Neue Methoden der geisteswissenschaftlichen Forschung – Eine Einführung in das Portal eAQUA

New Methods in the Humanities – The Portal eAQUA

Charlotte Schubert, Gerhard Heyer

Die folgenden Beiträge sollen einen Einblick in die Arbeit eines laufenden Projektes geben: in das Projekt eAQUA (**E**xtraktion von strukturiertem Wissen aus **A**ntiken **Qu**ellen für die **A**ltertumswissenschaft), das im Rahmen des BMBF Förderschwerpunktes „Geistes- und Sozialwissenschaften“ im Programm „Wechselwirkungen zwischen Geistes- und Naturwissenschaften“ seit 2008 gefördert wird. Ziel dieses Programms ist es, dass „in diesem Förderschwerpunkt [...] geistes- und naturwissenschaftliche Fächer in interdisziplinären Forschungsverbünden zusammen arbeiten und sich gegenseitig bereichern.“ Themen dieses Förderschwerpunktes sollen u.a. „die Dynamisierung und Synthetisierung von bestehendem und zukünftigem Wissen“, „Fortschritte in der Dokumentation von Schichten von Wissensordnungen“, die „Analyse von Wissenssystemen“ oder auch „Prozesse, die zur Entstehung von Texten führen“ sein.[1]

In dem Projektverbund eAQUA arbeiten Altertumswissenschaftler, Frühneuzeitler und Informatiker zusammen, um die Anwendung fortgeschrittener Werkzeuge aus dem Bereich des Text Mining für die beteiligten Fachdisziplinen erstmals experimentell zu erproben. Die Werkzeuge des Text Mining gehen über übliche Suchmöglichkeiten hinaus, indem sie semantische Zusammenhänge aufzeigen und ermöglichen eine schnelle Erschließung von Abhängigkeiten, Einflüssen und Transferwegen des Wissens in großem Umfang.

Diese Zusammenarbeit ist Teil eines größeren Transformationsprozesses der textbasiert arbeitenden Geisteswissenschaften. Der Bereich dieser wissenschaftlichen Produktion wird meist ‚Digital Humanities’ bzw. im Bereich der Altertumswissenschaften ‚Digital Classics’ genannt (vgl. Siemens/Schreibmann 2007). Zugespitzt kann man die Veränderung, die in dieser Bezeichnung zum Ausdruck kommt, durchaus so beschreiben, dass wir hier einen Umbruch von den non-digital zu den digital humanities vor uns haben, der natürlich auch ein Anwendungsfall von ‚contested order’ ist, insofern Berechtigung und Akzeptanz von Weg, Umbruch und Ziel bisher weder konstituiert noch methodisch etabliert sind. Der für die heutige geisteswissenschaftliche Forschung unhintergehbare Standard einer selbstreflexiven Verortung ist dafür noch nicht in Sicht. Es wäre vermessen, aus einem noch laufenden Forschungsprojekt heraus den Anspruch zu erheben, dies leisten zu können. Aber wir betrachten unsere hier vorgelegten Working Papers als Zeugnisse für den großen "struggle about order resp. the good order in the humanities" unter den Bedingungen des 21. Jahrhunderts innerhalb des hier sichtbar werdenden Transformationsprozesses der Geisteswissenschaften.

Der Projektverbund

Teil der Projektarbeit ist der Aufbau und die Entwicklung des Portals, dessen Anwendungsmöglichkeiten und Perspektiven hier aus der Perspektive der beteiligten geisteswissenschaftlichen Disziplinen im Rahmen der Working Paper Series als Arbeitsstand präsentiert werden. Konkreter Gegenstand des Projektes ist die Erarbeitung von spezifischen Methoden für die historischen Sprachen Griechisch und Latein, die einen Einsatz des Text Mining in den Altertumswissenschaften erlauben. So sind neue inhaltliche Zusammenhänge ebenso wie Ähnlichkeiten von Texten und Begriffen zu finden, um etwa die Autorenschaft oder Referenzen

[1] Vgl. Ausschreibung des BMBF. http://www.bmbf.de/foerderungen/7774.php (19/11/10).

und Zitate zu bestimmen. Insbesondere für nur fragmentarisch erhaltene Texte ist auf diesem Weg eine Präzisierung des Textbestandes selbst zu erwarten, indem Fragmente entweder ergänzt werden oder weitere hinzugefügt werden können.

Das Verbundprojekt eAQUA umfasst mehrere Teilprojekte, die mit ganz unterschiedlichen Quellencorpora und entsprechend unterschiedlichen Methoden arbeiten. Aus folgenden Teilprojekten werden hier Anwendungsmöglichkeiten präsentiert:

4.1 Projekt Atthidographen (Teilprojektleiterin: Ch. Schubert, wiss. Mitarbeiter: A. Bünte),

4.2 Projekt Platon (Teilprojektleiter: K. Sier, wiss. Mitarbeiterin: A. Geßner),

4.3 Projekt Metrik (Teilprojektleiter: M. Deufert, wiss. Mitarbeiter: J. Blumenstein / J. F. Gärtner),

4.4 Projekt Camena (Teilprojektleiter: W. Kuhlmann, wiss. Mitarbeiter: R. Gruhl),

4.5 Projekt Papyri (Teilprojektleiter: R. Scholl, wiss. Mitarbeiterin: M. Rücker),

4.6 Projekt Mental Maps (Teilprojektleiter: Ch. Schubert, wiss. Mitarbeiterin: R. Kath).

Aufgrund ihrer seit längerem bereits sehr weitgehend digitalisierten Quellencorpora können die textbasierten Altertumswissenschaften – in dem Forschungsverbund arbeiten Latinisten, Gräzisten, Althistoriker, Papyrologen, Epigraphiker – sowohl methodisch in der Anwendung dieser Verfahren als auch im Hinblick auf den Transfer in die Lehre eine Vorreiterrolle für alle Geisteswissenschaften übernehmen.

Bisher beschränkt sich die Nutzung der digitalen Quellen meist darauf z. B. nach bestimmten Formulierungen, Textstellen oder Namen zu suchen. Wir möchten jedoch anhand der hier präsentierten Einblicke in unsere laufende Arbeit zeigen, dass die Fortschritte in der Informationstechnologie eine weitergehende Nutzung der digitalen Textquellen als Rohstoff für die Schaffung von strukturiertem Wissen ermöglichen. Das Ziel unseres Projektes ist es, für die Altertumswissenschaft aus antiken Quellen spezifisches Wissen zu generieren und nach Abschluß des Projektes über ein Web-Portal (http://www.eaqua.net/portal/) unsere Methoden der praktischen Forschung nachhaltig zur Verfügung zu stellen. Dafür wird derzeit in der Kooperation zwischen Altertumswissenschaftlern und Informatikern die heute verfügbare *Text Mining Technologie* den Bedürfnissen und Anforderungen der Altertumswissenschaft angepasst. Im Ergebnis soll Folgendes erreicht werden:

- Nutzung von spezifischen digitalen antiken Textquellen,
- Nutzung geeigneter Text Mining Verfahren für die Generierung von strukturiertem Wissen wie beispielsweise Wissensnetzen für die Altertumswissenschaft,
- Bereitstellung des so generierten Wissens, darauf aufbauender Dokumentationen ihrer Nutzung als *best practice*,
- Iterative Anreicherung und Verbesserung der zugrundegelegten digitalen Textcorpora durch die Nutzung von projektspezifischen Ergebnissen.

Im Hinblick auf die beiden zuletzt genannten Ziele sollen die hier präsentierten Working Papers der erste Schritt sein.

Architektur des Portals eAQUA und Verfahren

Im Rahmen des Projektes ist eine Cocoon-basierte Architektur gewählt worden. Cocoon arbeitet komplett in der internen und externen Darstellung mit XML. Dadurch können Texte sowohl in TEI P5 epiDoc nicht nur dargestellt, sondern auch bereits existierende XSLT-Transformatoren wiederverwendet werden, um die Texte im Browser darzustellen.

In der internen Architektur treffen am deutlichsten die Wissenschaftswelten der Classical Studies und der Informatik aufeinander. Während Texte mit ihren vielfältigen Annotationen

wie heterogenen Dokumentstrukturen (Buch vs. Inschrift/Papyrus) oder dem textkritischen Apparat in XML gespeichert werden, können die Text Mining Daten aufgrund der Menge aus Performanz- und Lesbarkeitsgründen nicht in XML abgelegt werden. Die Cocoon-basierte Architektur bietet jedoch die Möglichkeit, je nach Anwendung die richtigen Texte mit den Text Mining Daten zu verknüpfen.

Da insbesondere das Preprocessing und das Aufbereiten der Textdaten oft sehr zeitaufwendig ist (nach zwei Jahren haben wir in unserem Projekt etwa 16 Mannmonate aufgewendet), werden sowohl SOAP- als auch REST-basierte Services angeboten, welche den Zugriff auf Daten als auch entsprechende Algorithmen und Tools ermöglichen.

Das eAQUA-Portal bietet anders als die meisten Portale eine reine REST-basierte Umgebung. Dies heißt, dass das gesamte Portal sowohl vom Menschen als auch von anderen Computern bzw. Portalen genutzt werden kann. Ein entsprechendes Nutzer-Interface wird derzeit auf der Rich-Client-Technologie Flash entwickelt. Es fragt die Daten auf die gleiche Weise vom Server ab, wie es auch ein anderer Computer in einer serviceorientierten Umgebung tun würde.

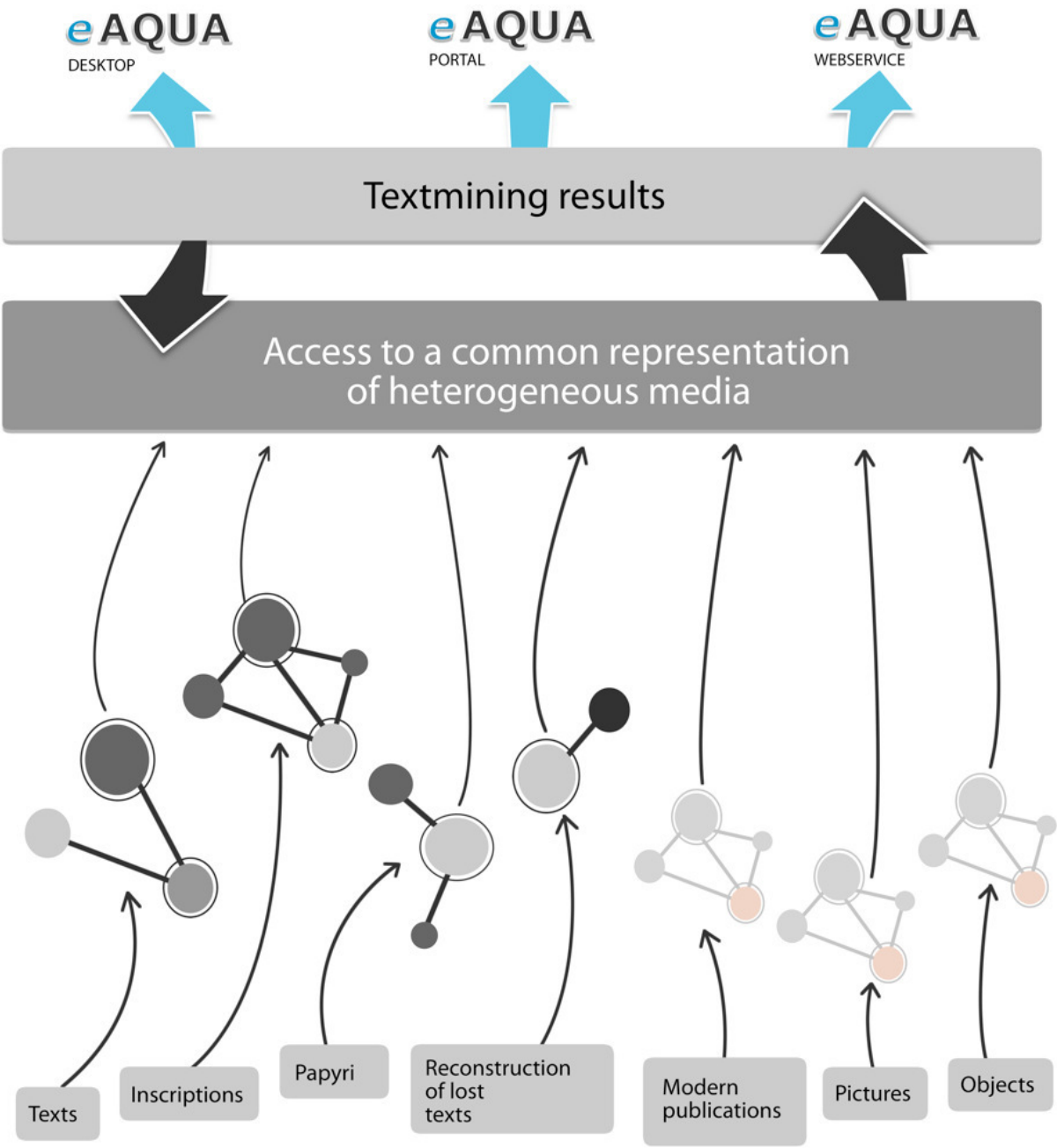

Abb. 1

Der Einsatz von Rich-Client-Technologien hat sich gegenüber reinen HTML-Seiten als sehr vorteilhaft erwiesen, da sowohl grafische Darstellungen als auch Benutzerinteraktionen besser umgesetzt werden können.

Neben zahlreichen Verfahren aus den einzelnen Teilprojekten, auf die im Zusammenhang dieser Teilprojekte im Detail eingegangen wird, liegt der Hauptfokus der Arbeit im Bereich der Automatischen Sprachverarbeitung auf den Ergebnissen aus dem Bereich Preprocessing bzw. Data Preparation, Semantik sowie Syntax (Vgl. Büchler, Heyer, Gründer 2008).

Im Bereich des Preprocessings ist eine ANT-basierte Toolchain entwickelt worden, die die Aufbereitung der Texte wie die Extraktion des Textes, das Segmentieren der Sätze oder auch das Tokenisieren der Wörter von einem Urzustand bis hin zur fertigen Datenbank übernimmt. Speziell für altgriechische Texte ist es somit möglich, große Teile dieser Toolchain auf neue Korpora anzuwenden.

Aus semantischer Sicht konnte eine wichtige Wechselwirkung aufgenommen werden, die sich daraus ergibt, dass für einen altertumswissenschaftlichen Fachwissenschaftler meist nicht das ohnehin Bekannte, statistisch Häufige im Vordergrund steht, sondern die seltenen Vorkommen von Erwähnungen und Belegstellen. In der praktischen Arbeit gibt es eine große Differenz zwischen mathematischen Signifikanzmaßen und nutzerspezifischer Relevanz bei der Bewertung einer Assoziation zwischen zwei Wörtern. Daher sind in zwei Bereichen – den Latent Relations sowie dem Association Chaining – Grundlagenarbeiten durchgeführt worden, um einerseits ein graphbasiertes Signifikanzmaß zu entwickeln, welches den Unterschied der Bedeutungskontexte misst und andererseits indirekte Assoziationen zu bestimmen, da nicht jede Information immer direkt im Text beobachtet werden kann.

Für die erfolgreiche und auf Innovation ausgerichtete Arbeit in einem interdisziplinären Projekt zwischen der Informatik und den Geisteswissenschaften sind neben guten Verfahren, entsprechenden Zugriffsformen (Benutzerschnittstelle) auf Text Mining Daten, auch methodische und konzeptionelle Sensibilität sowie eine z.T. hohe Abstraktionsfähigkeit nötig.

Wechselwirkungen und Anwendungsbereiche

Entsprechende visuelle Zugriffsformen auf die Text Mining Daten haben sich als elementar wichtig herausgestellt. Im Rahmen der Projektarbeit hat sich mehr und mehr die folgende Arbeitsformel herauskristallisiert: Viel Text erzeugt noch mehr Text Mining Daten, welche durch eine dedizierte Benutzeroberfläche selektiert und aggregiert werden müssen. So können beispielsweise diverse Algorithmen der Textwiederverwendung auf Textdaten angewendet werden und dies graphisch visualisiert werden (Graph). Die Visualisierung der zeitlichen Wiederverwendung einer Textstelle eines bestimmten Autors oder die Anzeige der Textseiten eines Werkes, die besonders häufig zitiert worden sind (Macroview auf die Textwiederverwendung) sowie die Anzeige verschiedener Varianten des gleichen Zitates (Microview) in graphischer Form (CitationGraph, flash layout) sind Analyseinstrumente, deren Einsatz in den Altertumswissenschaften neu sind.

Die sich hieraus ergebenden Anwendungsmöglichkeiten werden in den Beiträgen von A. Bünte (Teilprojekt 4.1 Atthidographen) am Beispiel der Verwendung des Wortes „Atthis“, A. Geßner (Teilprojekt 4.2 Platon) am Beispiel des CitationGraph für Platons Timaios und von Ch. Schubert am Beispiel eines Vergleichs verschiedener Suchstrategien (vollständig erhaltene Texte vs. Fragmente) beschrieben. Eine größere Perspektive spannt der Beitrag von R. Gruhl aus dem Teilprojekt 4.4 (Camena) auf, der am Beispiel der Auswertung frühneuzeitlicher Texte über die Visualisierung durch den Graphen einen Weg zu dem „Wissensnetz der Frühen Neuzeit. Von der virtuellen Bibliothek zur virtuellen Enzyklopädie“ aufzeigt.

Eine andere innovative Entwicklung weist der Arbeitsbereich der Plautinschen Metrik auf: Dort war die ursprüngliche Idee einen POS-Tagger für die Versmaß-Analyse zu trainieren, jedoch zeigte sich schnell, dass Verfahren wie der Viterbi-Algorithmus mit einem kleinen darunter liegenden Gedächtnis nicht zwangsläufig Verse gut taggen können. Vielmehr muss ein mit einem herkömmlichen Tagger annotierter Vers noch nicht einmal dem Versmaß – bspw. dem jambischen Senar – entsprechen. Ob dies der Fall ist, kann erst entschieden werden, wenn der Vers komplett annotiert ist. Diese neue Anforderung erfordert fortschrittliche Techniken auf der einen Seite. Auf der anderen Seite ergeben sich für die ASV interessante Wechselwirkungen, da solche Algorithmen für das POS-Tagging wiederverwendet werden können. Im Detail bedeutet dies, dass die Folge der benutzten POS-Tags erst am Ende des Satzes vergeben werden und nicht auf Basis eines kleinen Gedächtnisses von ein oder zwei Wörtern. Dies beleuchtet der Beitrag von J. Blumenstein, M. Deufert, J.F. Gaertner über die **e**lektronische Analyse der plautinischen Sprechverse aus dem Teilprojekt 4.3.

Bei der Bearbeitung der Papyri sind bisher verschiedene Verfahren der Rechtschreibkontrolle der Textvervollständigung bereits implementiert, eingesetzt und evaluiert worden. Bis zum Projektende werden nach aktuellem Stand etwa 25 Verfahren bzw. Methoden zur Verfügung stehen. Neben den Ansätzen aus der Rechtschreibkontrolle wie der Wortähnlichkeit, der syntaktischen oder semantischen Ähnlichkeit werden aktuell auch Methoden aus den Alter-

tumswissenschaften wie Datierungen oder Fundorte sowie Klassifizierungen benutzt, um auf das vorgeschlagene Wort Einfluss zu nehmen. Den Stand dieser Entwicklungsarbeit zu den Möglichkeiten der automatischen Textergänzung auf Papyri beschreibt M. Rücker (Teilprojekt 4.6 Papyri) in ihrem Beitrag.

Auch die Arbeit an dem Konzept der Volatilität (Visualisierung von Bedeutungsverschiebungen in großen diachronen Dokumentkollektionen) ist ein Feld, in dem sich die Perspektive zeigt, die sich aus der Wechselwirkung zwischen den Geisteswissenschaften und der Informatik ergibt: Die „Volatilität" von Konzepten, so wie sie von der Seite der Informatik im Sinne der Visual Analytics untersucht wird, hat beispielsweise ihre Entsprechung in den geisteswissenschaftlichen Ansätzen zur Deutungsmacht, die sich mit den diskursiven und symbolischen Deutungskämpfen um Begriffe oder historischen Ereignisse beschäftigen. R. Kath untersucht dies am Beispiel des Konzepts des „einfachen Lebens" in der Antike (Teilprojekt 4.8 Mental Maps).

Charlotte Schubert
Lehrstuhl für Alte Geschichte, Historisches Seminar, Universität Leipzig
schubert@uni-leipzig.de

Gerhard Heyer
Abteilung für Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig
gheyer@eaqua.net

Literatur

Bekanntmachung von Förderrichtlinien des Bundesministeriums für Bildung und Forschung (BMBF) „Wechselwirkungen zwischen Natur- und Geisteswissenschaften". http://www.bmbf.de/foerderungen/7774.php (19/11/10).

Büchler, M / Heyer, G. / Gründer, S. (2008) Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts. In: *e-Humanities an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science*.

eAQUA-Portal. http://www.eaqua.net/portal/ (19/11/10).

Siemens, R. / Schreibman, S. (2007) *A Companion to Digital Literary Studies*. Oxford: Blackwell.