

A use-case in Digital Humanities – Some reflections on the project eAQUA

Gerhard Heyer

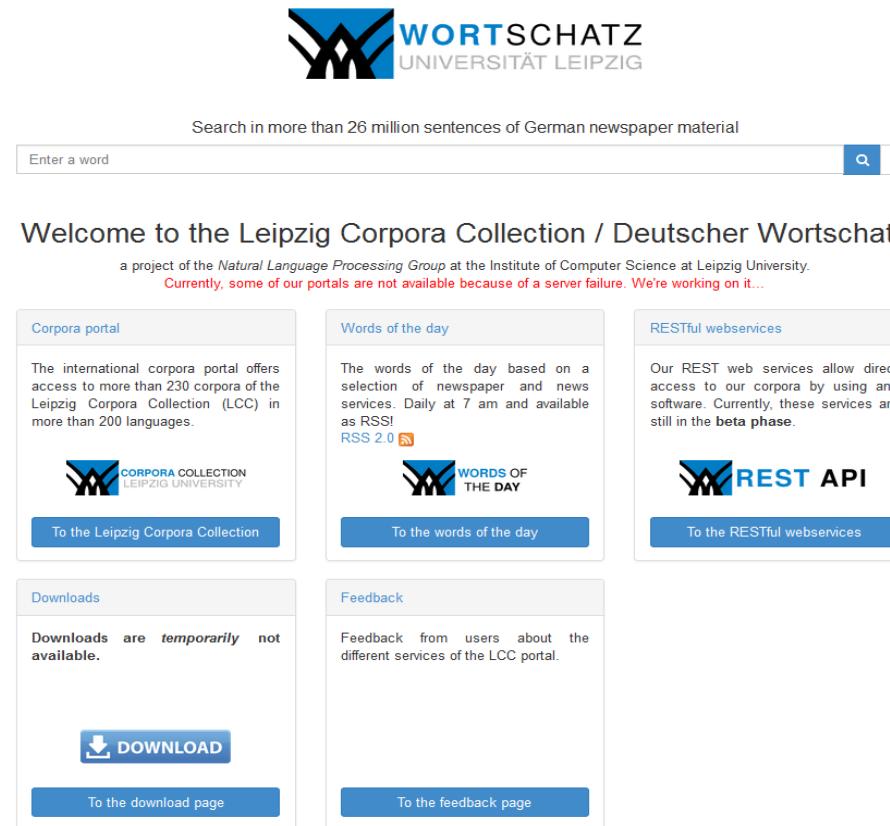
Universität Leipzig

heyer@informatik.uni-leipzig.de

The BMBF project eAQUA

- **Title**: „Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft“
- **Duration**: 1.4.2008 – 31.3.2011
- **Portal**: still running under <http://www.equa.net/>,
see also <https://de.wikipedia.org/wiki/EAQUA>
- **Participants**: ASV and classics department Leipzig University
- **Goals**: apply text mining to collections of classical text (Latin, Greek) in order to extract specific knowledge
- **Key developers**: Marco Büchler and Thomas Eckart

Wortschatz <http://wortschatz.uni-leipzig.de/en>



The screenshot shows the homepage of the Deutscher Wortschatz (German Vocabulary) website. At the top, there is a search bar with the placeholder "Enter a word" and a magnifying glass icon. Below the search bar, the text "Search in more than 26 million sentences of German newspaper material" is displayed. The main heading "Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz" is followed by a subtext: "a project of the *Natural Language Processing Group* at the Institute of Computer Science at Leipzig University. Currently, some of our portals are not available because of a server failure. We're working on it...".

The page features several sections with links:

- Corpora portal**: "The international corpora portal offers access to more than 230 corpora of the Leipzig Corpora Collection (LCC) in more than 200 languages." Includes a "To the Leipzig Corpora Collection" button.
- Words of the day**: "The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!" Includes an "RSS 2.0" link and a "To the words of the day" button.
- RESTful webservices**: "Our REST web services allow direct access to our corpora by using any software. Currently, these services are still in the beta phase." Includes a "To the RESTful webservices" button.
- Downloads**: "Downloads are *temporarily* not available." Includes a "DOWNLOAD" button and a "To the download page" button.
- Feedback**: "Feedback from users about the different services of the LCC portal." Includes a "To the feedback page" button.

At the bottom, a note states: "Data is automatically collected from carefully selected public sources. The example sentences are automatically selected and are not expression of this project. The authors are solely responsible for the content and opinions contained therein."

Corpora http://corpora.uni-leipzig.de/de?corpusId=pol_newscrawl_2011



Suche in 263 korpusbasierten monolingualen Wörterbüchern in 236 Sprachen.

🔍 ?

☒ Korpus: Polish (pol_newscrawl_2011)

Polish newspaper corpus based on material crawled in 2011
Sätze: 6,494,575 · Types: 1,549,233 · Tokens: 96,476,260 [mehr...](#)

[Deutsch](#) [Englisch](#) [Französisch](#) [Arabisch](#) [Russisch](#) [Alle...](#)

☒ Zufällige Wörter:

Kolejne [inicjatywy](#) [Miasta](#) [dojść](#) matka

Subprojects

- **Attidographers** – Classification of Attidographers
- **Plato** - Aftermath of Plato's writings
- **Plautine metric** – Metrical analysis of Latin comedies
- **CAMENA** - Knowledge network of Latin CAMENA corpus
- **Inscriptions** - Extraction of significant templates for different kinds of inscriptions like release documents (slave trading)
- **Papyri** - Classification of papyri (e. g. slave trading contracts) and text completion of fragmentary texts
- **Mental maps** - Building of century based mental maps and highlighting differences

Challenges

Challenges from Computer Science point of view

- **Integrating textual resources**

Conversion of different formats and standards

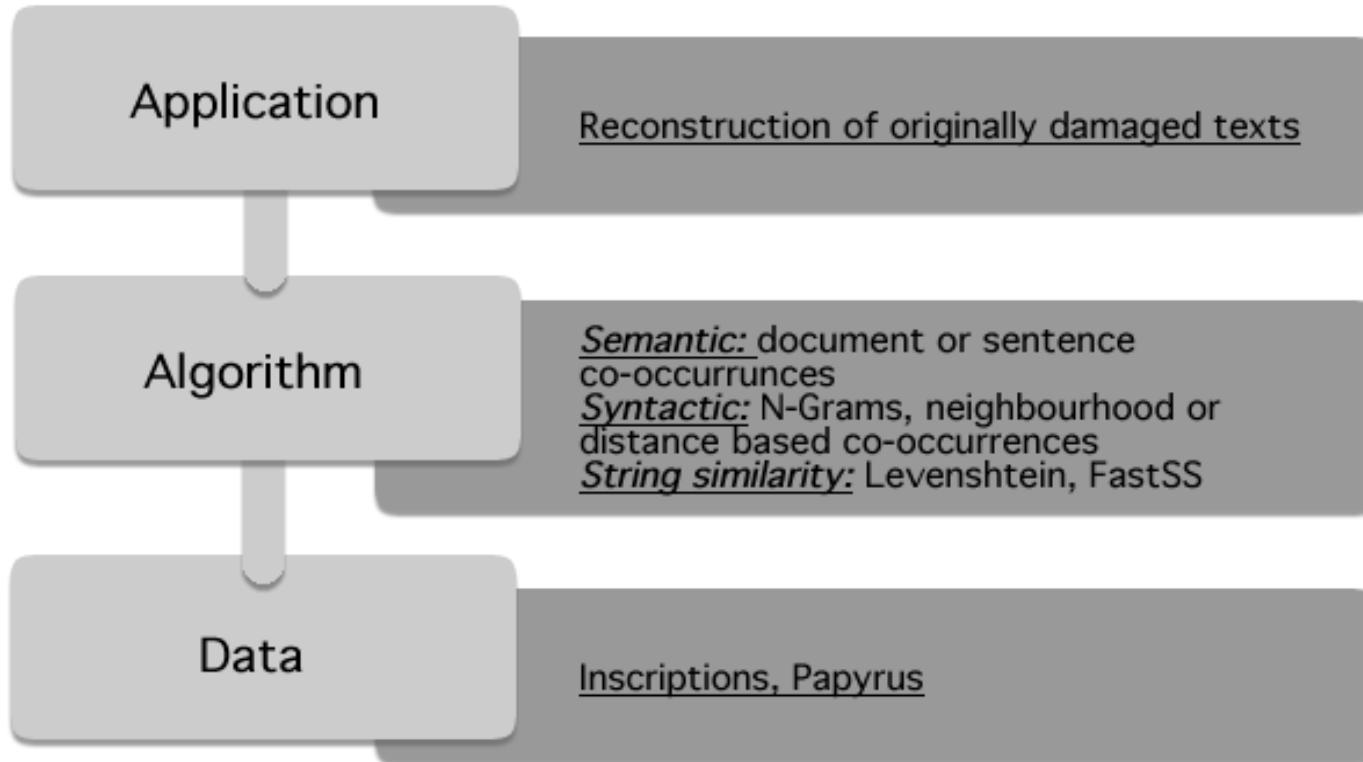
- **Software engineering issues**

Getting classicists to understand the intent and effects of text mining algorithms and clearly define their functional requirements

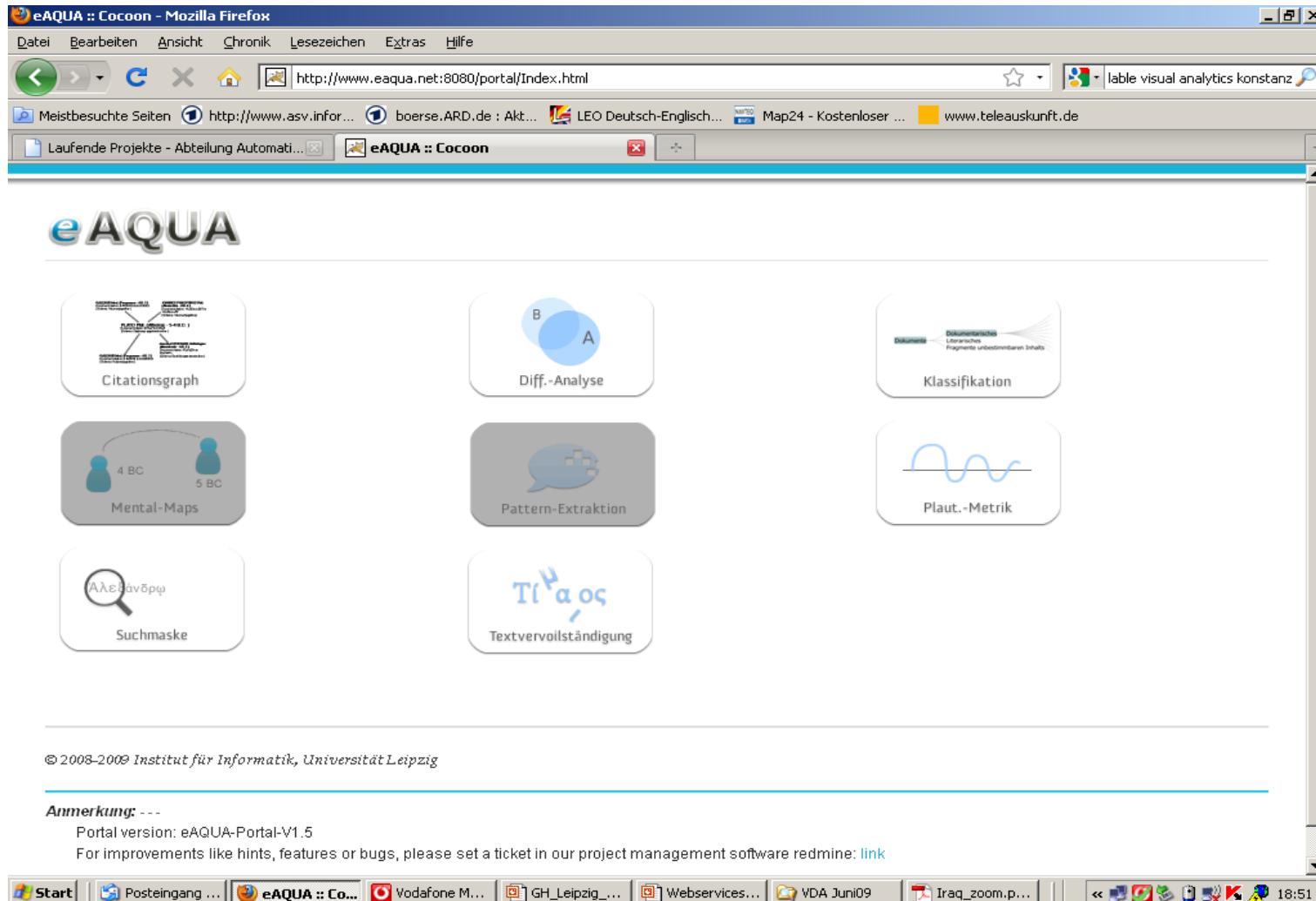
- **Semantic technologies that address the long tail**

Modifying and amending statistical and pattern based methods to effectively deal with rare events

Approach



Access to portal

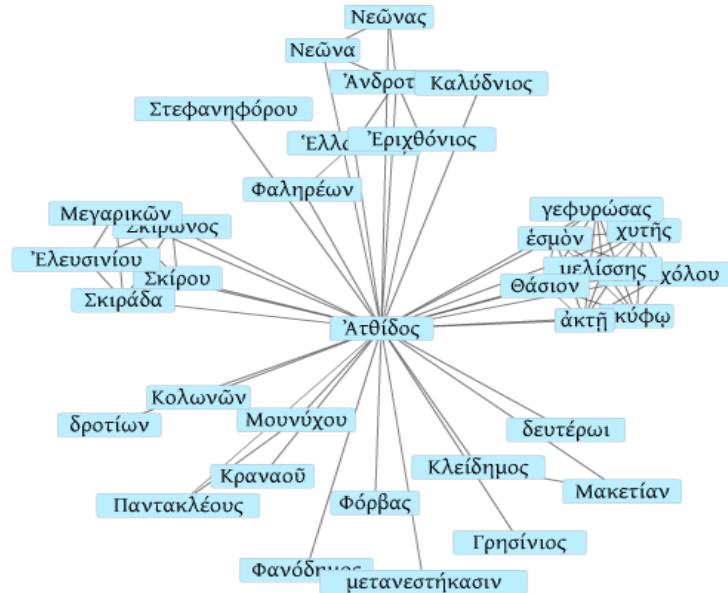


The screenshot shows a Mozilla Firefox browser window with the title "eAQUA :: Cocoon - Mozilla Firefox". The address bar displays "http://www.equa.net:8080/portal/Index.html". The main content area is titled "eAQUA" and features a grid of eight service icons:

- Citationsgraph**: A diagram showing a network of nodes connected by lines.
- Diff.-Analyse**: A diagram showing two overlapping circles labeled "B" and "A".
- Klassifikation**: A diagram showing a document icon pointing to a fan-like structure labeled "Liberarisch" and "Fragments unbestimmbaren Inhalts".
- Mental-Maps**: A diagram showing two people icons with "4 BC" and "5 BC" written next to them.
- Pattern-Extraktion**: A diagram showing a speech bubble icon.
- Plaut.-Metrik**: A diagram showing a wavy line icon.
- Suchmaske**: A diagram showing a magnifying glass icon over a Greek word.
- Textvervollständigung**: A diagram showing the word "Trieos" in blue and white.

At the bottom of the page, there is a copyright notice: "© 2008-2009 Institut für Informatik, Universität Leipzig". Below that, a "Anmerkung:" section contains the text: "Portal version: eAQUA-Portal-V1.5" and "For improvements like hints, features or bugs, please set a ticket in our project management software redmine: [link](#)". The taskbar at the bottom shows various open windows and system icons.

Example: Searching for „ΑΤΘΙΔΟΣ“



Co-occurrences to find „interesting“ neighbours

Signifikante rechte Kookkurrenzen für *Ἄτθιδος*

άκραχόλου (5); σκύφω (5); Παντακλέους (5); Μουνύχου (5); Φ (12); συγκυρκανήσας (4); Σκίρου (5); χυτῆς (5); Πραξίων (4); Ἐλευσινίου (5); Νεῶνας (4); Κραναοῦ (5); FHG (9); μελίσσης (6); γεφυρώσας (5); ἐσμὸν (5); Σκίρωνος (4); ἀκτῆ (5); μετανεστήκασιν (3); Μεγαρικῶν (4); Ἀτθίδος (5); Φαληρέων (3); μάντεως (5); καύματος (5); Ι (9); κεκλῆσθαι (5); ὡνομάσθαι (5); φησὶν (13); Βιβλία (6); Ἀρκεσινεῦσιν (2); Ἀνδανή (2); Ἀλικαρνάσσιος (2); Μινωίταις (2); κατησίμωκα (2); κατεστιλβωμένον (2); θουλῆς (2); ψιλοκιθαριστικήν (2); Δήμητρος (5); Μελίτη (3); δημιουργικάς (3); πῶμα (4); FGrHist (4); Ἀριστάρχειόν (2); Ἀμοργίοις (2); Ἐπίγονον (2); διαδικασίαι (2); Κλείδημος (3); νομάδας (3); γλυκύν (4); Μακετίας (2); Ἐλεγείαις (2); φησὶν (13); ἔξωκίσθησαν (2); γυναικονόμοι (2); ὄμωνύμωι (2); τινός (9); Κεκροπίδην (2); συγγενομένης (2); ἀπεχειροτονήθη (2); ἔάλωσάν (2); ἔής (2); Ἀγραυλον (2); Ἄλωπεκόννησον (2); διεξῆλθεν (3); ἐπεποιήκη (2); κιθάρισιν (2); μαγείρων (3); λύσιν (5); λίθου (5); πειθήμονα (2); κεδνῆ (2); ὄμογάλακτας (2); θέσμιον (2); Αίγιαλὸν (2); Κήρυξ (2); Ἀλιρρόθιον (2); κόβαλον (2); εύογκον (2); σποράδην (3); Θυμοίτης (2); Άλκιπηπην (2); δράματι (3); βωμολόχον (2); νομισθὲν (2); Εύδοξος (3); διαλέκτου (3); αὐτόχθονος (2); Ἀρεοπαγιτῶν (2); Δειναρχος (3); οἰκήσεις (3); θεωρικὸν (2); κιθαριστῆς (2); ἀπέτισε (2); θυγατρός (3); ἡγεμονεύων (2); Ξούθου (2); Ἐριχθονίου (2); φῦλον (3); Κεραμεικῷ (2); ιδίωμα (3);

- νομάδας is an interesting candidate
- Reference to quotation from the *Atthis* of Philochorus (3. / 2. century BC) in a normal word search hard to find
- Co-occurrence indicates that attidographers may have invented a phase in Athenian history to show an evolutionary development from *nomadism* to settled life

Citation analysis: Plato and Neo-Platonism



Assessment – What can we learn from the project?

- **What we did well**

- Sustainable portal, well documented
- Demonstration of feasibility to apply text mining in the domain of classics

[Marco Büchler, Gerhard Heyer und Sabine Gründer: eAQUA - Bringing modern Text Mining approaches to two thousand years old ancient texts. In: Proceedings of the 4th International Conference on e-Science (IEEE08), 2008]

Assessment – What can we learn from the project?

What we did not do well

- Balanced integration of requirements analysis and implementation
 - Relying on waterfall model of SE resulted in unclear specifications, ill-defined roles, and substantial delays
 - Agile paradigm would have been much more appropriate

Assessment – What can we learn from the project?

What we did not do well

- Apply tools available at that time without designing solutions to particular research questions
 - co-occurrences, differential analysis, ...
 - no word embeddings, topic models, vector space semantics, ...
- Taking tools offered by CS as a solution to research questions instead of a challenge to redesign the research question and rethink the research question in terms of computational requirements

Thanks for listening ...

Thanks for listening ...



Thanks for listening ...



Questions please!