C2DH Project Vision

(revised July 2017)

This is an update of the Project Vision which reflects the consequences of CVCE's integration into the University of Luxembourg but also progress made by the consortium in the last months. The initial document may be found

here

.

Changes to the project vision

Since work on the CVCE Backend and on ePublications will not be continued, we need to set new accents for the project vision. While we started with the highly specific use cases for EIS researchers working on the CVCE Backend and creating ePublications, we need to revisit the original use case and adjust it in direction of a more general-purpose approach to the exploration of digitized cultural heritage corpora. This does not however constitute a major deviation from the overall goals and techniques we have been working on so far.

The key objective for BLIZAAR's humanities use case remains the graph-based exploration of the CVCE corpus. Graph centrality scores, clusters and higher-level patterns in the data are less relevant for this and other use cases in cultural heritage. The data in the corpora we target is typically highly heterogeneous, contains gaps and biases which make a computational approach problematic and requires advanced skills in data analysis which our targeted users do not have. Rather, we are interested to explore the manifold links a given node has to other nodes with regard to the corpus, not the historical events it depicts.

Our targeted users will benefit from the system if:

1. it can show them how a given node (e.g. person, institution, topic) is present in the corpus, i.e. how it is linked to others,
2. if it helps them with the discovery of other relevant nodes similar to a recommender system
3. if it leads them to documents for further close reading.

Lessons learned so far

Need for simplicity

Our target user group has very little experience with data-driven visualisations or networks in general. A key premise is therefore the simplicity and comprehensibility of any techniques we apply. Standard force-directed graphs are very hard to comprehend for our users (interpretation of positioning, meaning of relations, network boundaries, legibility) and are therefore often met with skepticism or aversion. They should only be used where they play out their strengths and are relatively easy to interpret. As a direct consequence, we need to focus on:

- Focus on alternatives: matrices, hive plots, parallel coordinates, slope graphs…
- Vis based on easy-to-comprehend data and documentation of what is on display, no hairballs

- Only basic quantitative measures such as degree centrality
- A query-building system which allows users to benefit from complex algorithms such as Degree of Interest without needing an in-depth understanding of how they work

Node and edge types

We have identified the following node types to be relevant, all linked either by co-occurrences in documents or by presence in an ePublication:

1. Persons
2. Institutions
3. Topics
4. ePublication segments
5. Documents
6. Time

The relationships are derived from 1) the co-occurrences of document annotations (persons, institutions, topics) and 2) properties of the documents (time as stored in the metadata, being part of an ePub segment). The following undirected edges are of interest to us:

Person	Person	“is mentioned in” based on a projection of person-doc relationship
Person	Institution	“is mentioned in” based on a projection of person-doc relationship
Person	Topic	Person is mentioned in a document, topics have been assigned to documents
Person	Document	Person is mentioned in document
Person	ePub segment	Person is mentioned in a document which is part of 1 or more ePub segments
Person	Time	Person is mentioned in a document which has a creation date
Institution	see “Person”	
Topic	see “Person”	
ePub segment	Document	ePub segment contains document
Time	Document	Creation date of a document

An exploration of one node should yield information on how it relates to the others. A query for a person should e.g. tell us with which institutions, in which ePublication segments, with which topics and in which documents this person appears in the corpus.

Interactive node list

Network visualisations make it very hard to get an overview of which nodes a network actually contains. Lists, especially if they are sortable and segmented by node type can do a much better job here and help us identify the interesting nodes. Such an interactive node list could give an overview of nodes in a given graph and provide additional, quantifiable information about a node, e.g. centrality or presence over time.

List of matching documents

Parallel to the interactive node list a list of matching documents should be displayed which lets users jump from the visualisation to the content. The list should update according to user actions: A selected edge between a person and an institution in a graph should then display all document in which both of them are mentioned.

Entry points

The changes in the scope of the project vision and the progress in the development of demos means that we can focus on the following entry points:

Node-based. Users select one or more nodes (person, institution, topic) as input for the Degree of Interest algorithm and learn about co-occurring entities, associated topics, ePublications and their presence over time.

Keyword-based. Related to the previous approach but rather than explicitly selecting nodes, users auto-select all nodes which match a keyword phrase.

There are two ways to build a graph:

1. Select all nodes and parameters, run DOI or other algorithm, visualise subgraph
2. Iteratively add one node at a time to the graph. Each added node brings with it it's respective subgraph (ego network or DOI)

High-level tasks

1. Keyword-based search for nodes
2. Select nodes (annotated persons, institutions, topics or documents) from search results
3. Filter search results for time period and media type
4. Use DOI algorithm to create a subgraph based on user selected nodes and filters
5. Display an interactive list which contains nodes in the subgraph sorted by node type and displaying node properties such as presence in the corpus or subgraph over time, degree in the subgraph.
6. Display different visualisations of the subgraph (Hive plot, matrix, multilayers). Nodes selected in the nodelist are highlighted in the vis and vice versa.
7. Display list of documents which match selected nodes in the interactive node list and selected nodes and edges in the visualisations.
8. Select interesting nodes and edges in the interactive list or vis to see links to underlying documents
9. Use time slider to observe the changing "neighbourhoods" of nodes: With which persons, institutions are topics do they co-occur over time?

Question

Can we automatically guess which visualization methods will work best for a given subgraph? For example:

- The graph is very small: hive
- The graph is very dense: matrix
- ...

Initial project Vision