Archives Portal Europe. A new approach to the European Archival Heritage

by

Kerstin Arnold

presented by

Zoltán Szátucsek
Bogdan-Florin Popovici
Archives Portal Europe - Start and status
How it started

“Part two [...] deal[s] with institutional, technical and professional aspects of access to archives. Particular emphasis is placed on [...] finding aids and archival description; access on line and new research tools; setting up an Internet Gateway/Portal to documents and archives in Europe and cooperation with European networking projects in this field [...]”

Report on Archives in the Enlarged European Union, February 2005
What happened so far
Where we are at right now

- **Languages**: all official EU languages, plus Georgian, Hebrew, Icelandic
- **Countries**: 36 incl. overseas territories
- **Alphabets**: ქართული, Ελληνικά, български, עברית
- **Institutions**: 1,100+ actively contribute with content
- **Types of institutions**: National archives, City archives, Private archives, Museum archives, Church archives, University archives, Business archives etc....
- **Collections**: 600k+ Currently the largest online archival repository in the world
What the portal presents

- General information about more than 7,100 institutions
- 600,000+ Holdings guides
- 27,000 descriptions of records creators
- Collection level or item level with and without digital objects
- Source guides

Collections
Archival standards in Archives Portal Europe
Application of international XML standards

- **apeEAD and apeEAC-CPF**
  - Subsets of the international standards EAD 2002 and EAC-CPF
  - Tailored towards use in Archives Portal Europe
  - Based on use in partner institutions
Application of international XML standards

● apeMETS and apeMETSRights
  ○ Variation of the international standard METS
  ○ Created for the use in Archives Portal Europe

● EAG 2012
  ○ Developed during the APEnet and APEx projects
  ○ Based on its initial implementation in the Censo-Guía de Archivos de España e Iberoamérica
Continuous development

- APE’s profiles might change in order to
  - Enable new functionalities, e.g. when including licence and rights information
  - Allow for more precision in display, e.g. when distinguishing different types of digital objects
- Changes are either optional or will be introduced by automated and general batch processing
Enhancing retrieval
Challenges of a multilingual and multicultural research space
Multicultural and multi-institutional

30+ countries

Variations of institution types
Multilingual content

20+ languages

Multilingual interface

Multilingual search
Content remains in language of origin
User interface in variety of languages

The history of Europe - one click away

Archives Portal Europe provides you a single access to find, browse and discover information on archives about Europe held by thousands of cultural heritage institutions from more than 30 countries. We continually grow and expand our content, adding more countries, more institutions, and more archival collections.

Start your search

Type in your keywords and discover over 550,000 archival collections including hundreds of millions of documents, the persons and organisations that created and used these documents, and the institutions holding the archival material today.

Selected Language

English

Official translation

Automatic translation
Searching with multilingual terms

Start your search

Type in your keywords and discover over 550,000 archival collections including hundreds of millions of documents, the persons and organisations that created and used these documents, and the institutions holding the archival material today.

France Frankreich Frankrijk Франция Галія Francia Frankrig

Show advanced search options  ▼

Show only materials with digital objects  ▼

Search each term separately  ▼
No archive is an island - or is it?
Bridging the gap: multilingual access
Options

- Let the user try...
- Enhancing retrievability to bridge the gap = controlled access terms
Controlling access terms in archives...

...is difficult because:

- archival description tradition
- uncontrolled vocabulary
- random position in description
- static text, lack of connection to linked data
Solutions

A. Connecting existing topics
B. Cross-lingual automated search by topics
C. Detecting topic candidates for tagging
A. Making connections

- Selective and curated list of approx. 80 topics (UNESCO Thesaurus, UK Archival Thesaurus (UKAT))
- Topic labels translated by network contributors
- Importing finding aids = assign topic
- Importing source guide = inherit topic
Challenge

- Not all archival descriptions include subject headings
- Not all countries use source guides
- Content providers have to map their subject headings/source guides to the central topics
- Results in inconsistent coverage of topics
  - Currently predominantly for material from France (60 topics) and Germany (7 topics)
Solutions

A. Connecting existing topics
B. Cross-lingual automated search by topics
C. Detecting topic candidates for tagging
B. Searching by topic across languages

Goal: Identify material not yet tagged with a specific topic

○ Not requiring subject headings to be available in the metadata, but picking up on descriptive information e.g. in the title or content summary
○ Extension of current keyword search in the portal
○ For users and for content providers
B1. The manual way: crowd-sourced topic mapping

Scope and content

B1. The manual way: crowd-sourced topic mapping
B2. Automated detection tool

- Sample data currently included
  - Test sample of 636,000 documents of 12 topics
    - relatively small sample leaning towards specific subjects
  - Languages covered
    - Content predominantly in French and German, with some documents in Finnish, Latvian, and Polish
    - Additional languages for searching are English, Hebrew, Italian, Russian, Slovenian, Spanish, Swedish
B2. Automated detection tool

- Concept search
- Entity search
B2 a). Concept search

- Use of NLP and mathematical algorithms
  - Fast-Text word embeddings to capture the “meaning” ➔ representing user query as cross-lingual word embedding
  - Ranking all documents by measuring the “semantic” similarity of their vectors with vector of the query ➔ Using cosine similarity
B2. b). Entity search

- Mapping the entity searched for to its equivalent in Wikidata and VIAF (if present)
  - Retrieving name variations in other languages under study
  - Pre-processing name variations e.g. by leaving aside life dates or other characteristics that are (only) sometimes included in brackets
  - Searching for all name variations’ occurrences in the complete dataset
Cross-lingual Search on Archives Portal Europe

Welcome! This is the alpha version of a cross-lingual search tool for concepts and entities across the entire APE collection. To know more about the project, visit our GitHub.

Keyword(s)
- One or more
- Default is searching both, wherever they appear in the description
- You can use wildcard *
- You can use quotation marks “ “
- You can use BOOLEAN OPERATORS: AND, OR, ANDNOT

Language

How many results you want

Concept or Entity

BEMS = will search for the name of the entity in all languages available in Wikidata
### How the results look like

<table>
<thead>
<tr>
<th>Filename</th>
<th>Topic</th>
<th>Content</th>
<th>Country</th>
<th>Period</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armeekorps und Gardekorps der Preußischen Armee</td>
<td>Firstworldwar</td>
<td>Bd. 1 Armeekorps und Gardekorps der Preußischen Armee: F1537206 Enthält: Empfangsscheine Kriegsstammrollenauszüge Postkarten Urlaubsscheinepreußischen armeekorps gardekorps postkarten enthalt</td>
<td>Germany</td>
<td>-</td>
<td>0.474565</td>
</tr>
<tr>
<td>Infanterie-Regimenter der Preußischen Armee</td>
<td>Firstworldwar</td>
<td>Gefechtsberichte des Reserve-Infanterie-Regiments 98 Infanterie-Regiment der Preußischen Armee: F1538197 Enthält u.a.: Auszüge aus erbotenen französischen Befehlen und Kriegstagebüchern, mitgeteilt vom Generalkommando XIV. Armee-Korps.- Druck, 1915französischen preußischen armeekorps infanterieregimenter generalkommando</td>
<td>Germany</td>
<td>-</td>
<td>0.464275</td>
</tr>
<tr>
<td>Groener, Wilhelm (Generalleutnant, Reichswehrminister)</td>
<td>Democracy</td>
<td>Bd. 4 Groener, Wilhelm (Generalleutnant, Reichswehrminister): F1537461 Enthält: Kriegsgeschichte: Plan der französischen Heeresleitung zum Angriff auf Tann in Orleans 1870 Schlussuebungsreise, 1896 Kriegsgeschichte Taktik Befestigungslehrere Waffenl rh Verkehrsmittel Russisch Terrainlehe Staats- und Völkerrechtfranzösischen kriegsgeschichte heeresleitung groener wilhelm</td>
<td>Germany</td>
<td>-</td>
<td>0.462865</td>
</tr>
</tbody>
</table>
C. Detecting topic candidates

- Flag material related to a specific topic as a precursor to tagging the material
  - Not decided yet, who would do the tagging and where
  - Might be something that only happens in the aggregation process
  - Might be something that content providers take on board for integration at the source
How the detected topics are presented

Multilingual Detection for Archives Portal Europe

Welcome! This is the beta version of a multilingual detection tool for concepts and entities.

To know more about the project, visit our GitHub

Submit

<table>
<thead>
<tr>
<th>Content</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Anna Gurney): ['start': 12, 'end': 23, 'label': 'PER', 'score': 0.999872088432312], (William Hooker): ['start': 31, 'end': 45, 'label': 'PER', 'score': 0.998592663288116], (Royal Astronomical Society): ['start': 137, 'end': 163, 'label': 'ORG', 'score': 0.9148572963078817], (Frank Dyson): ['start': 188, 'end': 194, 'label': 'PER', 'score': 0.9712426656398621], (Emmanuel College): ['start': 200, 'end': 206, 'label': 'ORG', 'score': 0.99304095597183075], (Royal Society): ['start': 207, 'end': 213, 'label': 'ORG', 'score': 0.997593618335724], (Bedford College): ['start': 288, 'end': 294, 'label': 'ORG', 'score': 0.88109826502999], (Myra Curtis): ['start': 394, 'end': 400, 'label': 'ORG', 'score': 0.5369300842285156], (Pernel Strachey): ['start': 463, 'end': 479, 'label': 'PER', 'score': 0.9774636003780355], (Hertha Ayrton): ['start': 501, 'end': 507, 'label': 'PER', 'score': 0.881029694236755], (Margaret McNair Stokes): ['start': 606, 'end': 612, 'label': 'PER', 'score': 0.9317196754728], (Jane Ellen Harrison): ['start': 712, 'end': 718, 'label': 'PER', 'score': 0.999837683929443]</td>
<td></td>
</tr>
</tbody>
</table>
The future plans
Prospective extensions for automated tool

- Employ taxonomies from other LOD vocabularies
- Extend dataset (to eventually encompass the whole data set held in APE)
- Include more languages
Follow the development on GitHub

- https://github.com/ArchivesPortalEuropeFoundation/Topic-Detection
  - The project’s status and next steps
  - The code for the tool and the training data used
Any questions?

Thank you for your attention!

info@archivesportaleurope.net

www.archivesportaleurope.net  @archivesportal