**CHEEMA — Modeling**

<3 figures>

**Modeling Concepts to Improve the Search Capabilities on Ancient Corpora**

Cheema, M. F., Jänicke, S., Blumenstein, J,. Weilbach, C. and Scheuermann, G.

**Motivation**

A common task nowadays is the retrieval of information using search engines such as Google. Often, the results of a usual keyword search are not satisfying, and the user is forced to reformulate the query in order to improve the quality and to increase or reduce the number of search results.

When searching ancient texts for passages related to a specific topic, humanities scholars encounter similar problems. Considering *epilepsy* as an example, the traditional keyword search for the corresponding Latin term *morbus comitialis* (‘disease of the assembly’) yields incomplete results. Operating a truncated search for *morb\* comiti\** extracts more but less accurate results. It also yields false positives like a passage from Titus Livius (*Ab urbe condita libri* 41,18,16) containing both the words *morbus* and *comitia* but in a political context only. A further issue regarding the mentioned example is the existence of texts that relate to *epilepsy* but don’t use the corresponding terms explicitly (e.g., Lucretius, *De rerum natura* 3,487ff). Here, the disease is paraphrased via occurring symptoms such as *spuma* (‘foam’) and *concidere* (‘collapse’). Furthermore, searching for *epilepsy* only by the term *morbus comitialis* is incomplete because there are many different terms denoting the disease. Therefore, a simple keyword search is inadequate in finding appropriate results.

The prior goal of the eXChange1 projectis to extract more and more accurate results by extending the traditional keyword search with a so-called concept search. For this purpose, we develop a concept editor that supports the scholars in modeling their ideas of concepts. Using these concept models, we operate a concept search on ancient corpora and adequately rank the results, such that texts supporting the modeled concept are ranked higher. The significance of our concept modeling methodology is to empower the humanities scholar to construct concepts using her expert knowledge, and iteratively refine the model after analyzing the search results.

**Related Work**

Besides the aforementioned traditional keyword search, more sophisticated approaches have evolved over time. A *semantic search* (Guha et al., 2003) takes the contextual meaning of a term into account by automatically considering synonyms, word variation, and so on, whereas a *concept search* (Giunchiglia et al., 2009) automatically retrieves data that is conceptually similar to the formulated query. These search strategies can be supported with *topic maps* (Park et al., 2002) or *concept maps* (Novak et al., 2006), where certain topics and relationships can be modeled in a graph structure. Based upon reference texts such as glossaries, thesauri, and classification systems, these models can be automatically generated for modern languages. Appleford introduced human-generated *topic profiles* that consist of various allowed and disallowed terms forming a Boolean query (Appleford et al., 2013).2 The major difference of our suggested concept modeling compared to existing techniques is that concepts in information retrieval are formulated primarily using existing semantic knowledge from thesauri, topic maps, semantic networks, and so forth. Our suggested concept modeling allows the scholar to express her understanding of the concept model without the need of any semantic knowledge.

Teevan argues that even with a perfect search engine, a poorly constructed search question may not lead to the right answer, so the user needs to be provided with a directed search system (Teevan et al., 2004). This becomes quite relevant in case of ancient corpora that comprise a particular complexity and uncertainty, which leads to a potential evolution of concepts in ancient times. Also, the sparse presence of reference texts for ancient languages requires a semiautomated generation of concepts. For this task, we present a new methodology to assist humanities scholars in searching ancient texts by empowering them with a directed search environment in the form of a concept editor (discussed in the next section), where a scholar can model a concept using simple graph building tools. This was also a requirement of the collaborating humanities scholars as they wanted full control over the search process, which an automated search system fails to provide.

**The Concept Editor**

Figure 1 shows a screenshot of the concept editor3 containing the concept *fever*. Various shapes are provided for concept modeling. Initially, the scholar defines the main concept *fever*, drawn in blue. Other shapes representing subconcepts or terms can be drawn via drag&drop to support structuring the concept in a meaningful way.

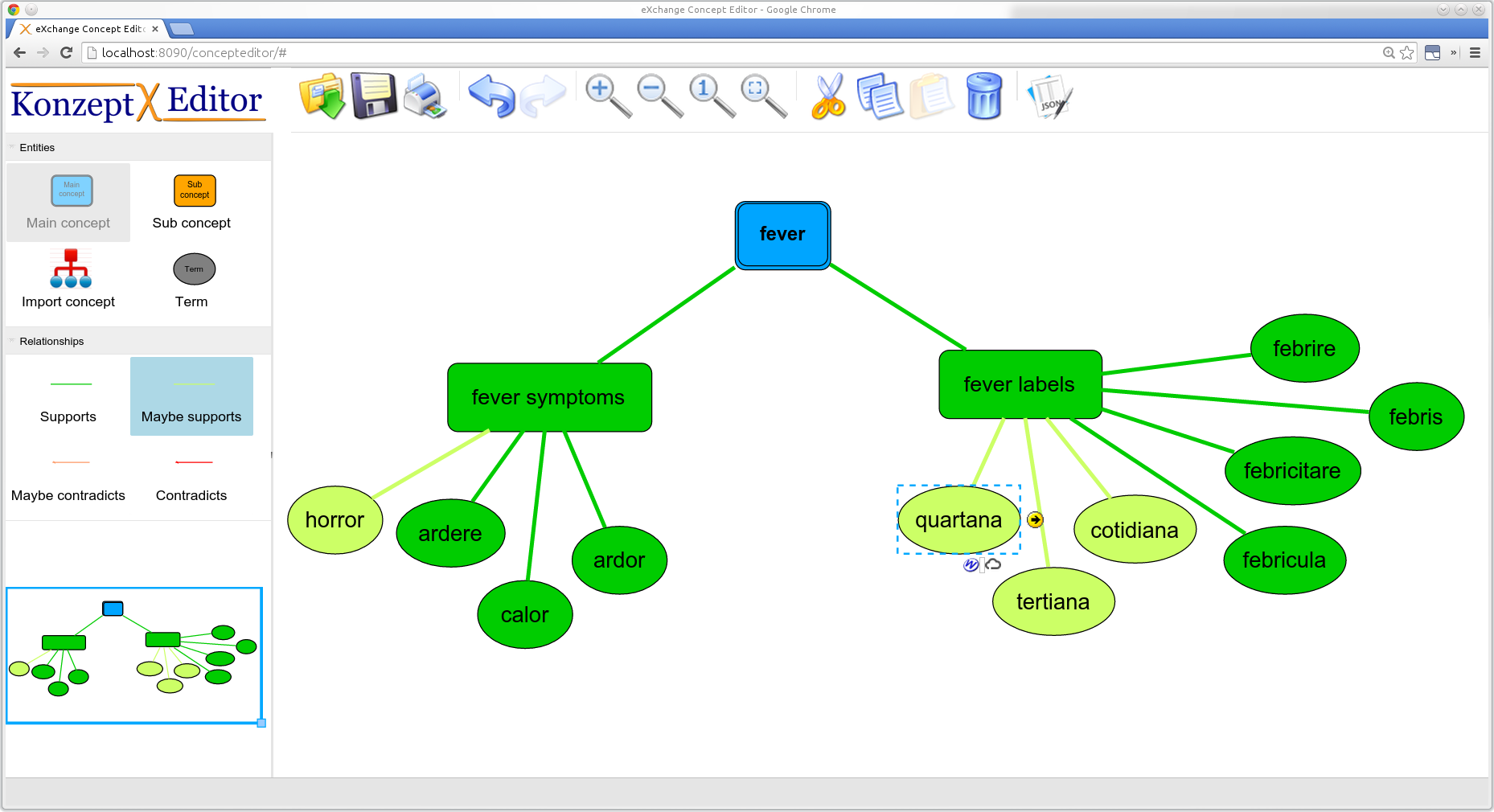


Figure 1. Concept editor.

The *fever* concept consists of the subconcept *fever symptoms*, which is connected to related Latin terms, and the subconcept *fever labels*, listing all known Latin terms used to describe fever. The editor provides four ways of connecting a shape to its predecessor discriminated by color. Green shapes support the given concept while red shapes are contradictory. Saturated colors represent definite knowledge of the scholars, whereas light shades represent assumptions. According to the model, texts containing the terms *febris* or *febrire* (‘to have fever’) were most likely used to address the concept *fever*. Terms like quartana may denote fever (‘quartan’) but also serve to name other things. Thus, they are colored in lighter shade.

Each term shape is connected to all possible spellings and word-forms contained in the database. A popup on demand provides a list and tagcloud that allow one to observe and select terms related to the corresponding concept. Figure 2 shows the selected word-forms for the term *quartana*. As *fever* in Latin is female, other gender forms of the adjective *quartanus* can be excluded. By taking advantage of language-specific grammar rules, it will help the scholar to avoid irrelevant results.

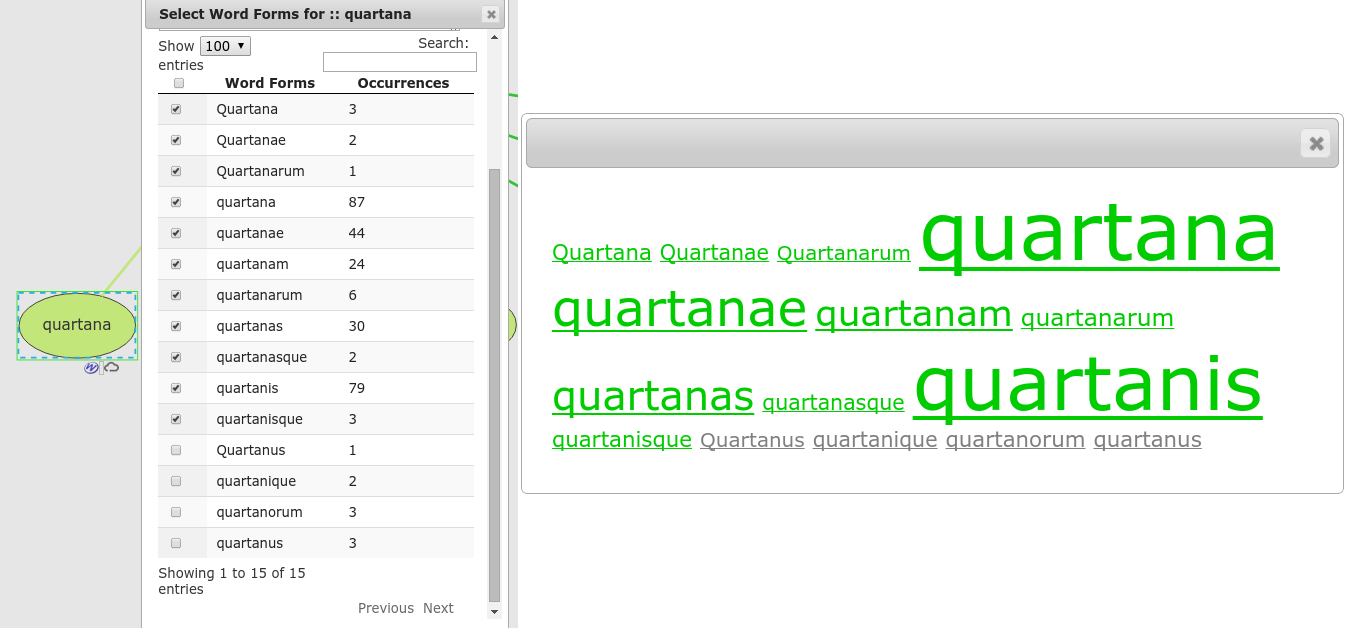


Figure 2. Selecting word-forms.

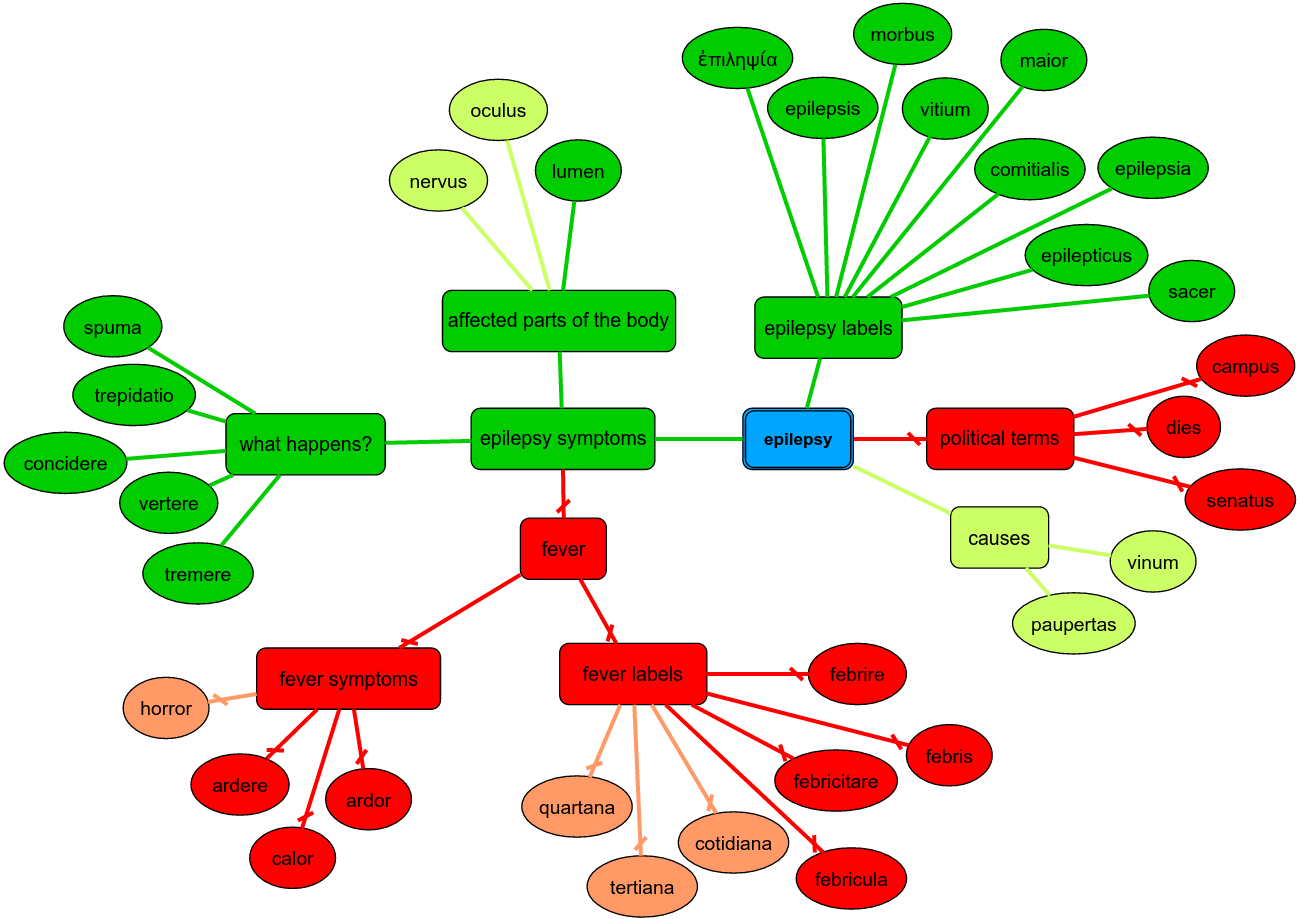


Figure 3. Concept *epilepsy*

The current model for the concept *epilepsy* is shown in Figure 3. It also contains subconcepts for labels and symptoms. Most of the Latin terms denoting epilepsy are represented by two single words (e.g., *morbus comitialis* or *morbus maior*). The subconcept *symptoms* is further subdivided, showing, e.g., *the affected parts of the body*. The unlikely symptom *fever* is imported as a contradictory subconcept, so all *fever*-related terms represent a contradictory relation to the concept *epilepsy*. Additionally, potential causes of epilepsy (e.g., *vinum* [‘vine’]) are given, and a list of contradicting political terms reduces the relevance of political texts for which the term *comitialis* is common.

After the concept is built, it is stored for persistence and forwarded to a concept search module. The selected word-forms and spellings of all terms are the basis for the search. For each text of the corpus, we compute a relevancy rank according to the proximity to the given concept. Each occurring ‘green term’ increases the rank of a text, and each occurring ‘red term’ decreases it. Finally, the result list contains all texts ordered by decreasing relevance.4

**Preliminary Results**

The results of the search for the concept *epilepsy* are partially consistent with the humanities scholar’s expectations. So, texts of authors popular for their medical works are highly ranked, e.g., *De medicina* by Aulus Cornelius Celsus and *Naturalis historia* by Pliny the Elder. Furthermore, a very low rank is calculated for the unrelated political text *Ab urbe condita libri*, written by the Augustan chronicler Titus Livius—one of the top results of the traditional keyword search.

The appearance of two texts with a high rank is surprising for the scholars: *De re coquinaria* by Apicius on cuisine and *De re rustica* by Columella on agriculture. That these texts appear among the results may be due to the term *vinum*. First, it can be assumed that it refers to epilepsy, but a detailed examination of the respective passages is required to find out if it occurs in remarks on epilepsy or, say, outline the relationship between nutrition and health (dietetics).

**Conclusion**

The benefit and potential of the presented concept search system compared to the traditional keyword search are not only in their multifacetedness but also the enabled persistence of the modeled concepts. Depending on the accuracy of search results, the scholar can iteratively improve results, either by adding or excluding terms from concepts. Retaining the control on the search and the ability to steer the search into a desired direction, now supported by the concept editor, was a major concern of the scholars.

Often, digitized ancient fragments are neither associated with an author nor with a certain title. This complicates their automated classification as authors are often related to specific genres. Based upon the modeling of generic concepts (e.g., the concept *political terms*), the potential genres of unclassified texts could be hypothesized.

As a side benefit of the concept modeling, we automatically collect and store the modeled hierarchies and relationships between (sub)concepts. Slightly enhancing the scholar’s modeling capabilities, we might be able to assemble thesauri for ancient languages in the future.

**Notes**

1. Digital Humanities project, eXChange: Exploring Concept Change and Transfer in Antiquity, http://exchange-projekt.de/.

2. Operating on social web sources, one is able to discover occurrences that match the given profile while dropping text passages that contain any of the disallowed terms. We consider that excluding text passages due to a disallowed term may result in the loss of important text passages.

3. The web-based concept editor is written in JavaScript and uses the libraries mxGraph (https://jgraph.github.io/mxgraph/) and jQuery (http://jquery.com).

4. In the current status, the search algorithm works on whole texts only. In the future, we will provide a list of text passages with a high density of related terms to further facilitate the scholar’s workflow. Furthermore, we will provide a transparent view of the results in the form of visual statistics illustrating the reasons for the ranking of a single text passage and the distribution of a concept’s related terms in general.

**References**

**Appleford, S. and Thatcher, J.** (2013). Using the Social Web to Explore the Online Discourse and Memory of the Civil War. In *Proceedings of the Digital Humanities 2013*, University of Nebraska–Lincoln, 16–19 July 2013.

**Giunchiglia, F., Kharkevich, U. and Zaihrayeu, I.** (2009). Concept Search. In *The Semantic Web: Research and Applications*. Berlin: Springer, pp. 429–44.

**Guha, R., McCool, R. and Miller, E.** (2003). Semantic Search. In *Proceedings of the 12th International Conference on World Wide Web*, Budapest, 20–24 May 2003, pp. 700–709.

**Novak, J. D. and Cañas, A. J.** (2006). The Theory Underlying Concept Maps and How to Construct and Use Them. Florida Institute for Human and Machine Cognition.

**Park, J. and Hunting, S., with Foreword by Engelbart, D. C.** (2002). *XML Topic Maps: Creating and Using Topic Maps for the Web.* Addison Wesley Longman, Boston.

**Teevan, J., Alvarado, C., Ackerman, M. S. and Karger, D. R.** (2004). The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria, 24–29 April 2004, pp. 415–422.