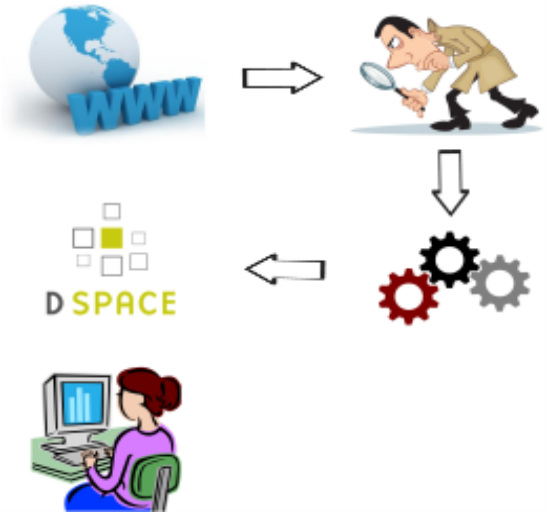
A NATIONAL DIGITAL LIBRARY OF SCIENTIFIC PUBLICATIONS USING DSPACE

Kostis Karozos, Giorgos Stoilos,
Giorgos Batistatos, Vasilis Vassalos

Department of Informatics,
Athens University of Economics and Business

Motivation

- The Greek National Documentation Center (NDC)
  - Collects and preserves digital data produced by the Greek scientific communities.
  - Automate the process.

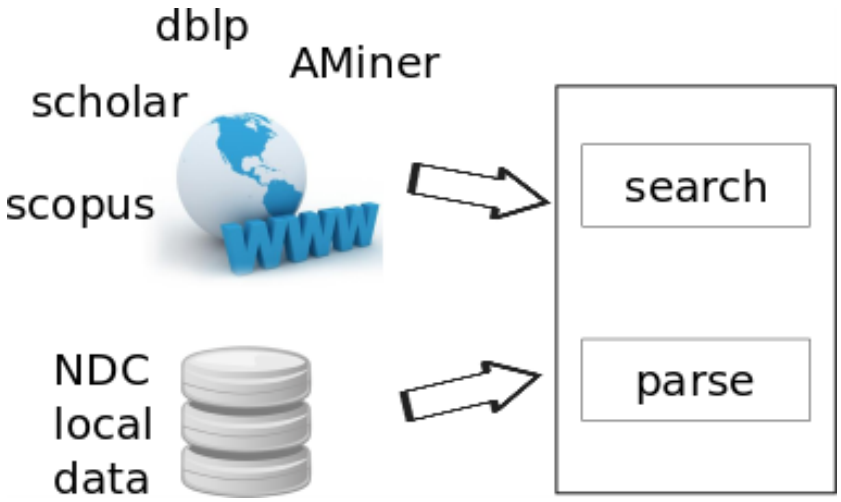


Challenges:

1) Collect publications
2) Check their "Greekness"
3) Process them (deduplicate, …)
4) Store them (DSpace)
5) Provide search services

Data Collection

- Is a hard problem
  - Find sources, analyse them, spot relevant information, parse data.
- We considered various sources
  - Web DBs/Pub. Sites
  - NDC local data
    - CDs, databases



Web Search Approach

- How do you start the search: two approaches
  - Seed of 3000 known Greeks (provided by NDC)
    - Search author's papers
    - Check co-authors: if Greek append them to list
  - Seed of known Greek Affiliations (provided by NDC)
    - Search based on institutions
- Used Jsoup API or web APIs if they are offered (e.g., scopus)

“Greekness” Detection Model

- We used three criteria:
  - first name (c_{fn}), last name (c_{ln}), and affiliation (c_{af})

au : author, w_1, w_2, w_3 : weights

$$greek(au) := w_1 \times c_{fn} + w_2 \times c_{ln} + w_3 \times c_{af}$$

- Maximise recall:
  - $thresh = 0{,}33$ and $w_i = 0{,}33$
  - If for some $c_* = 1$, then au is Greek

Criteria in Detail

- First Name (c_{fn}):
  - List with typical Greek first names (wiktionary): Γιώργος, Κώστας
  - Translated to Latin using ISO 843.
  - Not all Greeks follow an ISO ☺:

    Κώστας → Kostas/Costas Γιώργος → Giorgos/Yorgos/George
  - Alternative Writings: κ → κ/c, ϕ → f/ph, …
  - Soft Greekness: widely used Orthodox names (Russia) Sofia, Maria
    - If first-name is soft-greek, then $c_{fn} = 0{,}30$
    - Soft Greekness is not enough to pass the threshold

Criteria in Detail

- Last Name (c_{ln})
  - List from opengeodata.gr translated using alternatives
  - Again distinction strong vs weak greek last-names
    - Strong: typical Greek endings “os”, “is”, “ous”
    - Soft: all the rest which obtain a score of $c_{ln} = 0{,}7$
  - Soft first & soft last names pass threshold.
- Affiliation (c_{af})
  - List provided by NDC
  - Boolean criterion: $c_{ln} \in \{0,1\}$

Performance of Greekness Detection

- Evaluation using only first two criteria c_{ln} and c_{fn}:
  - Sample of 100 names
  - Compared against a system that uses n-grams

	greek(au)	n-grams
Recall	1,00	0,80
Precision	0,82	0,95
F1	0,89	0,88

- Evaluation of whole model
  - Sample of 1000 papers

	greek(au)
Recall	0,994
Precision	0,997
F1	0,985

Data Collection Results

- Run searching and Greekness model for 2 Months.
- Data Collected
  - Altogether: 1.1m records
  - Discarding identical: 580K
- Greek Scientists
  - Domestic: 167K
  - Abroad: 38K
  - Scientists with Greek affiliation but non-Greek name: 10K

Conjecture: collected about 65% of all Greek papers

Deduplication

- The same data coming from different sources.
  - Even though domain is controlled data are very often not exactly the same: s_1: vol="200", s_2: vol="In Press"
- Statistical methods for deduplication:
  - Python library Dedupe: Dedup./Entity Resolution
    - Affine gap similarity between fields
    - Classifier: logistics regression
- Trained using 5K records
- Evaluated on 106K records
- 580K → 480K

Precision	98.4%
Recall	99.9%

Record Storage

- We used DSpace
  - Used the Simple Archive Format (SAF)
    - Translated csv to saf.
  - Extended schema of Dspace
    - Some meta-data we extract are not in DC

    e.g., journal volume, paper pages, doi
    - In total we have 33 fields
- Various other customisations
  - Indexing over some of these fields

```
archive_directory/
    item_000/
        dublin_core.xml
        metadata_[prefix].xml
        contents
        file_1.doc
        file_2.pdf
    item_001/
        dublin_core.xml
        contents
        file_1.png
        ...
```

Search Services & Ext. Functionality

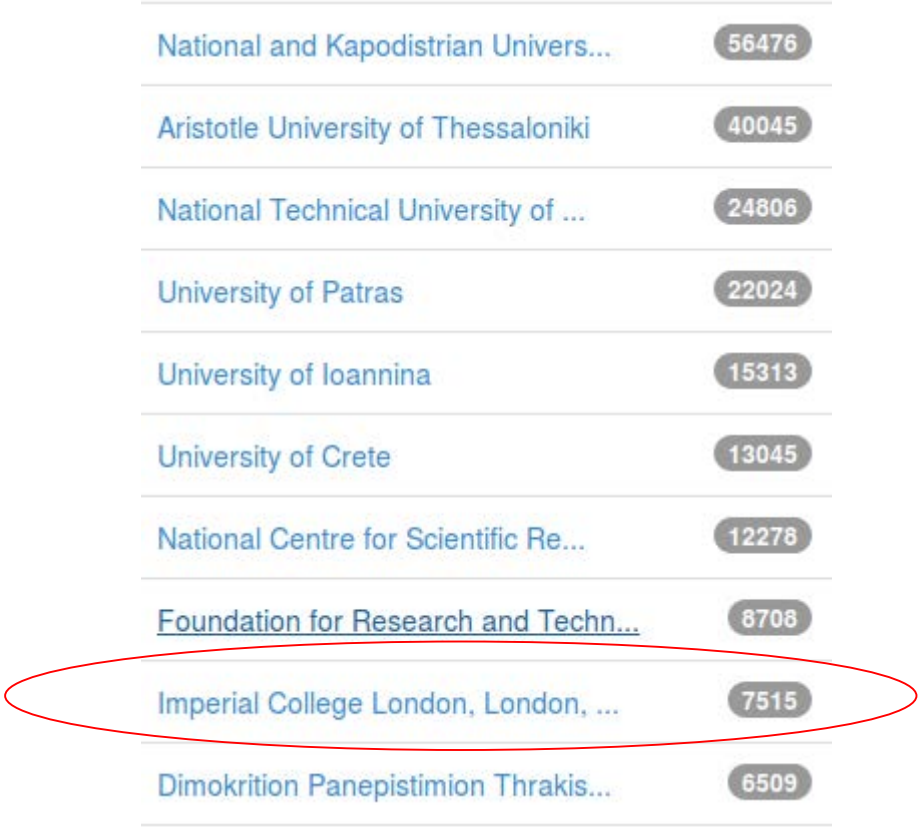
- DSpace out-of-the-box search services
  - Browsing indexed fields (authors, affiliations, ...)
  - Keyword search in meta-data
    - title:“Jupiter”
    - authorTag:<greek_abroad> (instructed DSpace to index)
  - Full-text & advanced searches (solr)
    - Wildcards: immunol*, Fuzzy searches: metric~0.8
- Extensions and additions
  - Extended .jsp to include new features.
  - Implemented additional search services.

Papers from 1900

Issue Date	Title	Author(s)
1904	On the stability of the volume by some organic liquids during the coagulation.	*Sigalas C; Sigalas C*
1904	Quantitative determination of phosphor in solutions	*Christomanos AC; Christomanos AC*
1904	Regarding the rotary power of normal and anti-toxic serums.	*Ferre G; Sigalas C; Ferre G; Sigalas C*
1904	New method for the representation of phosphor tribromide	*Christomanos AC; Christomanos AC*
1904	Description of phosphorous tribromide [Preliminary announcement]	*Christomanos AC; Christomanos AC*
1905	Concerning the solubility of phosphorous in ether and benzene	*Christomanos AC; Christomanos AC*
1905	STATE LIBRARY ADMINISTRATION	*Gillis J. L.; Gillis JL*
1906	IV. Beitrag zur kataraktbildung nach elektrischem schlag	*Bistis J.; Bistis J.; BISTIS J*
1907	A rare case of echinoccus of N. opticus.	*Papaioannou T; Papaioannou T*
1907	The knowledge of neuro keratine.	*Argiris A; Argiris A*

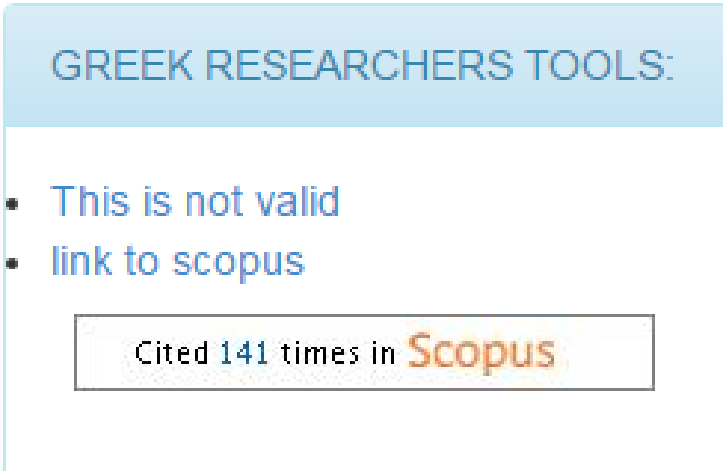
Universities with most publications

- Top four as expected
- Amongst the top 10 is the Imperial College!



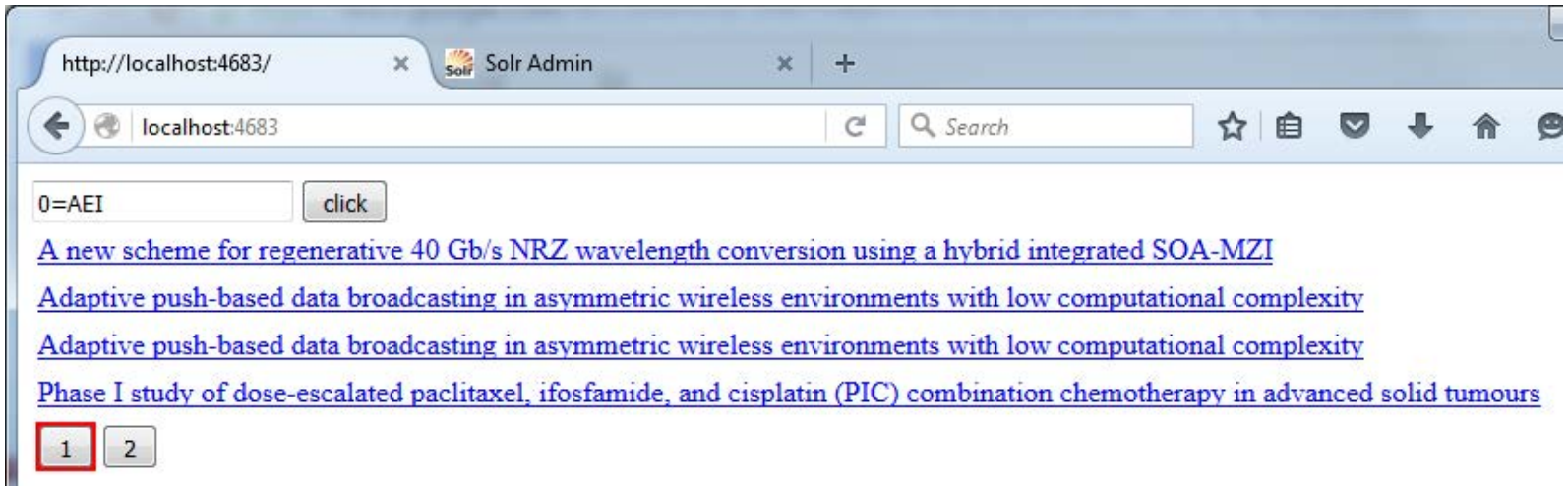
Additional Features fitted into DSpace

- Mechanism to mark that none of the paper authors is Greek (black-listing)
- Links to original records found on the Web.
- Current citation
  - This is dynamic



Additional Search Services

- Search according to type of affiliation
  - aff-types={higher-inst. (AEI), technical-inst. (TEI), hospital, ...}



- Retrieve all affiliations of an author (not possible in Dspace)

Thank You!

Questions