Anna Aurast*, Tobias Gradl, Stefan Pernes und Steffen Pielström

Big Data und Smart Data in den Geisteswissenschaften

DOI 10.1515/bfp-2016-0033

Zusammenfassung: Die Verbindung von großen Datenbeständen und geeigneten quantitativen Verfahren ermöglicht es, in den Geisteswissenschaften neue Fragestellungen zu formulieren und bereits bekannte Fragestellungen neu zu denken. Im Kontext des Cluster 5 wird die Möglichkeit genutzt, philologisch-kritische und historisch-vergleichende Perspektiven in die Modellierung großer Datenbestände einfließen zu lassen und in weiterer Folge neue, auf geisteswissenschaftliche Fragestellungen zugeschnittene Analysewerkzeuge zu entwickeln. Das beinhaltet auch die Möglichkeit der empirischen Überprüfung theoretischer Vorannahmen und Kategoriebildungen sowie eine Erweiterung und Veränderung der Modi akademischer Wissensproduktion – durch eine solche Modellierung wird schließlich auch eine unmittelbare Anwendbarkeit und Erfahrbarkeit geisteswissenschaftlicher Begrifflichkeiten möglich. Der Beitrag beschreibt zwei in Cluster 5 entwickelte Analysewerkzeuge: Dariah-DKPro-Wrapper, ein Softwarepaket zur automatischen linguistischen Analyse, zugeschnitten insbesondere auf literarische Texte in Buchlänge, sowie Cosmotool, ein Such- und Analysewerkzeug zur Exploration grenzübergreifender Lebensläufe anhand strukturierter und unstrukturierter Datenbestände.

Schlüsselwörter: Textanalyse; unstrukturierte Daten; Natural Language Processing

Big Data and Smart Data in the Humanities

Abstract: With networked large data collections and suitable quantitative methods new research questions become possible. Within the DARIAH-DE cluster 5 new philologico-critical and historical comparing perspectives enter the modelling of large data collections. Consequently new, tailored analytical tools can be developed. This includes the chance to test in empirical ways theoretical presuppositions and the extension and advancement of the modes of the production of scientific knowledge. The article describes two analytical tools developed within cluster 5: the Dariah-DKPro-Wrapper, a software package for the automated linguistical analysis, and Cosmotool, a search- and analysis tool for the exploration of transboundary lives by means of structured and unstructured data collections.

Keywords: Text analysis; unstructured data; natural language processing

***Kontaktperson: Anna Aurast,** aurast@ieg-mainz.de
Tobias Gradl, tobias.gradl@uni-bamberg.de
Stefan Pernes, stefan.pernes@uni-wuerzburg.de
Steffen Pielström, pielstroem@biozentrum.uni-wuerzburg.de

Die Einrichtung des Cluster 5, *Big Data in den Geisteswissenschaften*, erfolgt zu einem Zeitpunkt, der einen doppelten Wendepunkt darstellt: Einerseits stehen in den Geisteswissenschaften zunehmend große Datenbestände zur Verfügung, deren Erschließung sehr ähnliche Techniken erfordert wie „klassisches" *Big Data* in den Naturwissenschaften. Andererseits sind mittlerweile Größenordnungen und Komplexitätsgrade erreicht, die es unumgänglich machen, andere Organisationsprinzipien für solche großen Datenbestände zu finden – das Stichwort hierbei lautet *Smart Data*. Im Kern dieser Entwicklung hin zu Daten, die Ansätze zu ihrer computergestützten Modellierung und Auswertung bereits in sich tragen, stehen Formen der *semantischen Erschließung*. Ganz allgemein kann in diesem Zusammenhang davon gesprochen werden, dass nun in einer Vielzahl von Anwendungsgebieten (auch solchen, in denen linguistische Werkzeuge nicht traditionell zum Einsatz kommen) *computerlinguistische Verfahren* zur Norm werden.

Bei den heute als *Big Data* bezeichneten Datenbeständen handelt es sich gegenüber früher erhobenen Massendaten nicht um ein grundsätzlich neues Phänomen. Aufgrund der Größe und Breite aktueller semi- oder unstrukturierter Bestände ergibt sich jedoch ein handfester qualitativer Unterschied zu herkömmlichen Zähl- oder Messdaten, die entlang spezifischer Merkmalsdimensionen erhoben werden: Unstrukturierte Daten müssen erst anhand einer Vielzahl unterschiedlicher Dimensionen analysiert werden, um ihre bedeutungstragende Merkmale herauszufiltern. Die Eigenschaften solcher großen und unstrukturierten Datenbestände werden im Allgemeinen mithilfe der sogenannten *V's* beschrieben. Nach Laney (2001) handelt es sich bei *Volume* (Datenmenge), *Velocity* (Input/

Output Geschwindigkeit) und *Variety* (unterschiedliche Datentypen und -quellen) um die grundlegenden Merkmale von *Big Data* sowie der gesamten damit in Zusammenhang stehenden Entwicklung, wobei von einer fortwährenden Steigerung dieser Faktoren ausgegangen wird (ebd.). Inzwischen gibt es eine Reihe von Ergänzungen und alternativen Aufzählungen der *V's*, die jeweils andere Aspekte hervorheben und dabei eher konkrete Anwendungen in den Vordergrund stellen.[1] Ergänzend finden sich häufig *Veracity* (Richtigkeit) und *Value* (das generierte Wissen) sowie – insbesondere für die in Cluster 5 durchgeführten Arbeiten zutreffend – *Variability* (die Polysemie und Kontextabhängigkeit von Sprache) und *Visualization* (als Werkzeug zur Analyse und Kommunikation).

Um die dadurch entstehenden Herausforderungen für das Wissensmanagement zu bewältigen, werden in einer Reihe von Disziplinen wie z. B. *Information Retrieval, Natural Language Processing* oder Bioinformatik laufend neue Verfahren entwickelt. Diese Verfahren sind methodisch in einer Schnittmenge aus Statistik und Informatik, dem Bereich des *Machine Learning*, angesiedelt und können prinzipiell an die sprachlichen Gegebenheiten geisteswissenschaftlicher Forschungsdatenbestände, wie z. B. Sammlungen historischer, literarischer oder sakraler Texte, angepasst werden. Im Zuge dieser Entwicklung entsteht auch ein neues Begriffsinventar in jenen Textwissenschaften, die ihre Quellen im Rahmen von *Distant Readings*, z. B. als *Graphs, Maps* und *Trees*,[2] analysieren. Dies ist jedoch nicht als einseitiger Technologietransfer zu sehen, sondern birgt auch die Chance, *Big Data*-getriebene Modellbildung zu *re-humanisieren*.[3] Die Verbindung von großen Datenbeständen und geeigneten quantitativen Verfahren ermöglicht es bereits, theoretische Annahmen und möglicherweise auf anekdotischen Informationen beruhende Kategorienbildungen und Interpretationen einer empirischen Überprüfung zu unterziehen. Durch eine solche Modellierung wird schließlich auch eine unmittelbare Anwendbarkeit und Erfahrbarkeit geisteswissenschaftlicher Begrifflichkeiten und Konzepte möglich. Es handelt sich dabei um einen experimentellen Zugang zur Theoriebildung, der die Entwicklung von Analysewerkzeugen und die konzeptuelle Reflexion als eng miteinander verknüpfte Arbeitsbereiche auffasst.

Das interdisziplinäre Feld der digital arbeitenden Geisteswissenschaften bringt Ansätze hervor, die in der Lage sind, neuartige Fragen zu stellen und damit sowohl ihre Untersuchungsgegenstände als auch die Modi der akademischen Wissensproduktion zu erweitern und zu verändern. Im engeren Kontext quantitativer Ansätze und im Kontext des Cluster 5 im Speziellen besteht die Möglichkeit, philologisch-kritische und historisch-vergleichende Perspektiven in die Modellierung großer Datenbestände einfließen zu lassen und in weiterer Folge neue, auf geisteswissenschaftliche Fragestellungen zugeschnittene Analysewerkzeuge zu entwickeln.

1 Ausrichtung und Ziele des Clusters

Vor diesem Hintergrund hat DARIAH-DE Cluster 5 „‚Big Data' in den Geisteswissenschaften" eingerichtet. Ziel des Clusters ist die Erschließung, Weiterentwicklung und Vermittlung relevanter methodologischer Kompetenzen, die es ermöglichen, digitale Verfahren und Werkzeuge zur Analyse und Darstellung großer geisteswissenschaftlicher Forschungsdatenbestände einzusetzen. Der Fokus liegt hierbei vor allem auf der Senkung von Einstiegsschwellen: Interessierte Geisteswissenschaftler sollen auf möglichst einfache Art neue Forschungsmethoden nutzen können, die auf der Analyse großer Datenmengen basieren, so dass ein möglichst breiter Nutzerkreis von den Möglichkeiten der aktuellen technischen Entwicklung profitieren kann. In der Praxis bedeutet das, dass existierende Methoden und Tools weiterentwickelt und vor allem an die spezifischen Bedürfnisse digital arbeitender Geisteswissenschaftler angepasst werden. Begleitend zur technischen Umsetzung erstellt Cluster 5 Lehrmaterialien zu den entwickelten Tools und organisiert Workshops, um entsprechende Kenntnisse in die Forschungscommunity zu vermitteln und Feedback aus der Community in die Entwicklungsarbeit einfließen lassen zu können.

Im Mittelpunkt der Entwicklungsarbeit steht hierbei nicht unbedingt das *User Experience Design*, sondern die Erleichterung des Einstiegs in die Nutzung von Skriptsprachen. Im Gegensatz zur Entwicklung ausgereifter Softwarelösungen mit graphischen Benutzeroberflächen für spezifische Aufgaben erlaubt dieser Ansatz eine schnelle und unkomplizierte Anpassung der entwickelten Lösungen an neue Anforderungen und spezifische methodische Bedürfnisse einzelner Forschender. Der Fokus der Arbeit in Cluster 5 liegt damit vor allem auf dem *Empowerment*: Gefördert wird die Fähigkeit interessierter Forschender, neue Methoden nicht nur einzusetzen, sondern auch selbstständig weiterzuentwickeln und anzupassen.

1 Vgl. Hopkins und Evelson (2011) und Khan et al. (2014).
2 Moretti (2005).
3 Prescott (2015).

2 Use Cases

Die Arbeit im Cluster ist um Use Cases herum organisiert: Im einen geht es um die Extraktion komplexer, auf linguistischen Informationen basierender Features für die Analyse narrativer Techniken aus literarischen Texten, im anderen um die Erstellung und die historische Analyse von grenzüberschreitenden Biographien anhand von strukturierten und unstrukturierten Daten. Diese Use Cases stehen beispielhaft für Fragestellungen aus verschiedenen geisteswissenschaftlichen Forschungsfeldern, in denen große Mengen textbasierter, digitaler Forschungsdaten zur Verfügung stehen und am Computer analysiert werden können. Auf der Basis dieser Use Cases erfolgt die Entwicklungsarbeit in Cluster 5.

3 Narrative Techniken

Im Use Case „Narrative Techniken" soll exemplarisch demonstriert werden, wie eine große Sammlung literarischer Texte genutzt werden kann, um mithilfe quantitativer Verfahren die historische Entwicklung narrativer Techniken – und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien – zu analysieren. Dabei ist es das Ziel, ein Set von Best Practices, Beispiel-Workflows und allgemeinverständlichen Tutorials zu erstellen, die es Textwissenschaftlern ermöglichen, innovative, bereits vorhandene Werkzeuge flexibel auf ihre eigenen Daten anzuwenden.

Datengrundlage für die Entwicklung des Use Case ist ein Korpus bestehend aus rund 2000 deutschsprachigen Romanen, das sich vor allem aus Werken des 18. und 19. Jahrhunderts zusammensetzt und zum Teil in TEI-Format kodiert ist. Um die Übertragbarkeit und Robustheit der Lösungen weiter zu prüfen, werden die Werkzeuge auch auf eine Sammlung von 200 französischen Kriminalromanen des 19. und 20. Jahrhunderts angewandt.

Die wichtigste Grundlage für viele computerbasierte Analyseverfahren ist der Zugang zu linguistischen Informationen zu den untersuchten Texten. Nur wenn die grammatische Funktion der einzelnen Wörter in einem Satz bekannt ist oder die Eigennamen von Figuren in einem Roman als solche identifiziert werden können, lassen sich Eigenschaften, Handlungen und Interaktionen bestimmten Figuren zuordnen und das ist wiederum die Voraussetzung für eine inhaltliche Analyse. Solche Prozesse, die beim direkten Lesen einzelner Texte eher trivial erscheinen, für die Massenanalyse ganzer Textkorpora am Computer zu automatisieren, stellt die Forschung noch immer vor enorme Herausforderungen. Andererseits haben Computerlinguisten gerade auf der Ebene der Sprache in den letzten Jahren enorme Fortschritte gemacht. So stehen heute zahlreiche kleinere Softwaretools für verschiedene Aspekte der Sprachanalyse zur Verfügung. Für dieses Arbeitsfeld hat sich international die Bezeichnung *Natural Language Processing* (NLP) etabliert. Die bleibende Herausforderung besteht in der Kombination und Integration verschiedener solcher NLP-Komponenten zu einer auf ein komplexeres Problem zugeschnittenen Forschungsmethode.

Einen wesentlichen Beitrag zur Integration solcher Komponenten leistet das Darmstadt Knowledge Processing Software Repository (DKPro) der Technischen Universität Darmstadt. Es handelt sich um ein Apache UIMA basiertes Rahmenwerk, das einen einheitlichen Zugang zu einer Vielzahl von NLP-Werkzeugen ermöglicht und dadurch die Entwicklung eigener Systeme vereinfacht. Trotz des wesentlich vereinheitlichten und damit vereinfachten Zuganges, den DKPro einem Programmierer zu NLP-Werkzeugen bietet: Die Schwelle zur Nutzung ist hier immer noch die Fähigkeit, in C oder Java zu programmieren.

Um diese Schwelle mit Blick auf die digital arbeitenden Geisteswissenschaftler noch einmal deutlich herabzusetzen, hat die Technische Universität Darmstadt in ihrer Funktion als DARIAH-Partner den DARIAH-DKPro-Wrapper (DDW)[4] entwickelt. Dieses Java-Programm extrahiert aus einem gegebenen Text eine ganze Reihe von linguistischen Informationen und gibt sie in einem umfangreichen, komfortabel lesbaren Tabellenformat aus. Der DDW bündelt große Teile der Funktionalität des DKPro-Frameworks in einem einzigen Programm, das über eine Konfigurationsdatei gesteuert werden kann. Das Programm kann mit nur einem Kommandozeilenbefehl ausgeführt werden, was die Einstiegsschwelle zur Nutzung weitreichender NLP-Funktionen wesentlich herabsetzt, gleichzeitig aber ein hohes Maß an Konfigurierbarkeit und Anpassungsmöglichkeiten erhält.

Basierend auf dem DDW entwickelt die Universität Würzburg eine Reihe von Beispiel-Workflows in Form von allgemeinverständlichen Programmier-„Rezepten". Diese demonstrieren Textwissenschaftlern, wie sie mithilfe des DDW auf die Komponenten des leistungsstarken DKPro Rahmenwerks zugreifen und den Output mithilfe weit verbreiteter Skriptsprachen wie Python und R weiterverarbeiten, um z. B. den Schreibstil eines Autors näher zu charakterisieren, inhaltliche Aspekte anhand wiederkehrender Themen zu erkunden oder die Figurenkonstellation einer Erzählung als Netzwerk darzustellen. Diese Rezepte wer-

4 https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper.

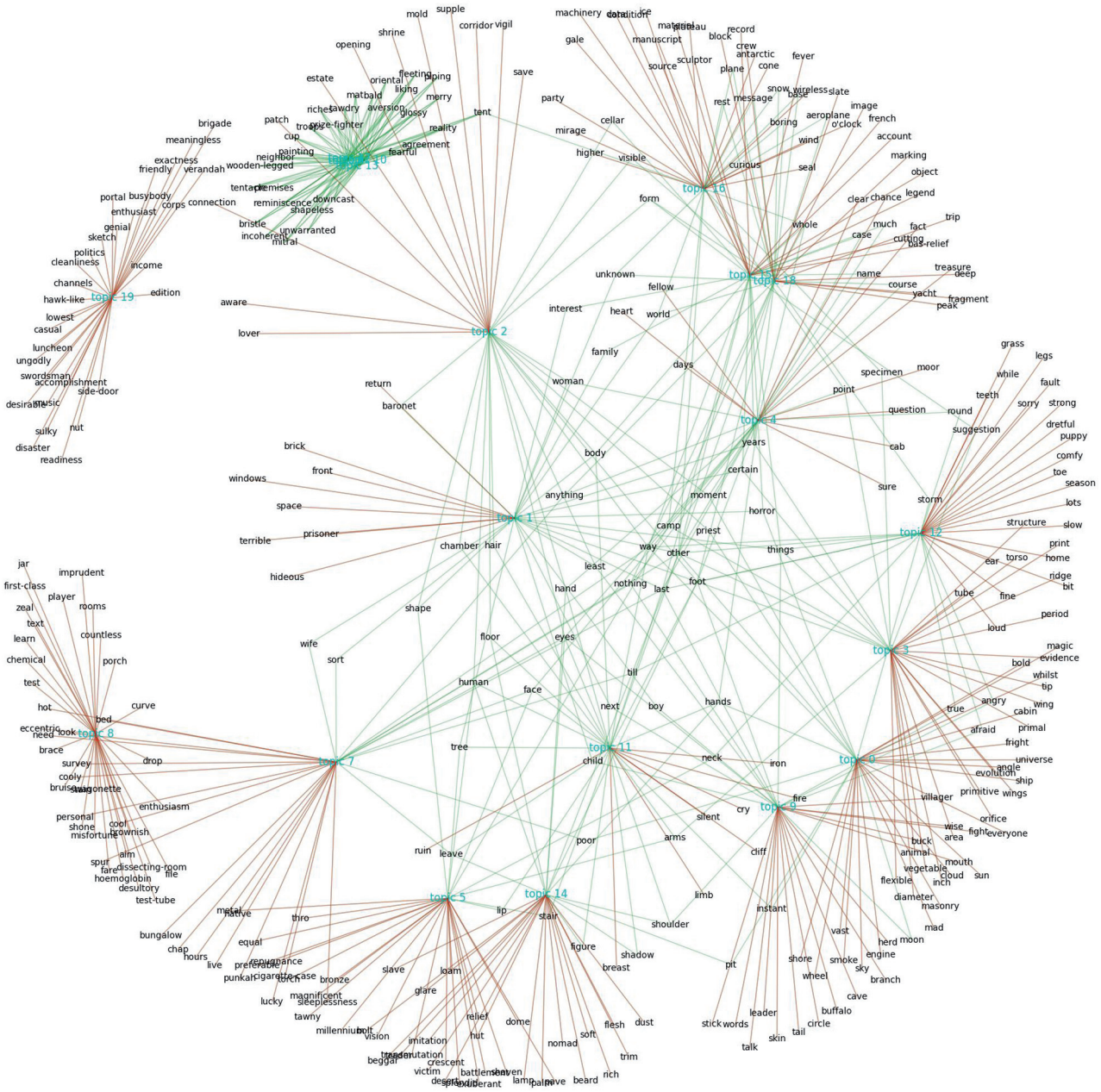

Abb. 1: Themennetzwerk von einem Kurzgeschichtenkorpus. Nach einem Beispielrezept erstellt mit dem DDW und LDA Topic Modeling

den in Form von Onlinetutorials der breiten Öffentlichkeit zur Verfügung gestellt und stetig ergänzt und weiter entwickelt. Hiermit leistet der Use Case einen nachhaltigen Beitrag zur Verbreitung und Vermittlung neuer, computerbasierter Forschungsmethoden in den textbasierten Geisteswissenschaften.

4 Biographien – Korrelationen zwischen Personen, Orten, Daten und Ereignissen

Der in einer interdisziplinären Kooperation zwischen dem Leibniz-Institut für Europäische Geschichte in Mainz (IEG) und dem Lehrstuhl für Medieninformatik an der Universität Bamberg bearbeitete Use Case „Biographien“ behandelt die Identifizierung von Korrelationen zwischen Personen, Orten, Daten und Ereignissen. Der Use Case ist inhaltlich eng verbunden mit dem am IEG angesiedelten

geschichtswissenschaftlichen Forschungsprojekt „Cosmobilities – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiographien des 19. Jahrhunderts“.[5]

Den konzeptionellen Ausgangspunkt von Cosmobilities bildete die Frage, inwiefern die Untersuchung grenzüberschreitender Lebensläufe in den europäischen Nationalbiographien unter Einbeziehung digitaler Ressourcen und anderen biographischen Materials gängige, nationalstaatlich geprägte Narrative über das 19. Jahrhundert hinterfragen und zugleich der historischen Forschung für diese Zeitperiode neue Impulse verleihen kann. Dabei sollte jedoch nicht lediglich das nationale Narrativ durch ein grenzüberschreitendes ersetzt werden. Das Projekt setzt vielmehr die Annahme voraus, dass das Handeln der hier untersuchten Akteure im Spannungsfeld sowohl von nationalen Zuschreibungen als auch erhöhter Mobilität bzw. Internationalität zu analysieren ist. Erst durch diese multiperspektivische Herangehensweise können mehrere, relational gedachte Handlungskontexte wie etwa lokale, nationale, imperiale oder globale Zusammenhänge in den Fokus rücken und so neue Forschungserkenntnisse fördern.

Aus technischer Perspektive besteht der Kern des Use Case in der Kombination von Informationen aus strukturierten und unstrukturierten Quellen. Als Datengrundlage werden hierzu im ersten Schritt strukturierte Daten aus Wikidata und die unstrukturierten Texte der deutsch- und englischsprachigen Wikipedia herangezogen, um eine – insbesondere für die Anwendung quantitativer Verfahren – ausreichende Anzahl biographischer Einträge[6] gewinnen zu können. Im weiteren Verlauf wird die Datenbasis um zusätzlich strukturierte und unstrukturierte Quellen erweitert, wie etwa die Allgemeine Deutsche Biographie (ADB), die Neue Deutsche Biographie (NDB) zusätzliche oder auch weitere europäische Nationalbiographien wie beispielsweise das Oxford Dictionary of National Biography oder das Polish Biographical Dictionary (= Polski Słownik Biograficzny). Aufgrund ihres Umfangs eignen sich die so gewonnenen Daten nicht unmittelbar für die Auswertung durch klassische geschichtswissenschaftliche Methoden. Stattdessen werden sie mithilfe eines eigens für das Forschungsprojekt entwickelten Such- und Analysewerkzeugs (*Cosmotool*[7]) erschlossen und verarbeitet. Mit diesem können Personen anhand festgelegter Kategorien (Geburts- und Sterbedaten, Verortung, Verknüpfung an bestimmte Ereignisse, zugeschriebene Berufe etc.) gesucht werden. So können beispielsweise Personen mit bestimmten Lebensdaten und Berufen identifiziert werden. Weitere Ereignisse wie etwa Geburt, Tod, Studium, Heirat, Kinder etc. werden in Relation zu Daten und Orten ausgewertet, wodurch erste Hinweise auf Mobilität erkennbar werden. Mithilfe einer Zeitleiste können die extrahierten Daten graphisch visualisiert werden. Eine Kartendarstellung macht die Mobilität der Personen sofort sichtbar. Mittelfristig wird das Cosmotool als Schnittstelle zwischen quantitativen und qualitativen Methoden der Betrachtung biographischer Daten dienen. Hierbei soll zum einen eine quantitative Vorauswahl von Profilen nach spezifizierbaren Kriterien für eine anschließende qualitative Betrachtung ermöglicht werden. Zum anderen könnten qualitative Einschätzungen – im Fall des Cosmobilities-Projektes insbesondere im Hinblick auf Mobilität und Grenzüberschreitungen – quantitativ evaluiert werden.

Der Use Case „Biographien“ ist aufgrund seiner interdisziplinären Struktur ein Knotenpunkt, an dem sich zwei verschiedene wissenschaftliche Richtungen und Fragestellungen kreuzen. Das *Cosmotool* ist dabei ein technisches Werkzeug, das von beiden Disziplinen zur Erkenntnisförderung genutzt wird. Neben der vorgestellten historischen Fragestellung, die den inhaltlichen Ausgangspunkt für die Arbeiten der Historiker am IEG an dem Use Case stellt, verfolgt die Bamberger Medieninformatik eine eigene Fragestellung, die sich mit der integrierten Analyse strukturierter und unstrukturierter Daten beschäftigt. Damit verspricht der Use Case am Ende zu neuen Forschungserkenntnissen sowohl in den Geschichtswissenschaften als auch in der Informatik zu führen und zusätzlich ein neues Such- und Analysewerkzeug (*Cosmotool*) bereitzustellen. Technische Lösungen werden dabei soweit möglich im Rahmen der generisch wiederverwendbaren Föderationsarchitektur[8] von DARIAH-DE implementiert, wodurch weitere, artverwandte Forschungsfragen von den Arbeiten um das Cosmotool profitieren könnten.

Der Use Case ist ein gutes Beispiel für gegenseitige Lernprozesse und Synergieeffekte innerhalb einer interdisziplinären Kooperation, die für beide Wissenschaftszweige als eine Bereicherung um neue Forschungsfragen betrachtet werden kann. Dabei sind weder die Informatik ein bloßer technischer Hilfsdienst für die Geschichtswissenschaft noch die Geschichtswissenschaft ein simpler Anwendungslieferant für die Informatik. Vielmehr ist das

5 http://www.ieg-mainz.de/Forschungsprojekte————_site.site.ls_dir._nav.17_f.69_likecms.html.

6 Der Prototyp verzeichnet derzeit 1,8 Mio. aus strukturierten Daten abgeleitete biographische Profile sowie knapp 0,5 Mio. aus biographischen Volltexten.

7 http://search.de.dariah.eu/cosmotool/search.

8 Gradl, Henrich und Plutte (2015).

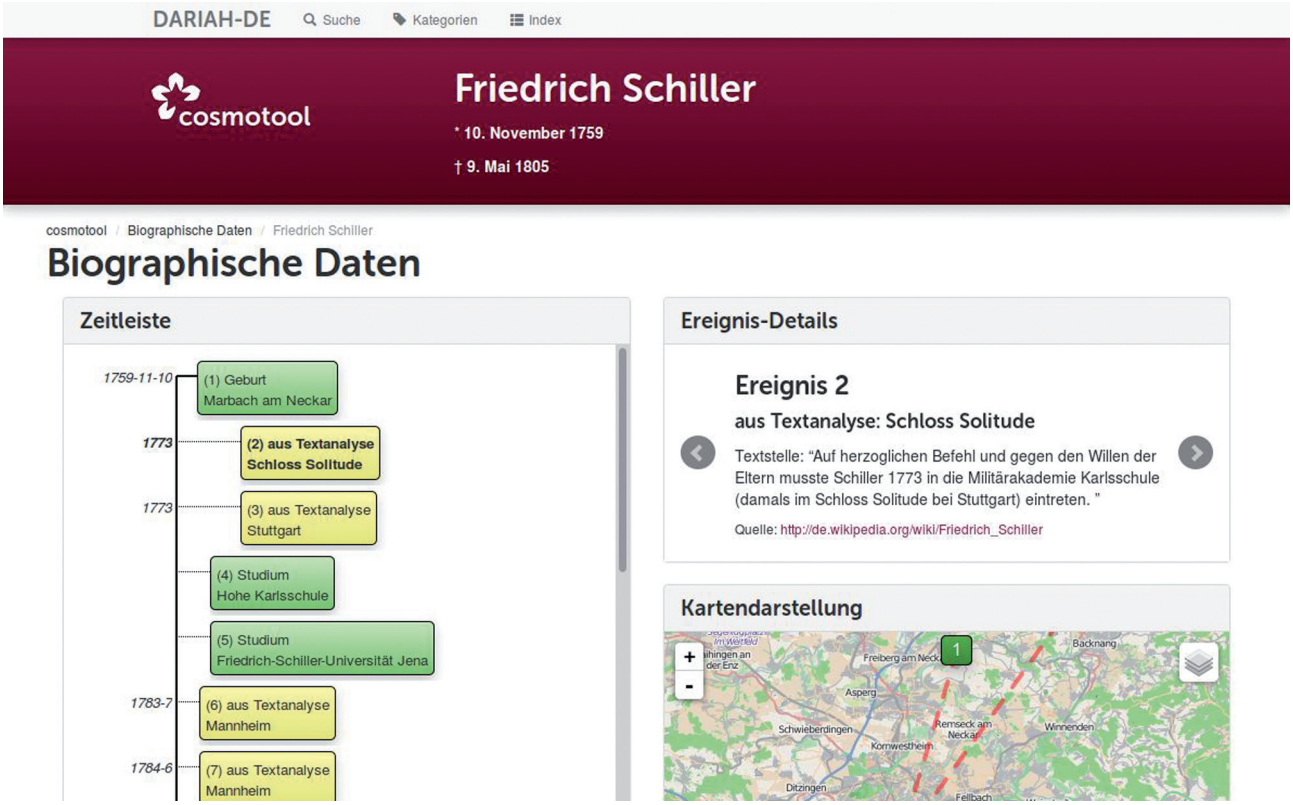

Abb. 2: Cosmotool: Biographische Daten zu Friedrich Schiller

Vorgehen durch einen Dialog der beiden Fachrichtungen gekennzeichnet, der als ein iterativer Prozess zu neuen Forschungserkenntnissen und bislang nicht identifizierten Fragestellungen in beiden Disziplinen stark beitragen kann. Eine Kooperation dieser Art benötigt jedoch das Wissen über die disziplinären Methoden und über unausgesprochene Annahmen der jeweils anderen Fachrichtung, um Missverständnisse zu vermeiden und um gelungenen Austausch und Erkenntnisgewinn zu ermöglichen.

5 Fazit

Beide Use Cases demonstrieren anhand beispielhafter Forschungsszenarien, wie digitale Datenbestände in für die Geisteswissenschaften innovativer Weise analytisch genutzt werden können. Mit ihrer Entwicklung verfolgt Cluster 5 zwei Ziele. Zum einen dienen sie beispielhaft der Inspiration und Anregung für künftige Forschungsprojekte, da sie demonstrieren, was in diesem Bereich bereits technisch möglich ist. Zugleich liefern sie aber auch gleich eine Reihe von Software- und Codekomponenten, die bei der Realisierung vergleichbarer Projekte direkt genutzt und in eigene Arbeitsabläufe eingebaut oder an die eigenen Bedürfnisse angepasst werden können. In der Entwicklungsarbeit wird darum auf die Quelloffenheit der Komponenten, ihre Verfügbarkeit und Modifizierbarkeit besonderer Wert gelegt. Gleichzeitig betonen die dazu entwickelten Schulungsmaterialien gerade diese Modifizierbarkeit sowie die Modularität der in den Use Cases verwendeten Techniken mit dem Ziel, nicht nur das technisch Mögliche zu demonstrieren, sondern gleichzeitig die Einstiegsschwelle für Forschende herabzusetzen und einfache Wege aufzuzeigen, wie derartige Methoden an ein breites Spektrum von Forschungsthemen angepasst werden können.

Literatur

Gradl, Tobias; Henrich, Andreas; Plutte, Christoph (2015): Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen. In: *Grenzen und Möglichkeiten der Digital Humanities*. Hg. v. Constanze Baum und Thomas Stäcker. Sonderband der Zeitschrift für digitale Geisteswissenschaften. Verfügbar unter http://www.zfdg.de/sb001_020.

Hopkins, Brian; Evelson, Boris (2011): Expand your digital horizon with Big Data. In: *Forrester*, 30.

Khan, M. Ali-ud-din; Uddin, Muhammad Fahim; Gupta, Navarun (2014): Seven V's of Big Data. ASEE Zone 1.

Laney, Doug (2001): 3D-Data Management: Controlling Data: Volume, Velocity and Variety. In: *META Group Research Note*, 6, 70.
Moretti, Franco (2005): Graphs, Maps, Trees: Abstract Models for a Literary History. In: *Verso*.
Prescott, Andrew (2015): Big Data in the Arts and Humanities: Some Arts and Humanities Research Council Projects. University of Glasgow Emblem Studies.

Anna Aurast
Leibniz-Institut für Europäische Geschichte (IEG)
Alte Universitätsstraße 19
D-55116 Mainz
aurast@ieg-mainz.de

Tobias Gradl
Otto-Friedrich-Universität Bamberg
Lehrstuhl für Medieninformatik in der Fakultät für Wirtschaftsinformatik und Angewandte Informatik
An der Weberei 5
D-96047 Bamberg
tobias.gradl@uni-bamberg.de

Stefan Pernes
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte
Institut für Deutsche Philologie
Universität Würzburg
Am Hubland
D-97074 Würzburg
stefan.pernes@uni-wuerzburg.de

Steffen Pielström
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte
Institut für Deutsche Philologie
Universität Würzburg
Am Hubland
D-97074 Würzburg
pielstroem@biozentrum.uni-wuerzburg.de