
ARCHIVES
PORTAL EUROPE
Foundation
Stichting Archives Portal Europe Foundation

ARCHIVES PORTAL EUROPE

Automated search in the archives: testing a new tool

11 February 2022

Today’s agenda

- Background and introduction to the tool
- Some notes on testing the tool
- Next steps

- - - Change to breakout room for part 2 - - -

- Free testing
  - Each participant by her-/him-/themselves
  - With option to ask questions in between
- Q&A

Background for developing the tool

Controlled access terms in archives

- Ideally
  - Assigned when creating the descriptive metadata
- In reality
  - Not always part of the archival description tradition
  - Not necessarily taken from controlled vocabularies (national or international)
  - Might only be available on collection level

Integration with archival systems

- Ideally
  - Designated fields for subject headings/named entities
  - Allowing for integration of controlled vocabularies
  - Enabling the addition of URIs for linked data vocabularies
- In reality
  - Not available at all or not as a designated field
  - Missing connection to (linked data) vocabularies

Encoding in archival metadata

- Ideally
  - Designated elements for subjects/named entities
  - Including literals and URIs from linked data vocabularies
- In reality
  - Not always in designated elements, but part of longer descriptive texts
  - Only giving the names, but no links

Search all content...

Democracy

Activism and struggles for democracy

Early Modern Period

Documents on the 1400-1800 period. This article is a currently a stub.

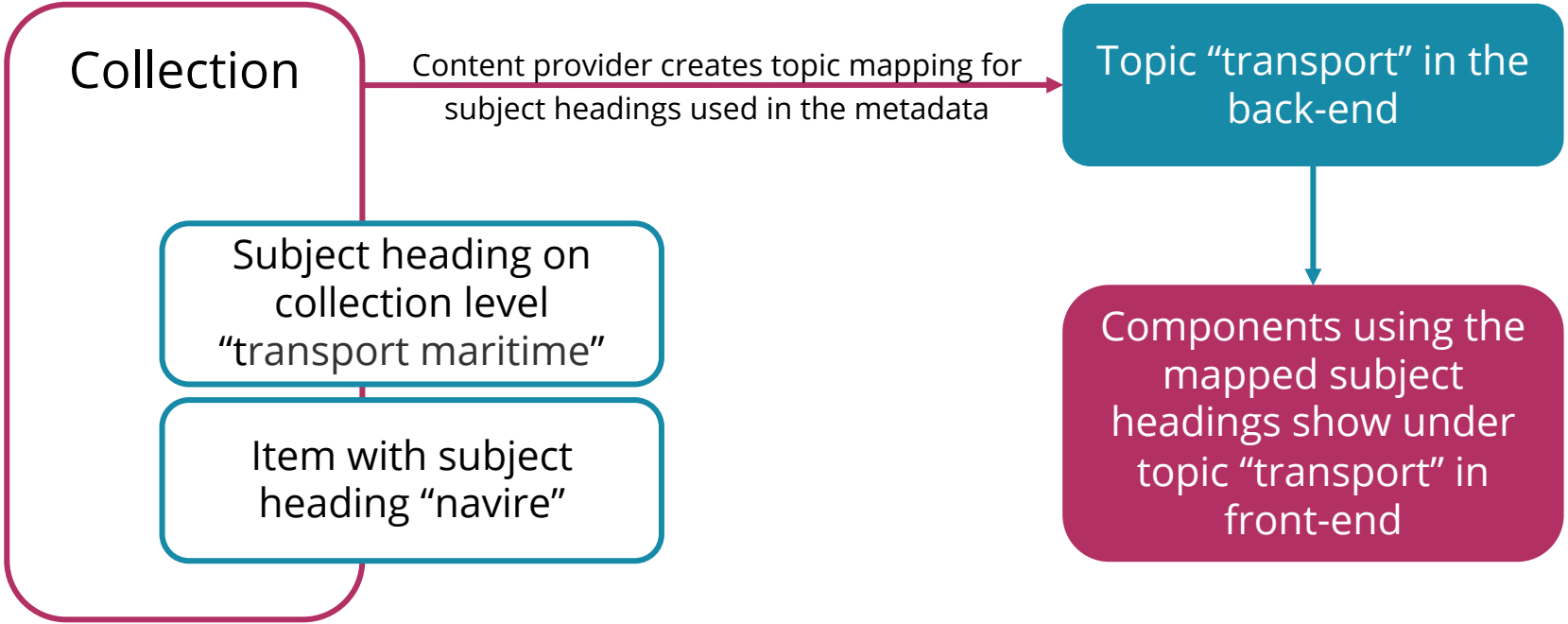
Economics

Properties, commerce, fiscality, accounting, etc.

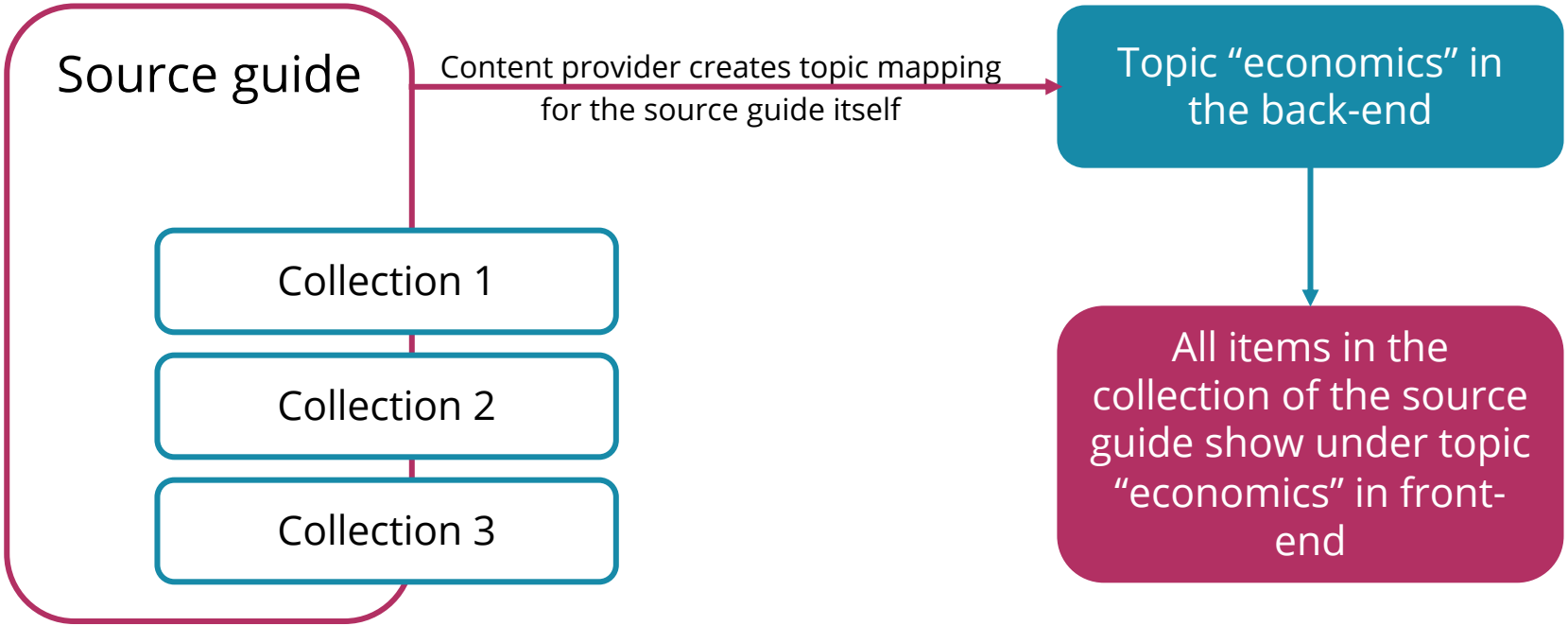
Making connections

- Selective and curated list of approx. central 80 topics
- Built from the UNESCO Thesaurus and from the UK Archival Thesaurus (UKAT)
- Topic labels translated by network partners
  - Thereby available in a variety of languages according to the language of the user interface of the portal

Use of existing subject headings



Connecting topics with the materials



Challenge

- Not all archival descriptions include subject headings
- Not all countries use source guides
- Content providers have to 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)

What the tool includes and does

Objectives - Step 1

- Identify material not yet tagged with a specific topic but related to it
  - 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

The data

- Currently included
  - Test sample of approx. 675,000 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

Concept search - how it works

- Using Fast-Text word embeddings
  - Capture the “meaning” of the document’s content
  - Pre-trained on Wikipedia and aligned in a common cross-lingual “semantic” space by the project MUSE

Concept search - how it works

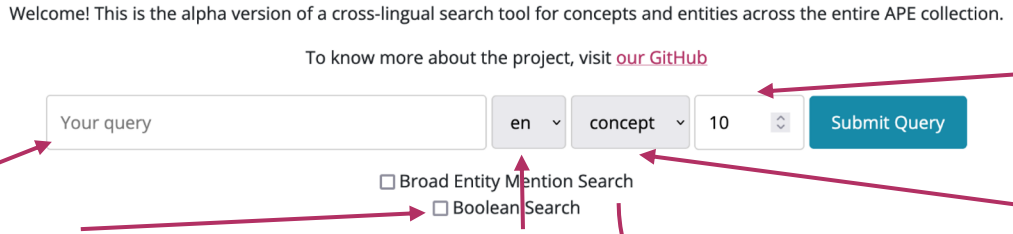
- Representing user query as cross-lingual word embedding
  - For results in languages different from query language
- Ranking all documents by measuring the “semantic” similarity of their vectors with vector of the query
  - Using cosine similarity

Entity search - How it works

- 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

How the tool looks like

Cross-lingual Search on Archives Portal Europe



How many results you want

Keyword(s)

- One or more
- Default is searching both, wherever they appear in the description
- You can use **wildcard ***
- You can use 3 BOOLEAN OPERATORS:
  AND
  OR
  " "

~~**ANDNOT**~~

Language

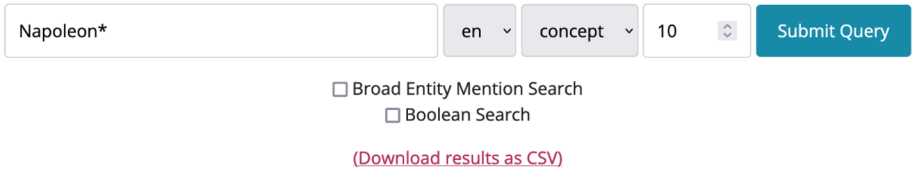
BEMS =
will search for the name of the entity in all languages available in Wikidata

Concept or Entity

Run the search

- Please note, that this might take a while
  - Especially for the entity search, which runs on-the-fly queries to Wikidata and VIAF
  - Alpha version with certain constraints re available channels
- You usually will get the amount of results specified in your search parameters (eg 10)
  - But they could be less - or even zero!
    - Especially for the entity search, which is more specific to the search term

How the results look like



Filename	Topic	Content	Country	Period	Score
Armeekorps und Gardekorps der Preußischen Armee	Firstworldwar	Bd. 1 **Armeekorps** und **Gardekorps** der **Preußischen** Armee:F1537206 **Enthält**: Empfangsscheine Kriegsstammrollenauszüge **Postkarten** Urlaubsscheinepreußischen armeekorps gardekorps postkarten enthält	Germany	-	0.474565
Heeresgruppen des Deutschen Heeres	Firstworldwar	**Kriegsgliederung** Heeresgruppe **Erzherzog Karl** von Österreich bzw. **Erzherzog Joseph Heeresgruppen** des Deutschen Heeres:F1536618 Enthält: Errichtung eines Stabsoffiziers der Flugabwehrkanonen bei der Heeresgruppe **Erzherzog Karl**erzherzog kriegsgliederung joseph heeresgruppen karl	Germany	-	0.465107
Infanterie-Regimenter der Preußischen Armee	Firstworldwar	Gefechtsberichte des Reserve-Infanterie-Regiments 98 Infanterie-Regimenter der **Preußischen** Armee:F1538197 Enthält u.a.: Auszüge aus erbeuteten **französischen** Befehlen und Kriegstagebüchern, mitgeteilt vom **Generalkommando** XIV. Armee-Korps.- Druck, 1915französischen preußischen armeekorps infanterieregimenter generalkommando	Germany	-	0.464275
Groener, Wilhelm (Generalleutnant, Reichswehrminister)	Democracy	Bd. 4 **Groener**, **Wilhelm** (Generalleutnant, Reichswehrminister):F1537461 Enthält: **Kriegsgeschichte**: Plan der **französischen Heeresleitung** zum Angriff auf Tann in Orleans 1870 Schlussuebungsreise, 1896 **Kriegsgeschichte** Taktik Befestigungslehre Waffenlehre Verkehrsmittel Russisch Terrainlehre Staats- und Völkerrechtfranzösischen kriegsgeschichte heeresleitung groener wilhelm	Germany	-	0.462865

How the search results are presented

- For each result you'll find
  - The title of the object (Filename)
  - The topic that it is currently tagged with (Topic)
  - The scope and content note (Content)
  - The language of description (Country)
  - The date(s) of creation (Period)
  - The score indicating how close the semantic relation is between your keyword and the result (Score)

How the search results are presented

- For each result you'll find
  - The words that have found to be a match to your keyword highlighted in variations of blue
    - The darker the blue, the stronger the semantic relation
- When searching for entities, you'll also see
  - Links to Wikipedia and/or VIAF, in case matches to your keyword have been found there (at the top of the table)

Next steps in the tool’s development

Objectives - Step 2

- 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

Prospective extensions

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

Collaborative extension of topics (1/2)

- In eventually bringing both objectives together, this tool could be used by researchers
  - To find and flag documents relevant to their research
    - Not necessarily bound to existing list of curated topics
    - But also identifying new topics of interest
  - To bookmark topic-specific documents and save these bookmarks in their user accounts of the portal
  - To expand automatically on their keyword searches

Collaborative extension of topics (2/2)

- Similarly, the tool could eventually be used by content providers
  - To review and refine their existing topic mappings
  - To identify and flag documents in other collections that relate to a specific topic
  - To use the resulting list of topic-relevant documents as a basis for adding subject headings at the source

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

Thank you for your attention!

We are changing to the second part of the event in a breakout room

info@archivesportaleurope.net

www.archivesportaleurope.net @archivesportal