

# Εξαγωγή σχεσιακών πληροφοριών από τη Βικιπαίδεια

Χρήστος Περιβολαρόπουλος

Τετάρτη 8 Ιουνίου 2016

Η εργασία αυτή έγινε υπο την επίβλεψη των κ. Κυριάκο Σγάρμπα του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών Πανεπιστημίου Πατρών και κ. Boris Katz απο το InfoLab του CSAIL MIT.

# Δομή

Το οικοσύστημα του START

WikipediaBase

Wikipedia Mirror

Επίλογος

# Ενας ωκεανός πληροφορίας

*As of 2014 Google has indexed 200 Terabytes (TB) of data.*

— <http://www.websitemagazine.com/>

*The English Wikipedia is now one of 292 Wikipedia editions and holds the largest amount of articles, with more than 5,167,046,*

— [wikipedia.org](http://wikipedia.org)

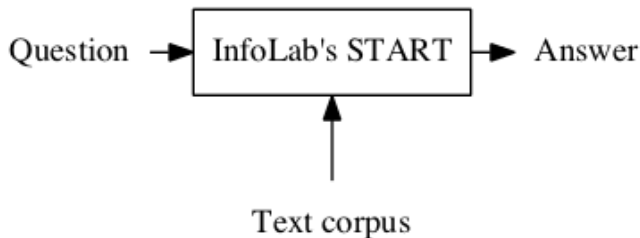
# Εμπόδια στη χρήση της πληροφορίας

- ▶ Αδομητη πληροφορία κατανοήσιμη μονο απο ανθρώπους — πχ. ελεύθερο κείμενο, διαγράμματα, video
- ▶ Προϋπόθεση από το χρήστη να γνωρίζει τη δομή της πληροφορίας
- ▶ Προϋπόθεση από το χρήστη να γνωρίζει ειδικά εργαλεία και/ή τα μαθηματικά μοντέλα.

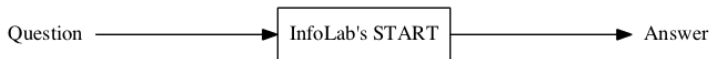
# InfoLab's START

Πρόσβαση σε πολύπλοκες πληροφορίες απαντώντας σε ερωτήσεις σε φυσική γλώσσα.

# Βασική λειτουργία του START.



# Ανάγκη πρόσβασης στο διαδίκτυο.

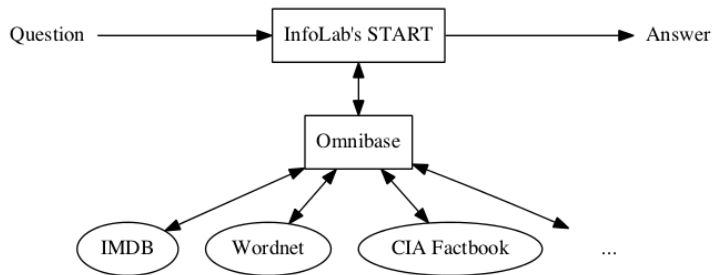


?

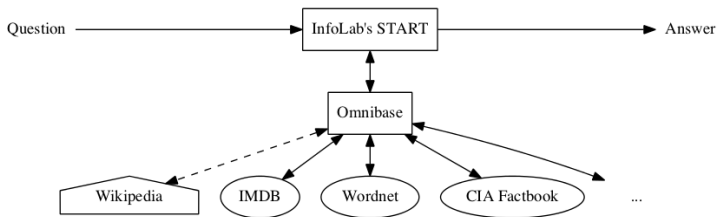




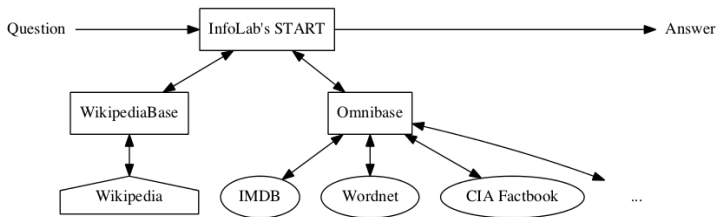
# Omnibase



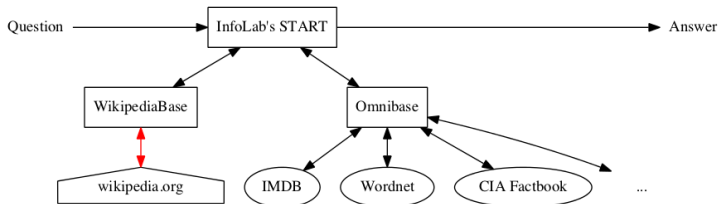
# Η wikipedia χρήζει ιδιαίτερης προσοχής



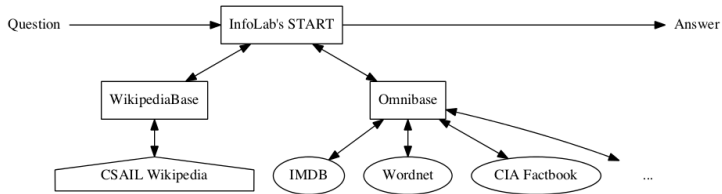
# WikipediaBase



# Έντονη χρήση του wikipedia.org



# Wikipedia Mirror



# Πληροφορίες που διαχειρίζεται κυρίως το WikipediaBase

- ▶ Χαρακτηριστικά οντοτήτων  
(get "wikibase-person" "Barack Obama" (:ID "BIRTH-DATE"))  
=> ((:yyymmdd 19610804))
- ▶ Κατηγορίες/κλάσεις οντοτήτων  
> (get-classes "Hillary Rodham Clinton")  
("wikibase-term" "wikipedia-paragraphs"  
"wikibase-person" "wikipedia-officeholder"  
"wikipedia-person")
- ▶ Συνώνυμα

# Infobox

## Gerhard Gentzen



Gerhard Gentzen in Prague, 1945.

<b>Born</b>	November 24, 1909 <a href="#">Greifswald, Germany</a>
<b>Died</b>	August 4, 1945 (aged 35) <a href="#">Prague, Czechoslovakia</a>
<b>Nationality</b>	<a href="#">German</a>
<b>Fields</b>	<a href="#">Mathematics</a>
<b>Alma mater</b>	<a href="#">University of Göttingen</a>
<b>Doctoral advisor</b>	<a href="#">Paul Bernays</a>

# Infobox

Λαμβάνουμε:

- ▶ Χαρακτηριστικά από τον πίνακα του infobox
- ▶ Την κατηγορία της οντότητας από τον τύπο του infobox.

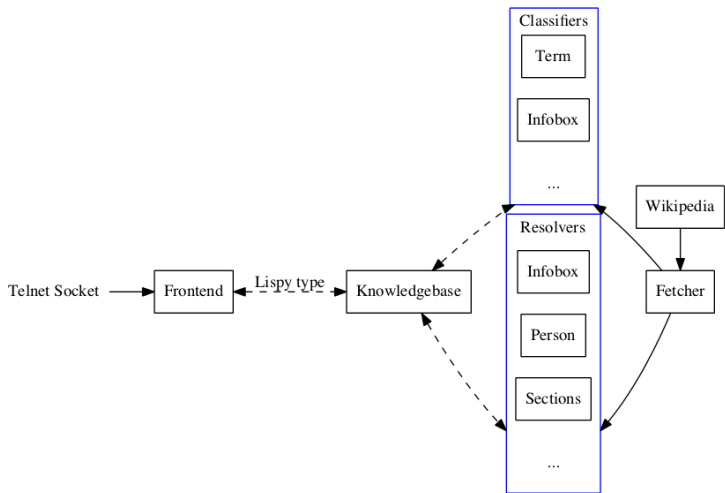


# Ιεραρχία των infoboxes

Προέρχεται από τη σελίδα List of infoboxes. πχ.  
Template:Infobox officeholder είναι υπο-κλάση του  
Template:Infobox person και έτσι:

```
> (get-classes "Hillary Rodham Clinton")  
("wikibase-term" "wikipedia-paragraphs"  
 "wikibase-person" "wikipedia-officeholder"  
 "wikipedia-person")
```

# WikipediaBase data pipeline



# Resolvers

- ▶ Infobox — Χαρακτηριστικά απο το infobox
- ▶ Person — Αν το άρθρο αναφέρεται σε άνθρωπο, το γένος του, η ημ. γέννησης κτλ
- ▶ Sections — Αυτούσιο κείμενο του άρθρου
- ▶ Term — συντεταγμένες, εικόνες, αριθμό, κύρια ονόματα, περίληψη άρθρου, URL και αριθμός λέξεων
- ▶ Error — Ένδειξη ότι δεν βρέθηκε το χαρακτηριστικό.

# Classifiers

- ▶ Term — πάντα wikipedia-term
- ▶ Infobox — απο τον τύπο του infobox
- ▶ Person — Ευρετικές για το αν το άρθρο αναφέρεται σε πρόσωπο.

# Provider/Acquirer model

Διαχωρισμός της λογικής που σχετίζεται με το χειρισμό δεδομένων από τη λογική που καθορίζει τις πηγές της.

$$\bigcup_{p \in \text{providers}} \{o : o \in \text{provided}(p)\}$$

ή dict πρόσβαση

$$\bigcup_{p \in \text{providers}} \{(k, v) : (k, v) \in \text{provided}(p)\}$$

# Συνώνυμα

Περισσότερα από ένα σύμβολα μπορεί αν αντιστοιχούν στην ίδια οντότητα.

Μερικά σύμβολα αντιστοιχούν σε άρθρα της wikipedia αλλά όχι σε οντότητες χρήσιμες στο START.

# Παραδείγματα συνωνύμων

- ▶ Raven (Journal)  $\equiv$  Raven
- ▶ Russian language/Russian alphabet  $\equiv$  Russian alphabet  $\equiv$  Russian language
- ▶ Obama  $\equiv$  Barack Obama
- ▶ Beatles  $\equiv$  The Beales

# Παραδείγματα μη αποδεκτών συνωνύμων

- ▶ File:Venn0001.svg
- ▶ List of infoboxes
- ▶ Alexander\_Pushkin#Legacy
- ▶ Abraham House  $\equiv$  A. House  $\not\equiv$  House



# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007

# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
 $((30, 9, 2006), (18, 7, 2007))$

# Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))
```

► tag:month, tag:fullname

# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric

# Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))
```

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric
- ▶ tag:year, tag:4digit

# Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))
```

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric
- ▶ tag:year, tag:4digit
- ▶ {day,numeric} of {month,fullname} {tag:year}

# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric
- ▶ tag:year, tag:4digit
- ▶ {day,numeric} Of {month,fullname} {tag:year}

# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric
- ▶ tag:year, tag:4digit
- ▶ {day,numeric} Of {month,fullname} {tag:year}
- ▶ {day,number}.{month,number}.{year}



# Αναγνώριση χρονικών διαστημάτων

> I am on the 12th bus,  
I will be here from 30th of september 2006 to  
18.7.2007  
((30, 9, 2006), (18, 7, 2007))

- ▶ tag:month, tag:fullname
- ▶ tag:day, tag:numeric
- ▶ tag:year, tag:4digit
- ▶ {day,numeric} Of {month,fullname} {tag:year}
- ▶ {day,number}.{month,number}.{year}
- ▶ {date} **to** {date}

# Τι είναι το Wikipedia Mirror

Ένα πρόγραμμα που παράγει κλώνους της wikipedia που τρέχουν σε ένα τοπικό μηχάνημα.

# Διαδικασία

# Διαδικασία

- ▶ Στήσιμο του server stack

# Διαδικασία

- ▶ Στήσιμο του server stack
- ▶ Εγκατάσταση του Mediawiki

# Διαδικασία

- ▶ Στήσιμο του server stack
- ▶ Εγκατάσταση του Mediawiki
- ▶ Κατέβασμα και αποκωδικοποίηση των wikipedia dumps

# Διαδικασία

- ▶ Στήσιμο του server stack
- ▶ Εγκατάσταση του Mediawiki
- ▶ Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- ▶ Φόρτωση των dumps στη wikipedia

# Διαδικασία

- ▶ Στήσιμο του server stack
- ▶ Εγκατάσταση του Mediawiki
- ▶ Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- ▶ Φόρτωση των dumps στη wikipedia
- ▶ Ρύθμιση του mediawiki να μιμείται τη wikipedia



# Διαδικασία

- ▶ Στήσιμο του server stack
- ▶ Εγκατάσταση του Mediawiki
- ▶ Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- ▶ Φόρτωση των dumps στη wikipedia
- ▶ Ρύθμιση του mediawiki να μιμείται τη wikipedia
- ▶ Ρύθμιση της βάσης δεδομένων για επιδόσεις συγκρίσιμες με wikipedia.org

# Mediawiki Stack

- ▶ Wikipedia configuration, database (MySQL) etc
- ▶ Mediawiki
- ▶ PHP
- ▶ Apache
- ▶ Linux

# Mediawiki Stack

- ▶ Wikipedia configuration, **database (MySQL)** etc
- ▶ **Mediawiki**
- ▶ **PHP**
- ▶ **Apache**
- ▶ **Linux**

**Bitnami!**

# Περιεχόμενα βάσης δεδομένων

Μηνιαία snapshots ολόκληρης της βάσης (εκτός απ' τα extensions και τους χρήστες).

`https://dumps.wikimedia.org/enwiki/latest/`

# Mwdumper

Μετατροπή απο XML σε MySQL: **mwdumper**

# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων  
απο την XML library (xerces).

# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- ▶ Γνωστό bug άλλα δύσκολα αναπαραξιμο.

# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- ▶ Γνωστό bug άλλα δύσκολα αναπαραξιμο.
- ▶ Μονό του το προβληματικό άρθρο δεν έσπαγε.



# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- ▶ Γνωστό bug άλλα δύσκολα αναπαραξιμο.
- ▶ Μονό του το προβληματικό άρθρο δεν έσπαγε.
- ▶ Καλυμμένο με κενά έσπαγε.

# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- ▶ Γνωστό bug άλλα δύσκολα αναπαραξιμο.
- ▶ Μονό του το προβληματικό άρθρο δεν έσπαγε.
- ▶ Καλυμμένο με κενά έσπαγε.
- ▶ Αφαιρετέο από το XML αρχείο δούλευε.

# xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- ▶ Γνωστό bug άλλα δύσκολα αναπαράξιμο.
- ▶ Μονό του το προβληματικό άρθρο δεν έσπαγε.
- ▶ Καλυμμένο με κενά έσπαγε.
- ▶ Αφαιρετέο από το XML αρχείο  
δούλευε.Αυτοματοποιήσουμε αυτή τη διαδικασία.

# Επιδόσεις

- ▶ **CPU:** Xeon E5-1607 3GHz 4-Core 64 bit
- ▶ **Main memory:** 64G
- ▶ **HDD:** (spinning disk) 500GB + 2Tb

Επιδόσεις δημιουργίας της βάσης (ανάλογα τη φορά):

- ▶ Προεπεξεργασία των dumps 10 λεπτά
- ▶ Δημιουργία της βάσης 10 ώρες

# Runtime Επιδόσεις

Απαγορευτικές για την ομαλή λειτουργία του START. Για το άρθρο του Barack Obama:

- ▶ Χωρίς βελτιστοποιήσεις 10s κατ ευθείαν στη βάση
- ▶ Με κάποιες βελτιστοποιήσεις 7s κατ ευθείαν στη βάση

# Ρυθμίσεις MySQL

Βελτιστοποιήσεις στη βάση που επιχειρήθηκαν:

- ▶ `innodb_buffer_pool_size`  $\in$  (16M, 5GB)
- ▶ Απενεργοποίηση της `fsync`
- ▶ `innodb_io_capacity` ήταν μεγαλύτερη από το bandwidth του δίσκου.
- ▶ `SET AUTOCOMMIT = 0; SET FOREIGN_KEY_CHECKS=0;` κατά τη διάρκεια των dumps.

# Ευχαριστίες

Ευχαριστώ τους καθηγητές μου κ Σγάρμπα, κ Katz και κ Φακωτάκη.



# Mandatory Dijkstra quote

*The question of whether Machines Can Think... is about as relevant as the question of whether Submarines Can Swim.*

— *Edsger W. Dijkstra*

# Links



<https://github.com/infolab-csail/WikipediaBase>



<https://github.com/infolab-csail/wikipedia-mirror>



[https://github.com/fakedrake/overlay\\_parse](https://github.com/fakedrake/overlay_parse)



<https://bitnami.com/>



<https://en.wikipedia.org/>



<https://dev.mysql.com/doc/refman/5.5/en/innodb-para>