Outline: Problems to be addressed by BLIZAAR

29.01.2016, MDU

The main problem we identified during the preparation of BLIZAAR is that humanities researchers in the European Integration Studies department (EIS) at the CVCE struggle to get a sense of **what information is available** in the backend and to **identify gaps**. CVCE is currently working on a faceted search functionality for its collections and ePublications which may complement any solution developed within the scope of BLIZAAR. Outlined below is an expanded description of problems already outlined in the proposal (p.8) and a first sketch of network models.

Data

Named entities have been annotated, a Neo4j database stores links between entities based on co-occurrences in CVCE documents and ranks them based on the number of co-occurrences and Jaccard distance. Currently (including duplicates and fragments) we have:

People: 15.800
Institutions: 4100
Places: 3400
Time period: 1945 – ca. 2010

Detailed document metadata such as author, creation/publication date, title, publisher etc. is available.

Every document (text, photo, audio, video…) is part of one or more ePublications which are organized in hierarchy. See this example:
http://www.cvce.eu/recherche/unit-content/-/unit/02bb76df-d066-4c08-a58a-d4686a3e68ff

Another CVCE project, EIBIO, which is closely connected to histoGraph, aggregates machine-readable data on people's biographies and links them to the institutions for which they have worked for. At this stage, some 200 biographies have been processed (time stamped). I plan to use 3 months of the budgeted student time to increase this dataset.

Our work on histoGraph provided us with a dense dynamic, multimodal network of entity co-occurrences which is too large to be visualised as a node-link-diagram and contains the information in ca. 30 ePublications (which themselves are eclectic hand-picked selections of documents which can comprise of 100s of objects).

We have already shared the histoGraph database with you, additional data is of course available for you.

Diversity and continuous coverage

Goals

Evaluate to which extent ePublications adhere to some of the CVCE internal quality standards. Relevant in the context of BLIZAAR: For each entity (person, institution) there need to be

a. A balance of sources from different types of archives
b. A balance of different types of media (photo, text, video…),
c. These need to cover the full time period of the respective subsection of the ePublication.

Researchers from the EIS department need to be able to 1) quickly assess whether conditions a), b), c) are met within a specific ePublication and 2) be made aware of the sections within the ePublications where they don't.

Example

There might be a gap in the continuous coverage of the activities of an entity, say the European Central Bank for 2011. Such gaps are usually found by chance or following a specific search for them. Continuous coverage over time is entity-dependent: The European Central Bank was established in 1998, therefore a lack of references to it before this year does not constitute a gap.

Theme-like networks: search

Goals

Researchers from the EIS department would like to have themes associated with each document. Such themes would usually be assigned manually by EIS. They depend on the interests of a researcher, there can therefore not be an definitive or objective catalogue of themes. Ideally, each researcher can define his/her theme based on specific research questions and interests. In contrast to the faceted search engine CVCE is currently developing, a theme-based approach should allow 1) the discovery of relevant documents and entities which would be missed by a known-item search and 2) provide a higher-level perspective on CVCE document collection (and the history it represents) as opposed to the document/entity-focus of the search engine.

Proposed approach

Standard topic modelling techniques may not be the best solution in our case given the heterogeneity of our corpus (diversity of document types and languages, the date range across 70 years) and the need to pre-define a number of topics.

I suggest to explore the extent to which we can address the need for themes using entity co-occurrences, clustering and transitivity. The Eurovoc thesaurus (http://eurovoc.europa.eu/drupal/?q=abouteurovoc&cl=en) may be the point of departure. Eurovoc was originally designed to help catalogue the output of EU

institutions. All entries are hierarchically organized according to subject specificity and consider synonyms. Eurovoc itself can be conceptualised as a multi-layer network with multiple types of nodes which are organised hierarchically. See their website for detailed descriptions of entity types and relations between them.

Example: The entry "internal migration" (http://eurovoc.europa.eu/1916) has the higher-level term "social questions" above it and "seasonal migration", "commuting", "interurban migration", "nomadism" (…) below it. The entry "seasonal migration" has the related terms "seasonal worker" and "tourism" associated with it.

A combination of one or more Eurovoc entries (including their associated entries) together with named entities from CVCE documents, a time range and a free keyword search could be an effective way for users to characterize the theme they are interested in. Eurovoc's hierarchical organisation would allow users to opt for broader or more focused searches; the histoGraph entity co-occurrence network may help to expand the search space based on closely associated entities.

In its simplest from, this selection yields a list of user-selected keywords and associated keywords. Documents which mention a higher number of these keywords have a higher chance to be of relevance for the respective theme.

A more complex approach could consider transitivity and clustering. Keywords which cluster may indicate a sub-theme. Named entities (or just words in general) in the documents which co-occur very often with these keywords may prove to be a worthwhile extension of the original keyword list.

[We could also think of the user-selected keywords as a multimodal network which is placed on top of the histoGraph multimodal network. It is an exciting thought but makes things complex very quickly.]

All in all, from a user perspective this process could include (some of) the following steps:

1) Characterize a theme based on Eurovoc entries, histoGraph entities and free keywords. Alternatively: Select one of the themes EIS researchers used to for the existing ePublications
2) Fine tune this selection, focus on certain keywords/entities, exclude others
3) Validate theme construct using retrieved documents ("Do these documents cover what I am interested in?")
4) Explore and further modify theme construct based on retrieved documents
5) Gain a higher level perspective of the theme: Changes over time, position in the overall network of entities and documents, gaps

The challenge is to identify the most relevant entities/documents among the large number of candidates. In essence, this requires us to qualify the nature of the co-occurrence and the role of a document in the context of a given theme. Possible approaches could include:

a) Centrality and clustering
b) User intervention and a live-update of the theme-like network

c) Contextualisation using EIBIO
d) Collection of user activities
e) Dedicated crowd-sourcing tasks

Example
A user selects "NATO" + "Poland" + "date range: 1975--1995". Likely associated keywords for this example would be:

- in Eurovoc "NATO countries", "OTAN", "Poland", "Eastern Europe", "regions of Poland"…
- in histoGraph data: "Lech Walesa", "Leszek Miller", "Danuta Hübner"…

These entities constitute a way to identify documents which do not mention explicitly "NATO" and "Poland" together but are still relevant in this context. This would also allow interaction between search keywords and available content.
Finally, weighting the significance of certain keywords would make it possible for users to fine tune their theme-searches and explore different aspects of their theme.

Theme-like networks: exploration

Goal
Clusters of co-occurring Eurovoc entries and histoGraph named entities could also be the basis of an interactive dynamic map of all entities, documents and their relations provided by the CVCE. There is likely no one way to grasp and represent the complexity of these relations but specific use cases are conceivable:

a) Explore the overlap between user-selected themes
b) Explore the overlap between CVCE-created themes (ePublications)
c) Identify unwanted gaps between themes
d) Check for continuous coverage
e) Understand where a new document which is about to be added to the Backend fits in this map and which gaps it closes
f) Understand the context of a given document, how it fits in this map and which gaps it closes
g) Receive recommendations for other documents
h) Categorize types of recommended documents which allow users to explore other themes associated with their original theme. Where applicable, link back to existing ePublications.