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 we 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 15K known Greek Affs (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 ($c_{af}$)
  - $p$: paper, $w_1, w_2$: weights

$$greek(p) := \max(w_1 \times c_{fn} + w_2 \times c_{ln}, c_{af})$$

- Maximise recall:
  - $thresh = 0.33$ and $w_i = 0.33$
  - If for some $c = 1$, then $p$ is relevant
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 opengeodeta.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

- Evaluation of whole model
  - Sample of 1000 papers

<table>
<thead>
<tr>
<th></th>
<th>greek(au)</th>
<th>n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>1,00</td>
<td>0,80</td>
</tr>
<tr>
<td>Precision</td>
<td>0,82</td>
<td>0,95</td>
</tr>
<tr>
<td>F1</td>
<td>0,89</td>
<td>0,88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>greek(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0,994</td>
</tr>
<tr>
<td>Precision</td>
<td>0,997</td>
</tr>
<tr>
<td>F1</td>
<td>0,985</td>
</tr>
</tbody>
</table>
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: $s_1: \text{vol} = \text{“200”}$, $s_2: \text{vol} = \text{“In Press”}$

- 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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>98.4%</td>
</tr>
<tr>
<td>Recall</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

580K $\rightarrow$ 480K
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
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

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

- Top four as expected

- Amongst the top 10 is the Imperial College!

<table>
<thead>
<tr>
<th>University Name</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>National and Kapodistrian University</td>
<td>56476</td>
</tr>
<tr>
<td>Aristotle University of Thessaloniki</td>
<td>40045</td>
</tr>
<tr>
<td>National Technical University of Pireas</td>
<td>24806</td>
</tr>
<tr>
<td>University of Patras</td>
<td>22024</td>
</tr>
<tr>
<td>University of Ioannina</td>
<td>15313</td>
</tr>
<tr>
<td>University of Crete</td>
<td>13045</td>
</tr>
<tr>
<td>National Centre for Scientific Research</td>
<td>12278</td>
</tr>
<tr>
<td>Foundation for Research and Technology</td>
<td>8708</td>
</tr>
<tr>
<td>Imperial College London, London, ...</td>
<td>7515</td>
</tr>
<tr>
<td>Dimokrition Panepistimion Thrakiou</td>
<td>6509</td>
</tr>
</tbody>
</table>
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

GREEK RESEARCHERS TOOLS:

- This is not valid
- Link to Scopus

Cited 141 times in Scopus
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