histograph - Graph-based exploration crowdsourced indexation

Lars Wieneke
Centre virtuel de la Connaissance sur l'Europe
lars.wieneke@cvce.eu

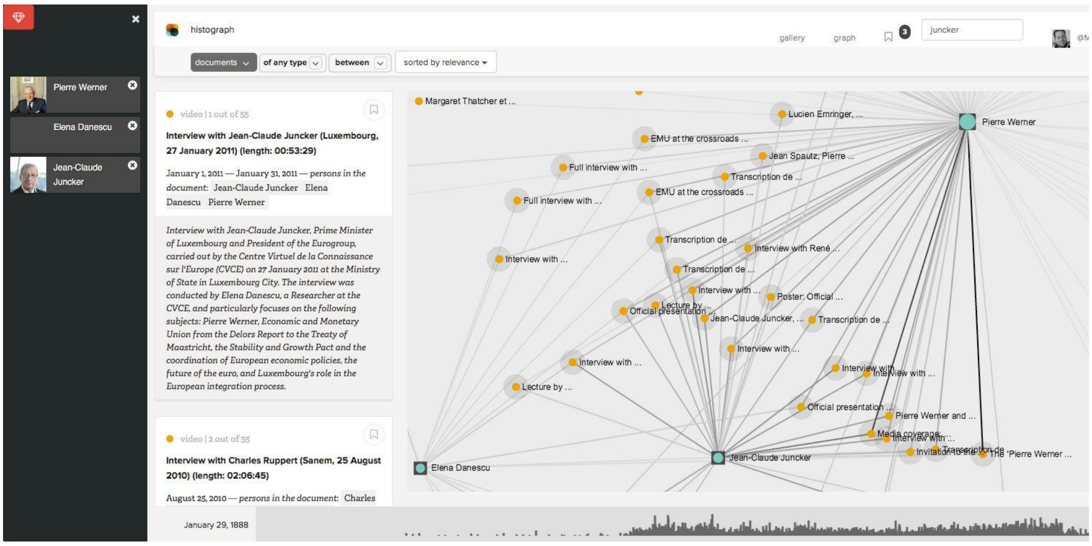
Daniele Guido
Centre virtuel de la Connaissance sur l'Europe
daniele.guido@cvce.eu

Marten Düring
Centre virtuel de la Connaissance sur l'Europe
marten.during@cvce.eu

histograph (http://histograph.eu) is an open source application for the graph-based exploration and crowd-based indexation of multimedia documents. histograph was initially developed as one of two demonstration applications for the FP7-funded research project CUbRIK, which aimed to find new pathways for the combination of human and machine computation in multimedia search. histograph combines the graph-based exploration of larger cultural heritage collections with crowd-based indexation.

For the CVCE use case, the tool opens up a new perspective on the CVCE's collections, which comprise approximately 20 000 digitised text documents, photos, audio recordings and videos. histograph adds an explorative approach to the hierarchically organised, expert-curated collections: users can decide what interests them and find their own path through the collections.

histograph uses a Neo4j graph database to store relations between entities. This approach facilitates queries that would be computationally expensive in relational databases but are easily available in graph databases, such as the calculation of paths between entities that are not directly connected. Below we show the result of a query for all documents which connect three people. The left column now shows the list of selected people and the middle column the list of documents that mention two or more of the selected entities. The right column provides a graph of the co-occurrences and lets users select and further explore any relations which are of interest to them.



video | 1 out of 55

Interview with Jean-Claude Juncker (Luxembourg, 27 January 2011) (length: 00:53:29)

January 1, 2011 — January 31, 2011 — *persons in the document:* Jean-Claude Juncker Elena Danescu Pierre Werner

Interview with Jean-Claude Juncker, Prime Minister of Luxembourg and President of the Eurogroup, carried out by the Centre Virtuel de la Connaissance sur l'Europe (CVCE) on 27 January 2011 at the Ministry of State in Luxembourg City. The interview was conducted by Elena Danescu, a Researcher at the CVCE, and particularly focuses on the following subjects: Pierre Werner, Economic and Monetary Union from the Delors Report to the Treaty of Maastricht, the Stability and Growth Pact and the coordination of European economic policies, the future of the euro, and Luxembourg's role in the European integration process.

video | 2 out of 55

Interview with Charles Ruppert (Sanem, 25 August 2010) (length: 02:06:45)

August 25, 2010 — *persons in the document:* Charles

Figure 1: Screenshot from the application which demonstrates the retrieval of shortest paths between persons.

histograph enables an interest-driven exploration by users and provides them with an effective way to retrieve and explore any relationships which are of interest to them. Compared with the museum-like order of the traditional CVCE collections, histograph models rather resemble a visit to an archive which holds the promise of serendipitous discoveries.

histograph combines tools for the automatic detection of named entities (people, places, institutions and dates), enrichment with DBpedia and VIAF with crowd-based annotations. By default, every automatically detected entity is indicated as pending validation by a human user. Automatic entity detection works very well overall but will always remain imperfect in places. To address this, histograph depends on human validation and error correction. All annotations can be in one of three stages: not validated, validated or disputed. In addition, users are encouraged to fix mistakes themselves by annotating new entities and by flagging wrong entity types, fragments, duplicates or erroneous annotations. To avoid accidental annotations and reduce the risk of vandalism, histograph treats every annotation as a suggestion pending confirmation by other users.

We operate with two types of crowd tasks: tasks targeted at a generic crowd, which means that anyone is able to provide input, and harder, more challenging tasks, which target expert users. Users qualify for these expert tasks on the basis of their previous actions. For example, a user who annotates many documents associated with Pierre Werner will be presented with a task to validate related annotations by others and to identify unknown entities in related documents. They thereby implicitly become experts on certain persons, topics or documents. All edits can be validated by peers and thereby confirmed but always remain available for future changes.