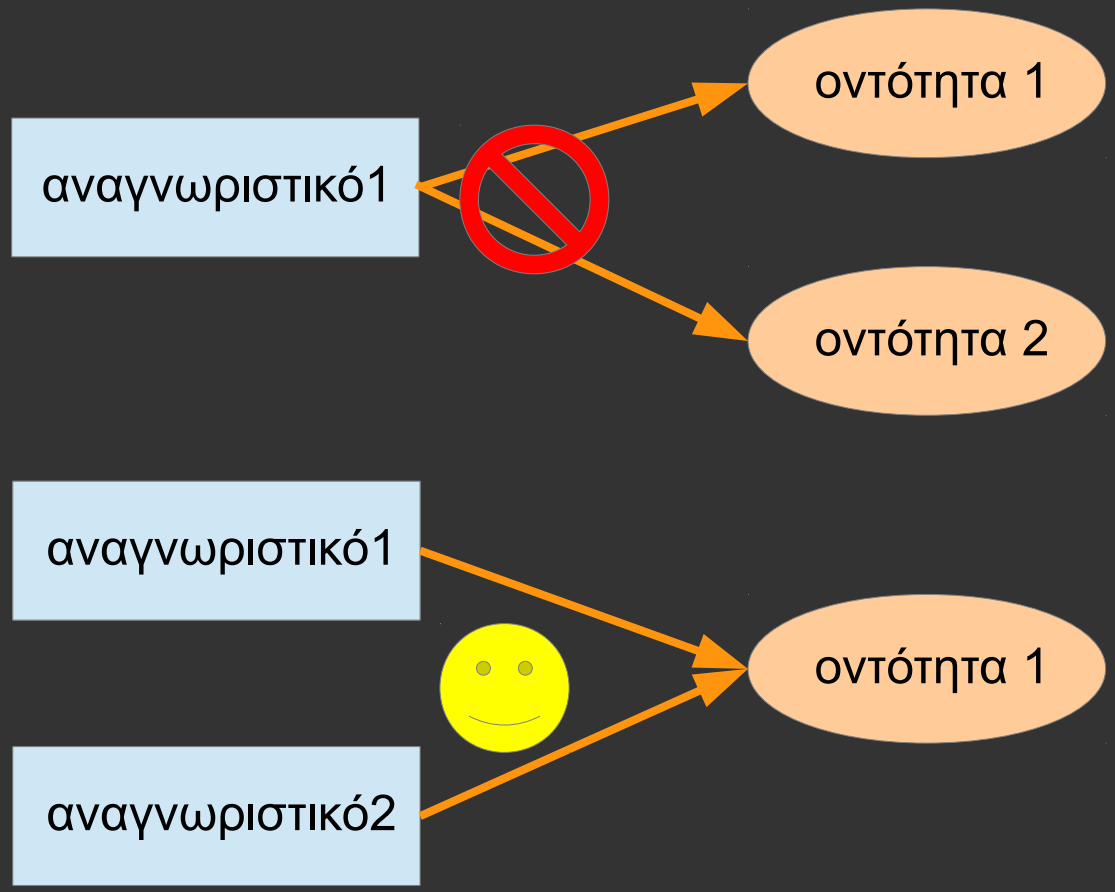
Αναγνωριστικά Οντοτήτων

- Στον Σημασιολογικό Ιστό θέλουμε να συνδυάζουμε δεδομένα από πολλαπλές πηγές και παραγωγούς
  - Συνεπώς, το ζητούμενο είναι η σφαιρική αναγνώριση των οντοτήτων
    - Ένα αναγνωριστικό να υποδηλώνει μία και μόνο οντότητα, σε παγκόσμιο επίπεδο
- Στα παραδείγματά μας μέχρι τώρα
  - Χρησιμοποιούμε τοπικά αναγνωριστικά
  - Όμως: ένα αναγνωριστικό όπως π.χ. sem_web μπορεί να χρησιμοποιείται από τρίτους για τελείως διαφορετική οντότητα!
- Η (πικρή) αλήθεια: τα αναγνωριστικά στα παραδείγματά μας μέχρι τώρα αξίζουν όσο και οι ονομασίες των ανώνυμων κόμβων..
  - Ήρθε η ώρα να στηριχτούμε σε πρότυπα!

Σημείωση: Αμφισημία και Ταυτοσημία

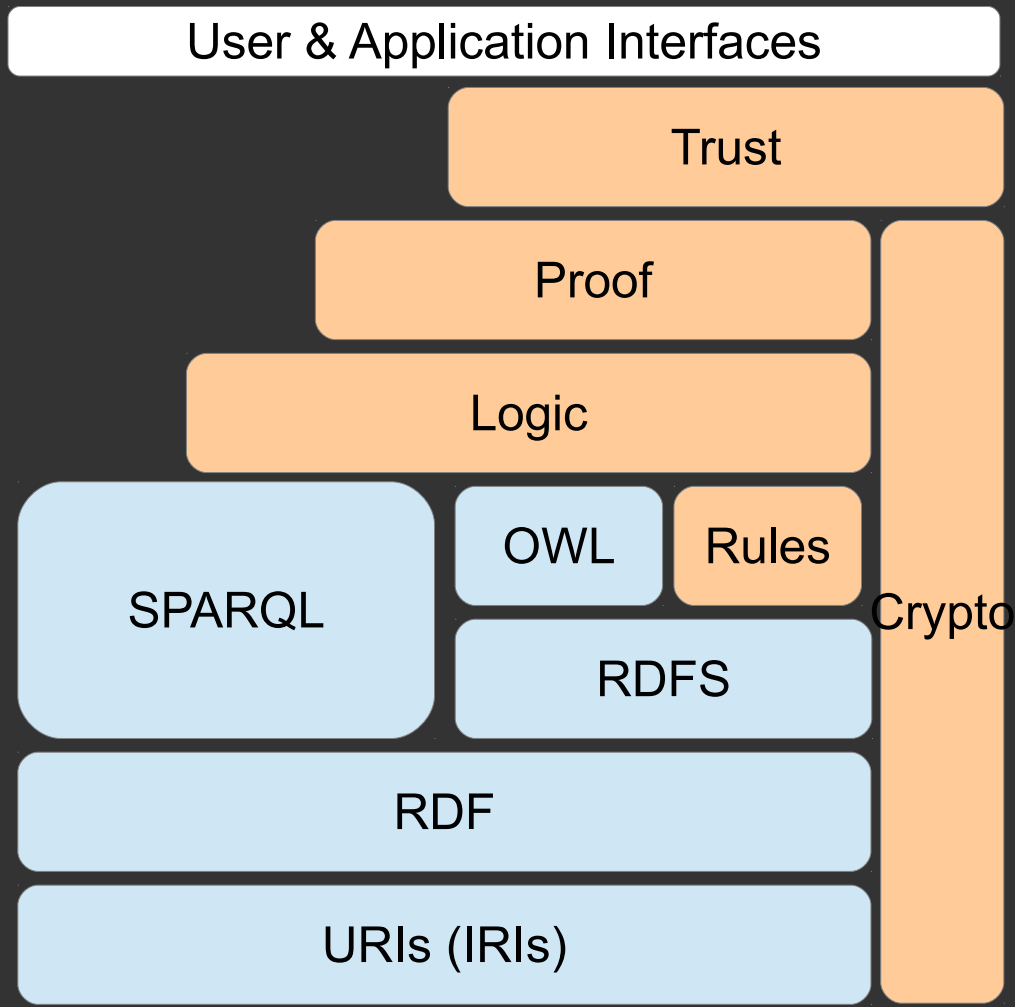
- Επιτρέπεται ένα αναγνωριστικό να αναφέρεται σε δύο διαφορετικές οντότητες;
  - ΟΧΙ!!! Ένα αναγνωριστικό προσδιορίζει μοναδικά μια οντότητα
- Δύο διαφορετικά αναγνωριστικά μπορούν να αναφέρονται στην ίδια οντότητα;
  - ΝΑΙ!!! Αυτό είναι απόλυτα επιτρεπτό
  - π.χ. τα `http://dbpedia.org/resource/Mount_Olympus` και `http://sws.geonames.org/734890/` αναφέρονται στην ίδια οντότητα (το βουνό Όλυμπο)
- Πρακτικά: στο δικό μας σετ δεδομένων καλό είναι να χρησιμοποιούμε μόνο ένα αναγνωριστικό για την ίδια οντότητα

Αμφισημία και Ταυτοσημία σχηματικά



Η δεύτερη περίπτωση είναι νόμιμη!

Τα επίπεδα του Σημασιολογικού Ιστού

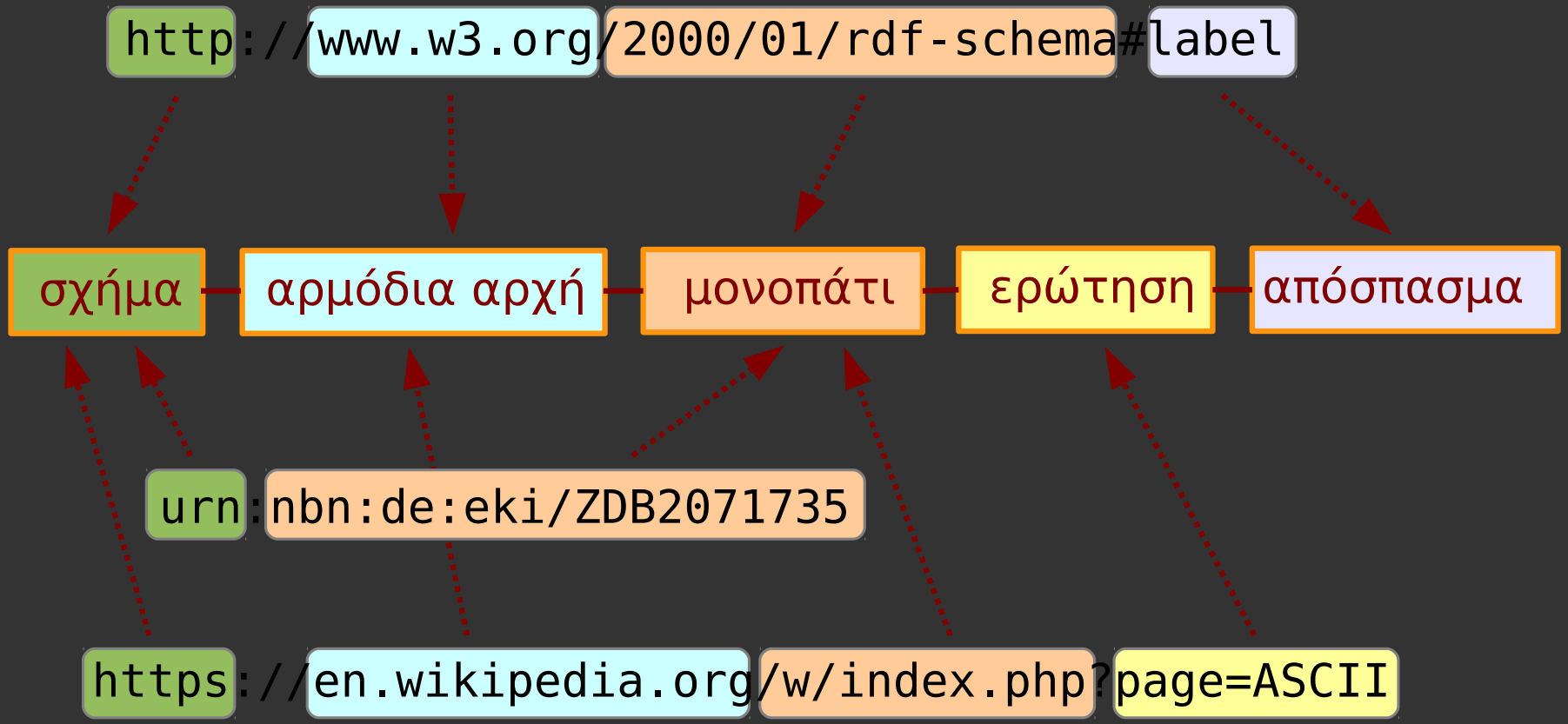


Και τα πρότυπα των χαμηλότερων επιπέδων, καθιερωμένα από τον οργανισμό W3C (World Wide Web Consortium)

Uniform Resource Identifiers (URIs)

- URI: ένα συμπαγές string με καλά ορισμένους κανόνες σύνταξης
  - που αναγνωρίζει μια οποιαδήποτε οντότητα (εδώ ονομάζεται “resource”)
  - Μονοσήμαντα και με παγκόσμια ισχύ
  - Η οντότητα μπορεί να είναι οτιδήποτε: όχι μόνο μια πληροφοριακή πηγή (όπως ένα έγγραφο ή μια ιστοσελίδα) αλλά και άνθρωπος, πράγμα, έννοια, συναίσθημα, κλπ
- Προσοχή: το URI δεν είναι η οντότητα αλλά αναφέρεται στην οντότητα

Γενική σύνταξη URIs



Κάθε υποκατηγορία URI μπορεί να έχει τη δική της υπο-μορφή

Μορφές URIs

- Μια μεγάλη κατηγορία URIs μοιάζουν με διευθύνσεις στο Web (URLs)
  - π.χ. http://dbpedia.org/resource/Lodovico_Giustini
- Μια δεύτερη κατηγορία URIs έχει τελείως διαφορετικό σχήμα:
  - urn:uuid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6
  - urn:nbn:de:eki/ZDB2071735
  - URN:ISBN:978-82-8140026-9
    - Αυτά ονομάζονται Uniform Resource Names (URNs)
- Και οι δύο τύποι απαιτούν μια εκδούσα αρχή για την μονοσήμαντη και σφαιρική ανάθεση μέρους του URI

Μοιάζουν αλλά δεν είναι URLs

- Στον κλασσικό Σημασιολογικό Ιστό το URI:

`http://dbpedia.org/resource/Lodovico_Giustini`

- απλά αναγνωρίζει την οντότητα “Lodovico Giustini” (το πρόσωπο)
- και επιτρέπει να κάνουμε δηλώσεις σχετικές με την οντότητα αυτή
  - όταν το URI εμφανίζεται ως υποκείμενο ή αντικείμενο σε τριάδες
- Δεν απαιτείται η ύπαρξη πληροφορίας σ’αυτή τη διεύθυνση!
  - Το αντίθετο μάλιστα: δεν πρέπει να υποθέσουμε κάτι τέτοιο
  - Πόσο πρακτικό μπορεί να είναι αυτό;
    - Το θέμα θα μας απασχολήσει αργότερα, στα πλαίσια των Συνδεδεμένων Δεδομένων – Linked Data

Τι ισχύει για τα URNs;

- Η αρχική ιδέα ήταν η κατασκευή ιδανικών URIs
  - Με αποσύνδεση της τοποθεσίας από το αναγνωριστικό
  - Σε αντίθεση με τα “μη καθαρά” http://.. URIs
    - που, αν και δεν είναι υποχρεωτικό, συχνά “μπερδεύουν” την αναγνώριση μιας οντότητας με την προσπέλαση της σχετικής πληροφορίας
- Στην πράξη αποδείχτηκε ότι τα URNs είναι δύσχρηστα
  - Η προσπέλαση σχετικής πληροφορίας είναι κυρίαρχη στην εποχή του Web!
  - Με τα URNs η προσπέλαση είναι αδύνατη:
    - πώς ξέρουμε ότι για το urn:nbn:de:eki/DNB991052625 θα πάρουμε πληροφορία...
    - ...από το `http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&IKT=8132&TRM=DNB991052625`;
- Σήμερα, όλο και περισσότερο χρησιμοποιούνται http URIs στη θέση τους

URIs και IRIs

- Σύμφωνα με τη αρχική σύνταξη των URIs [RFC3986] οι επιτρεπόμενοι χαρακτήρες σε ένα URI ανήκουν στο 7-bit ASCII (απλοί λατινικοί χαρακτήρες)
- Οι πιο κάτω χαρακτήρες είναι δεσμευμένοι και πρέπει να κωδικοποιούνται με %xx

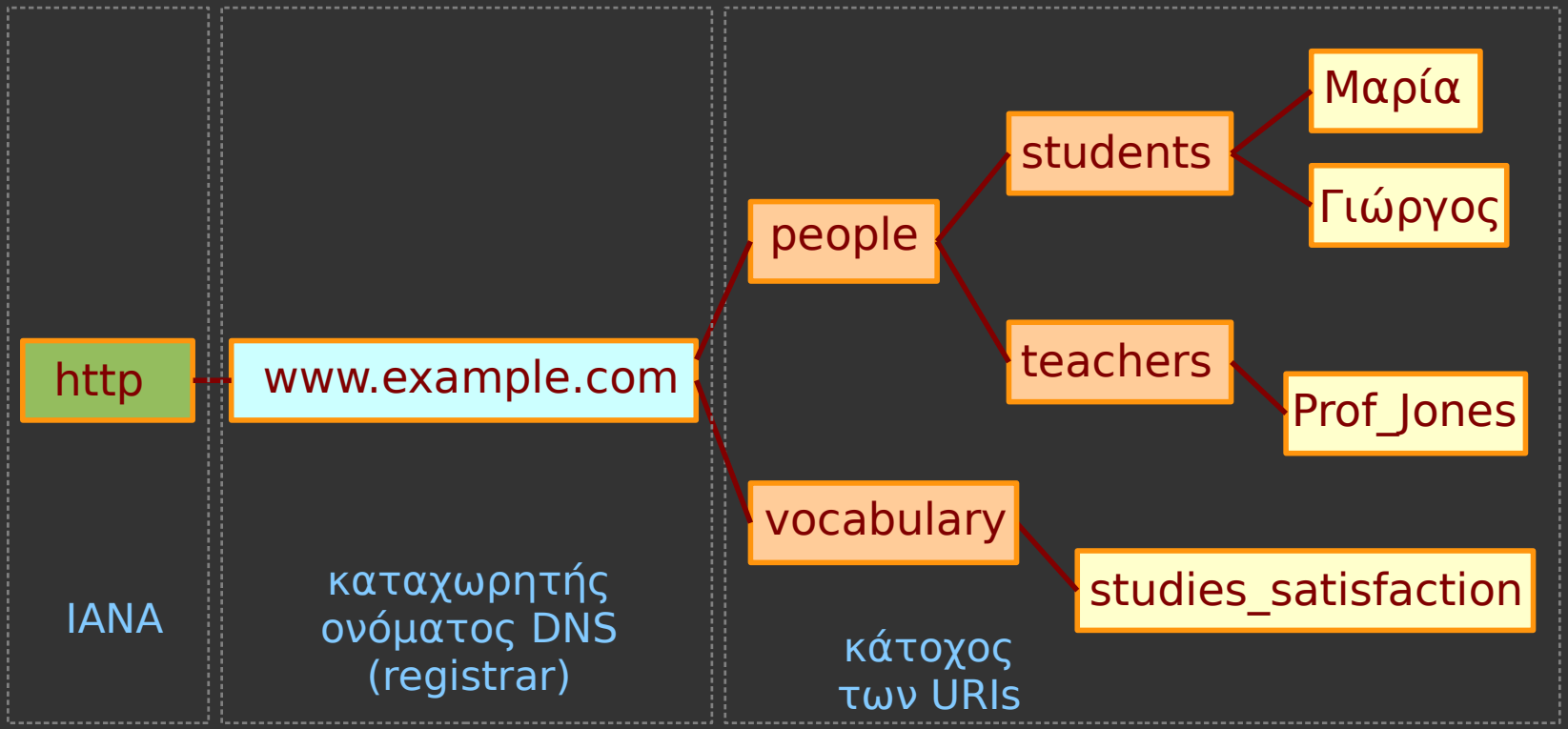
```
: / ? # [ ] @ ! $ & ' ( ) * + , ; =
```

- Επίσης τα http URIs πρέπει να κωδικοποιήσουν και τους χαρακτήρες

```
< > " space { } | \ ^ `
```

- Τα σύγχρονα πρότυπα του Σημασιολογικού Ιστού χρησιμοποιούν τον όρο Internationalized Resource Identifiers (IRIs, [RFC3987]), όπου επιτρέπεται κάθε χαρακτήρας Unicode (εκτός των δεσμευμένων)
  - Σε επόμενα χρησιμοποιούμε ισοδύναμα τους όρους URI και IRI

Ποιος διαχειρίζεται τα URIs;



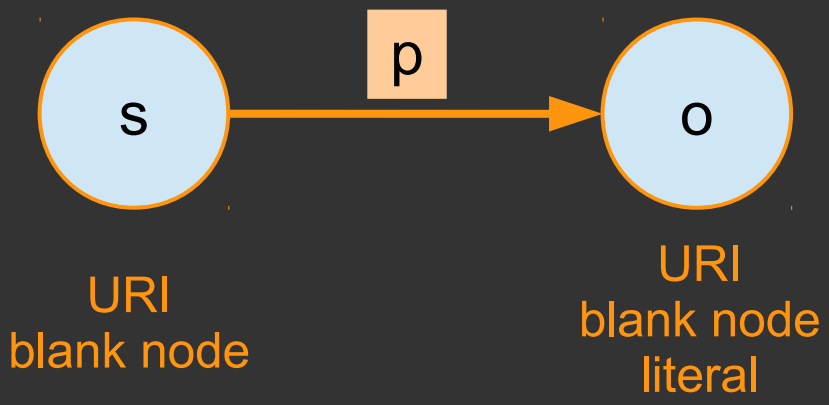
Η ιεραρχική διαχείριση εγγυάται τη μονοσήμαντη αναγνώριση!

Χώροι ονομάτων (Namespaces)

- Κάθε URI χωρίζεται σε δύο μέρη:
  - Πρόθεμα (prefix) που ορίζει τον ”χώρο διευθύνσεων” για κάθε ομάδα URIs
  - Τοπικό μέρος (local part), αναφέρεται στην οντότητα καθαυτή
  - http://ex.com/resource/ entityA
    - ”slash (/) namespace”
  - http://ex.com/vocab# termX
    - ”hash (#) namespace”
  - ”hash” και ”slash” URIs: ισοδύναμα ως προς τη χρήση ως αναγνωριστικά
    - Τα πράγματα αλλάζουν όταν χρησιμοποιούνται και για προσπέλαση (θα το δούμε αργότερα)

Χρήση των URIs στις τριάδες (s,p,o)

- Αντικαθιστώντας τα “ασθενή” αναγνωριστικά σε subject και object
- Με “ισχυρά” αναγνωριστικά URIs
- Η έννοια που μεταδίδεται είναι σαφέστερη από πριν
  - Τι άλλο μπορεί να γίνει;
  - Τι συμβαίνει με το predicate της τριάδας;

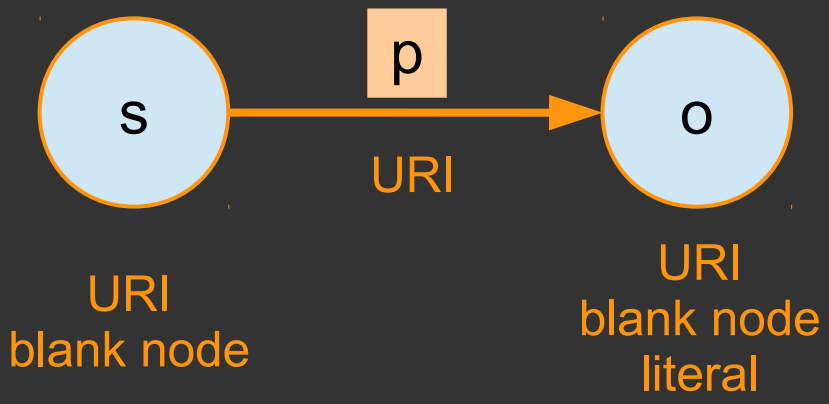


Λεξιλόγια (Vocabularies): Εισαγωγή

- Στον Σημασιολογικό Ιστό, τα λεξιλόγια είναι ομάδες URIs, σε έναν κοινό χώρο ονομάτων, για την περιγραφή όρων σχετικών με το εκάστοτε πεδίο εφαρμογής
  - Παράδειγμα: έστω (φανταστικό) λεξιλόγιο για την περιγραφή εργασιακών σχέσεων, το οποίο περιέχει τα URIs:
    - http://ex.com/evocab#Employee
    - http://ex.com/evocab#Employer
    - http://ex.com/evocab#salary
    - http://ex.com/evocab#worksAt
    - κ.ο.κ.
- Τα URIs ενός λεξιλογίου χρησιμοποιούνται σε μεγάλο βαθμό (αλλά όχι μόνον) ως predicates των τριάδων

Ευρέως γνωστά λεξιλόγια – Γιατί;

- Η χρήση ευρέως γνωστών λεξιλογίων στα σημασιολογικά δεδομένα
  - Επιτρέπει την κατασκευή έξυπνων εφαρμογών που μπορούν να “κατανοήσουν” τη σημασία των δεδομένων
    - και των σχέσεων μεταξύ δεδομένων
- Ένα γνωστό λεξιλόγιο δρα ως κοινός σημασιολογικός παρονομαστής
  - Έτσι, πριν φτιάξουμε το δικό μας, πρέπει να αναζητήσουμε ήδη υπάρχοντα λεξιλόγια!



Τώρα και τα predicates είναι URIs!

Η σειρά σας!

- Ξεκινήστε από τα αρχικά δεδομένα του ωρολογίου προγράμματος (σε μορφή πίνακα)
  - Μετατρέψτε όλα τα αναγνωριστικά που θα γίνουν URIs
    - Αντικαταστήστε όλους τους μη επιτρεπόμενους χαρακτήρες
    - Θυμηθείτε: δουλεύουμε με IRIs
- Στη συνέχεια, τροποποιήστε το πρόγραμμα που κατασκευάζει το csv των τριάδων
  - Μετατρέψτε τα αναγνωριστικά των s και o σε URIs
    - namespace http://host/sw/you/resource/
  - Μετατρέψτε τα αναγνωριστικά των p σε URIs
    - Φτιάξτε το δικό σας λεξιλόγιο (προσοχή: σε πραγματικές συνθήκες αυτό δεν συνιστάται!)
    - namespace http://host/sw/you/myvocab#

host = η διεύθυνση θα δοθεί στο εργαστήριο

you = το login name σας στο Ιόνιο (π.χ. p03jdoe)