eAQUA - Bringing modern Text Mining appraoches to two thousand years old ancient texts

Marco Buechler, Gerhard Heyer and Sabine Gründer
Natural Language Processing Group
Institute of Information Science
University of Leipzig Johannisgasse 26, 04103 Leipzig, Germany
{mbuechler, heyer, sgruender}@informatik.uni-leipzig.de

Abstract

In this paper we give an overview of our work on a new research project, which brings together ancient texts and modern methods from the field of text mining. The project is structured so that is comprises data, algorithms, and applications. In this paper we first give a short introduction of the current state of the art. After that we describe what eAQUA will do and what is our methodology.

1. Introduction

In the last years numerous digitalizing projects were funded. As a result of this digitalisation wave there now exist vast numbers of small corpora. Basically, the digitalisation process stopped when the objects were scanned and stored as an image. Other projects used OCR software to extract text from the pictures. Currently, there seem to be as many data formats or user interfaces as there are use case scenarios. Often then, the tools and ressources are designed and suitable just for a dedicated research question. Typically, researchers in the field of Classics are working with more than one tool or web site at once and have to frequently switch between them.

Nowadays the usage of digital sources in Classical Studies is mostly restricted to looking for e. g. dedicated phrases, text positions or names. But beyond these typical Information Retrieval applications, it would be promising to use improvements in information technologies for an advanced study of digital text resources and for generating structured knowledge resources for the community of Classical Studies.

The **eAQUA** project (Extracion of structured knowledge from Ancient Resources for the Classical Studies [19]) aims at generating specific knowledge from ancient texts and wants to provide this knowledge via an open eAQUA portal.

eAQUA itself has only a small number of own corpora. For this reason the portal will use a standardized data interface for importing other corpora. The main focus of this project is to break down research questions from the field of Classics in a reuseable format fitting with NLP algorithms and to apply this type of approach to the data from the Ancient sources.

That means, the surplus that eAQUA aims to bring to the field of Classical Studies should not consist in providing a search engine for a special text database. Rather, our main aim is to find out how concrete research questions can be operationalized to existing tools of the NLP community.
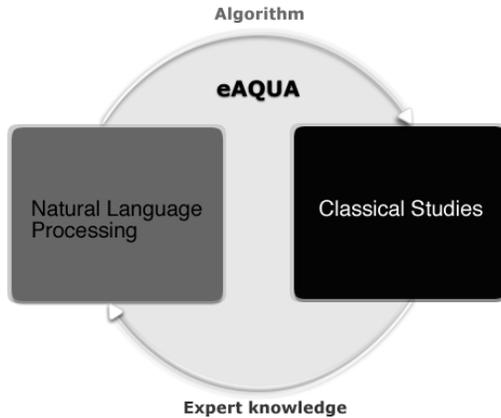
A bridge between Classical Studies and Information Science must be grounded on a lot of interaction and efforts on both sides. In figure 1 a plan for a process of iterative interaction is shown. This interaction between both sides faces many challenges coming up from different perspectives and working methods of researchers on both sides.

2. State of the art

In the past years a huge number of digitalizing projects (see section 3.3) spread out a lot of data. Tools like Diogenes [1], View & Find [24] and Lector [23] can partially use these data. Most of these tools are standalone and local applications. More modern projects like Hopper [15] and the Archimedes Project [25, 2] are server based. Some of the tools can be accessed for free, others have a fee.

Most applications only give access to the data via an user interface. Some of these projects like Hopper [15] are using modern technologies from the information retrieval field. Other tools are still using full text searches.

The Perseus Project is currently one of most active projects in the community of Classical Studies. Besides of offering a large amount of text data, it will provide to the community tools like Morpheus (a morphological analyser for ancient Greek and Latin) or the Latin Treebank [11, 10].

Figure 1. Iterative interaction cycle between two different research methods: Classical Studies and Information Science

Figure 2. The approach comprising data, algorithms, and applications shortly described and applied to the correction of the damaged papyrus of figure 3.

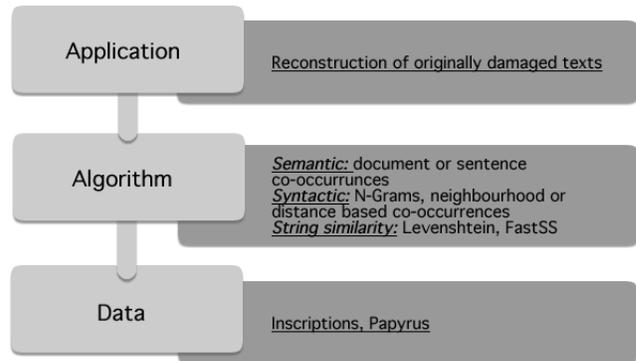## 3. Data, Algorithm, Application

Based on existing resources and current research questions there can be fixed a simple vision. It would consists of an unified approach comprising Data, Algorithms and Applications. eAQUA addresses both, algorithms and applications. The Data will be seen as existing and will be imported by eAQUA's standardized interface.

- *Application:* In the eAQUA context, application can more scientifically be understood as a research question.

- *Algorithm*: Algorithms from the field of NLP, Picture Mining or text picture analyses will be used to operationalize the research questions (e. g. Co-occurences analsysis, N-gram statistics, clustering and classification approaches, time series analysis or simple citation usages).

- *Data*: Text and picture data which were digitalized in different projects before (e. g. Perseus corpora (Boston), Anarche (Cologne, Germany), Camena Termini (Heidelberg, Germany)).

Figure 2 represents a simple instance of this for the reconstruction or the respective correction of a papyrus or an inscription.

3.1. Application

Research in Classical Studies are often data driven, which means that researchers are looking for phrases or words, collecting such results and interpreting them. The working method of information scientists is more model driven. Dedicated text fragments or passages fit to a model or they have a low significance and will therefore be truncated.

eAQUA needs a trade-off of both working methods. On the one hand, Classics researches needs access to digital libraries that is faster in comparison with accessing printed versions. On the other hand, more model driven analytics can give a more formal point of view to research questions.

Inside eAQUA user interfaces will be built to support research questions in the fields of

- Semantic reconstruction of lost works of the Atthidographs and their classifications,

- After-effects of Platon's work: How are phrases of Platon cited directly or as paraphrases afterwards?

- Classification of papyri (e. g. slave trading contracts),

- Extraction of significant templates for different kinds of inscriptions like release documents,

- Plautin metric: What are good models for rhythm metrics of Latin comedies?

- Text completion of fragmentary texts.

3.2. Algorithms

In the field of algorithms, rule based and probabilistic approaches are both used. Rule based approaches have already strong support in the Classical Studies community.
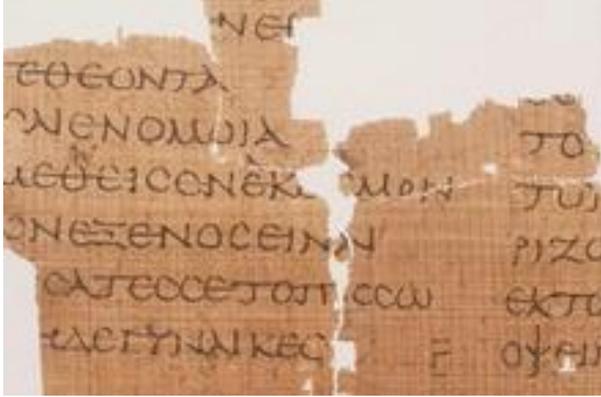
Figure 3. A damaged papyrus which needs reconstruction of lost fragments.

But a stronger focus will be set on the probabilistic approaches [18].

Tools like the ASV Toolbox [12] or Medusa [13, 14] will be used to break down the above mentioned research questions. In eAQUA algorithms from the following fields will be needed

- different kinds of time series analysis,

- detecting collocations and multi word expressions,

- different levels of computing paraphrases,

- semantic/context and string based spell checking for generating suggestions,

- feature and annotation based classification,

- hard and soft clustering (e. g. graph based clustering),

- co-occurrences for detecting semantic relations,

- distance based co-occurrences for computing phrases and templates,

- modified unsupervised POS-Tagger for tagging rhythm metrics.

3.3. Data

As a result of the digitalisation wave of the last years a lot of ancient data exists digitally. Unfortunately, most of these data differ in format. On the one hand, this addresses different media types like text or picture. On the other hand, texts (e. g. complete texts, inscriptions or papyrus) are mostly stored in proprietary and non standardized formats. One standard, which eAQUA will support, is the epiDOC standard [16]. epiDoc is based on TEI [4] and extends TEI P4

for epigraphic annotations like the Leiden Convention. In contrast to modern texts which are complete, ancient texts or fragments of texts have possibly been destroyed over time (see figure 3). Such missing parts of letters or words were annotated by the editors of the texts with the most probable letter or word. These annotations follow the Leiden Conventions.

Table 1 gives a small overview of the spectrum of ancient data collections.

Text collections like Camena or the TLG are very suitable for information retrieval or text mining algorithms because they mostly include non-damaged and corrected texts. More critical for the use of NLP algorithms are papyrus and inscriptions. The degree of demolition varies considerably. Consequently, the quality of the output of the NLP algorithms can considerably vary too.

Furthermore, table 1 reflected an interest of the community in pictures and real world objects (e. g. sherds or mintages as an abstact construction of those entities), too [17]. In this context meta data models like CIDOC-CRM [5] are used for the annotation of pictures or objects. The set of text types like inscriptions, papyri or pictures and objects can be extended by a virtual semantic representation of reconstructed texts based on indirect citations.

In face of this situation a platform like eAQUA needs to handle different text types at once. That is why a common represantation of all those different types is needed (see figure 4). Current researches in the field of distributed Topic Maps are highlighting approaches for identifying the same topic in different data resources. This identification is possible even if the name of the topic is not the same [22, 6].

With respect to tasks like citation of phrases or the semantic reconstruction of documents a solid database with semantic annotations (e. g. synonyms) like WordNet [7] or GermaNet [3] is needed. A specific feature of such a GraeceNet is the time dependency of word meanings in ancient texts (see more in section 3.4).

The start of GraeceNet is involved in current eAQUA activities. Beside semantic term expansion as described above, GraeceNet will also be used for a semantic search and for the purpose of evaluation. The importance of the latter aspect is due to the fact of there existing only rare data (e. g. Latin Treebank [11, 10]) for evaluation of NLP algorithms on ancient texts.

3.4. Challenges

Internally the eAQUA project we will handle data from 3,000 B.C. to 600 A.C. On the one hand, millions lines of text can be used for (probabilistic) text mining approaches. On the other hand, a time range of more than 3,500 years and a non-global civilisation at this time makes data very sparse. The role of geographical distances were completely

Name	Language	Media type
Arachne [17]	Greek	picture, objects
Bibliotheca Teubneriana Latina (BTL)	Latin	text
Camena [20]	Latin	text
Duke Data Bank of Documentary Papyri	Greek, Latin	papyrus
Göttinger Digitalisierungszentrum [21]	different languages	text
PHI5	Latin	text
PHI7	Greek	inscriptions, papyrus
Papyrus portal[26]	Greek	papyrus
Perseus texts [15]	Greek & Latin	text
Codex Sinaiticus [8]	Greek	Christian Bible texts
Thesaurus Linguae Graecae (TLG)	Greek	text

Table 1. A small overview to existing text and picture corpora.

different than today, in the sense that semantics and knowledge were locally bounded.

The time dependency of ancient data is critical too. In order to understand a word or a process of the 5th century A.C. the mental map of this time is needed. Words change their meanings and the ambiguity degree of different meanings over time. This change of meanings or of the relations between words can come up either as part of the evolution of a language or in result of artificial influences, e. g. a political system at a dedicated time. For projects like GraeceNet this strong time dependency of meanings must be taken into account.

Another problem for the application of text mining methods to ancient texts is the representation of characters. Different character distributions on the basis of the TLG and the PHI corpora have shown that about 3% of the characters (0.002% running characters) can not be displayed. Such characters can loosely be classified as musical and mathematical characters (acrophonic numerals) [27, 9].

Generally, tools like Hopper [15] or Diogenes [1] will be used in order to look for text positions of a word or phrase. Depending on iterative improvements in the field of algorithms the number of found text positions can change. For this reason a platform like eAQUA needs a mechanism we are calling *Saved Searches*. A Saved Search freezes a state of search result and store it under a own URL which a researcher can use for citation.

With NLP components improving over time, results set by Saved Searches would change. Hence, changes from the original Saved Search have to be incrementally displayed and can be compared along the time line. Additionally, a large enough set of Saved Searches can be used internally for evaluations of the used approaches.

3.5. Social Networking

It is commonly assumed that the community of ancient researchers is small and well arranged. But given the possibilities of modern computer technologies, new synergies between researcher seem to become possible. With respect to this, eAQUA will have built in a profiler for semantic interests of users. This tool will allow researchers to compare their profiles among each other and motivate concrete cooperation and scientific exchange. Of course this idea is not just applicable to data or research terms, but adaptable to the use of algorithms, algorithm combinations or data algorithm usages too.

4. Conclusions

In a pre-computer time researchers of Classical Studies tried to solve research questions (application) concerning ancient texts or pictures (data) completely manual. Caused by the increasing coverage of computer technologies and several digitalizing projects this working method dislocates to the computer.

A manual search on a corpus is not sufficient for giving answers to more complex research questions. Tools from the NLP field can be used to support scientific work here. Hence, more complex, unified approaches to Data, Algorithms, and Applications are needed. For the development of such an approach two completely different research methods or strategies of researchers of Classical Studies and Information Science have to be combined.

In the eAQUA project an iterative interaction cycle between both Classics researches and information scientists will be established, as described in figure 1. This interface is strongly forced by social and project management tasks. A second interface between algorithms and data is a more technical one. This interface consists of standardisation of different text formats, epigraphic annotations (e. g. if frag-
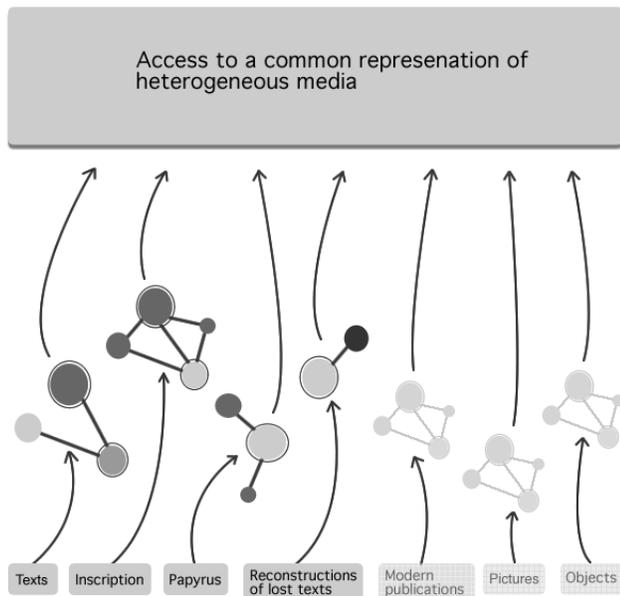
Figure 4. Different types of texts like papyrus, inscriptions or semantic reconstructions of lost documents will be stored in a common representation layer (e. g. Topic Maps). The user will access this semantic layer. Alongside texts, different medias like Arachne's pictures and objects could also be part of the search results.

ments of words are lost) and a common representation of different media types like pictures.

References

[1] Diogenes. World Wide Web electronic publication, 1999. http://freshmeat.net/projects/diogenes/.

[2] The Archimedes Project. World Wide Web electronic publication, oct 2004. http://archimedes2.mpiwg-berlin.mpg.de/archimedes_templates.

[3] Germanet. World Wide Web electronic publication, sep 2008. http://www.sfs.uni-tuebingen.de/lsd/.

[4] TEI: Text Encoding Initiative. World Wide Web electronic publication, sep 2008. http://www.tei-c.org/index.xml.

[5] The CIDOC CRM - Conceptual Reference Model. World Wide Web electronic publication, sep 2008. http://cidoc.ics.forth.gr/.

[6] Topic Maps Lab. World Wide Web electronic publication, sep 2008. http://www.topicmapslab.de/.

[7] Wordnet - Princeton University Cognitive Science Laboratory. World Wide Web electronic publication, sep 2008. http://wordnet.princeton.edu/.

[8] Codex Sinaiticus. World Wide Web electronic publication, sep 2009. http://www.codex-sinaiticus.net.

[9] Thesaurus Linguae Graecae - Unicode Project. World Wide Web electronic publication, sep 2009. http://repositories.cdlib.org/tlg/unicode/.

[10] D. Bamman and G. Crane. The Latin Dependency Treebank in a Cultural Heritage Digital Library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 33–40, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[11] D. Bamman and G. Crane. The Latin and Ancient Greek Dependency Treebanks. World Wide Web electronic publication, sep 2009. http://nlp.perseus.tufts.edu/syntax/treebank/.

[12] C. Biemann. Toolbox Homepage. World Wide Web electronic publication, jan 2007. http://wortschatz.uni-leipzig.de/ cbiemann/software/toolbox/index.htm.

[13] M. Buechler. Flexibles berechnen von kookkurrenzen auf strukturierten und unstrukturierten daten. Master's thesis, Leipzig University, Germany, July 2006.

[14] M. Buechler. Medusa Release Homepage. World Wide Web electronic publication, jan 2006. http://aspra25.informatik.uni-leipzig.de/medusa/.

[15] G. Crane. The Perseus Project. World Wide Web electronic publication, 1985. http://www.perseus.tufts.edu/hopper/.

[16] T. Elliott. EpiDoc: Epigraphic Documents in TEI XML. World Wide Web electronic publication, sep 2008. http://epidoc.sourceforge.net/.

[17] R. Foertsch. Arachne - Objektdatenbank und kulturelle Archive des Forschungsarchivs für Antike Plastik Köln und des Deutschen Archäologischen Instituts. World Wide Web electronic publication, aug 2008. http://www.arachne.uni-koeln.de/.

[18] G. Heyer, U. Quasthoff, and T. Wittig. *Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse.* W3L-Verlag, Herdecke, Bochum, 2006.

[19] G. Heyer and C. Schubert. eAQUA - Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft. World Wide Web electronic publication, apr 2008. http://www.eaqua.net/.

[20] W. Kühlmann. TERMINI - Vernetzter Wortschatz lateinischer Wissensliteratur der Frühen Neuzeit. World Wide Web electronic publication, oct 2004. http://www.uni-mannheim.de/mateo/termini/index.html.

[21] N. Lossau. Göttinger Digitalisierungszentrum: GDZ. World Wide Web electronic publication, aug 2008. http://gdz.sub.uni-goettingen.de/.

[22] L. Maicher. *Autonome Topic Maps - Zur dezentralen Erstellung von implizit und explizit vernetzten Topic Maps in semantisch heterogenen Umgebungen.* PhD thesis, Natural Language Processing Department, University of Leipzig, Leipzig, Germany, 2007.

[23] R. Maier. Lector 2007. World Wide Web electronic publication, 2007. http://www.maierphil.de/lector/.

[24] B. Meißner. SPITBOL Programming by/for Classicists: Accessing and Analyzing Classical Texts . World Wide Web electronic publication, 1995. http://www2.altertum.uni-halle.de/v_f.html.

[25] M. Schiefsky. Archimedes Project. World Wide Web electronic publication, oct 2004. http://archimedes.fas.harvard.edu/.

[26] R. Scholl. Papyrusportal. World Wide Web electronic publication, aug 2008. http://www.papyrusportal.de/.

[27] I. Unicode. The Unicode Character Code Charts By Script. World Wide Web electronic publication, sep 2009. http://www.unicode.org/charts/.