Bilanz Lehrmaterialien

Big Data in den Geisteswissenschaften (R 5.4.2)

Version 07.04.2016
Cluster 5
Verantwortliche Partner IEG Mainz

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: <Final>
Verfügbarkeit: <DARIAH-DE-public>
Autoren:

<Claudia Falk, IEG Mainz>

Revisionsverlauf:

Datum	Autor	Kommentare
21.03.2016	Claudia Falk	Erster Entwurf
07.04.2016	Claudia Falk	Finalisierung des Reports

Inhaltsverzeichnis:

1. Einleitung....4
2. Bilanz Lehrmaterialien....5
2.1 Use Case Narrative Techniken: Tutorial zur *NLP Based Analysis of Literary Texts* 5
2.2 Use Case Biographien: Tutorial für das Erstellen von kontrollierten Vokabularen mithilfe des Labeling System....7
3. Fazit....9
4. Webseitenverzeichnis....10

1. Einleitung

Dieser Report stellt die Ergebnisse vor, die hinsichtlich der Inhalte, Darstellungsformen und der medialen Umsetzungsmöglichkeiten der Lehrmittel zum Thema “Big Data Methodik in den Geistes- und Kulturwissenschaften” erzielt wurden. Im Gegensatz zu dem prototypisch-generischen Konzept einer Lehr- und Lernmittelsammlung und dem damit verbundenen Disseminationskonzept, das in R 5.4.1 *Lehrmaterialiensammlung* präsentiert wurde, bietet der vorliegende Report einen konzentrierten Überblick über die erreichten Ziele und die erarbeiteten Lehrmittel. Bis zum Ende der Laufzeit von DARIAH II im Februar 2016 wurden im AP 5.4 folgende Arbeitsschritte umgesetzt:

- die Use Cases wurden als prototypische Anwendungsfälle für Big Data Methoden in den Geisteswissenschaften bearbeitet,
- darauf basierend wurden Lehrmaterialien für die in den Use Cases genutzten Tools und Methoden erstellt,
- und es wurden Workshops durchgeführt, die dazu dienten, die Use Cases und die darin verwendeten Technologien und Methoden (Expertenworkshops) sowie die angewendeten Verfahren (Methodenworkshops) zu vermitteln.

Im Folgenden sollen die in Cluster 5 entwickelten Lehrmaterialien vorgestellt werden. Bei den Lehrmaterialien handelt es sich um zwei Tutorials: das Tutorial, das in Use Case 1 (*Narrative Techniken*) entwickelt wurde, stellt die Anwendung des *Natural Language Processing* (*NLP*) auf literarische Texte vor (*NLP Based Analysis of Literary Texts*),[1] für den Use Case 2 (*Biographien*) wurde ein Leitfaden zum Erstellen kontrollierter Vokabulare mithilfe des *Labeling System* entwickelt.[2] Auf eine Zusammenfassung der veranstalteten Workshops wird verzichtet, da diese Informationen bereits in M 5.4.2 *Workshops 2014* (tatsächlich Dokumentation der Planung der für 2015 geplanten Workshops, da 2014 keine Veranstaltungen durchgeführt wurden) und M 5.4.4 *Workshops 2015* (Durchführung und Dokumentation der Experten- und Methoden-Workshops, die 2015 stattfanden) ausführlich aufbereitet wurden.

1 https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper/blob/master/doc/tutorial.adoc
2 http://labeling.i3mainz.hs-mainz.de/

2. Bilanz Lehrmaterialien

2.1 Use Case Narrative Techniken: Tutorial zur *NLP Based Analysis of Literary Texts*

Im Rahmen des Use Case 1 *Narrative Techniken und Untergattungen im deutschen Roman* wurde eine große Sammlung literarischer Texte mithilfe quantitativer Verfahren untersucht. Ziel dieser Analyse war es, darzulegen, wie sich narrative Techniken historisch gewandelt und darauf basierende literarische Kategorien entwickelt haben. Dazu wurden Verfahren zur automatischen Erkennung bestimmter Merkmale, z.B. Eigennamen oder Passagen direkter Rede, eingesetzt. Mithilfe dieser Merkmale, die im Anschluss als *Features* zueinander in Bezug gesetzt wurden, ließen sich Texte gruppieren und narratologische Kategorien überprüfen.
Erklärtes Ziel des Use Case war es, zur Entwicklung der Werkzeuge Lernmaterialien zu erstellen, die sowohl Anwender mit weniger Erfahrung im Umgang mit Programmiersprachen als auch Experten nutzen können. Die technische Umsetzung umfasste zwei Bereiche. Sämtliche *NLP* Vorverarbeitungsschritte wurden mit Hilfe des Darmstädter *DKPro Frameworks* als eine generische *Pipeline*, d.h. als Abfolge aufeinander aufbauender Verarbeitungsschritte, umgesetzt und als Plain-Text-Format ausgegeben. Dieses Format kann programmatisch einfach eingelesen und auch mit herkömmlicher Office Software geöffnet werden. Ab diesem Punkt begann die Weiterverarbeitung der Ergebnisse, im Use Case prototypisch umgesetzt anhand der Skriptsprachen *Python* und *R*. Da die Werkzeuge des Use Case und die Lernmittel auf Programmcode ohne graphische Benutzerführung basieren, sind für die NutzerInnen Kenntnisse über die Grundlagen der Programmierung, der Statistik und des *Natural Language Processing (NLP)* hilfreich. Sie sind allerdings nicht zwingend notwendig, da für alle relevanten Konzepte und Methoden weiterführende Informationen bereitgestellt werden.
Um die NutzerInnen in die Textanalyse mithilfe linguistischer Annotationen einzuführen, wurde in Cluster 5 ein entsprechendes Tutorial[3] entwickelt. Dieses Tutorial beschreibt zunächst wie linguistische Annotationen mithilfe eines einfachen Java-Programms, das die Fähigkeiten verschiedener aktueller *NLP* Werkzeuge miteinander verbindet, erzeugt werden können. Die erstellten Annotationen sind dann in einem standardisierten Ausgabeformat abrufbar und lassen sich für komplexe empirische Textanalysen zu Stil und Inhalt in einer Skriptsprache nutzen. Die Workflows basieren auf Werkzeugen aus dem *DKPro Framework* für *NLP*, die in einer einzigen Datei vorliegen. Damit ermöglichen es die Workflows auch NutzerInnen, die keine Arbeitserfahrung mit dem Framework haben, praktisch mit den Tools zu arbeiten.
Im ersten Teil des Tutorials wird dargelegt, wie die *Pipeline* aufgesetzt wird. Dazu wird ausgeführt, welche Systemanforderungen gewährleistet sein müssen, wie Java

3 https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper/blob/master/doc/tutorial.adoc

installiert wird und wie der Download der *Pipeline* funktioniert. Es folgen Informationen zum Start der *Pipeline*. Die NutzerInnen erhalten Einblick in den Gebrauch der Kommandozeile, in die Verarbeitung einer Textdatei und in den *File Reader*, wobei zusätzlich über den *Text Reader* und den *XML Reader* sowie die *Reading Directories* informiert wird. Zum Start der *Pipeline* zählen auch die Auswahl der Sprache, die Optionen, die mit der Kommandozeile verbunden sind (z.B. kann über die Kommandozeile eine Hilfefunktion abgerufen werden) sowie Informationen zur Fehlerbehebung.
In einem nächsten Schritt erklärt das Tutorial, welche Komponenten es gibt. Ausgehend vom einfachsten Bestandteil der Segmentierung folgen Abschnitte zum Taggen von Sprachbestandteilen (z.B. Artikel, Verben, Nomen usw.), zur Lemmatisierung und zum *Parsing*. Diese letztere Komponente dient etwa dazu, den Text gemäß vordefinierter Regeln in bestimmte Strukturen einzuteilen. Hinzu kommen die Funktionen der Namenserkennung (*Named Entity Recognition*) und der semantischen Rollenzuweisung, mit deren Hilfe es möglich ist, alle im Text vorhandenen Namen herauszusuchen bzw. einem Verb argumentative Aspekte beizufügen.
Das Tutorial informiert zudem darüber, wie die *Pipeline* konfiguriert werden kann. So wird erklärt, wie man alle Funktionen der Pipeline aktivieren, eine eigene Konfigurationsdatei schreiben, *Argument Parameter* nutzen und letztlich eine eigene *Pipeline* programmieren kann. Weiterhin wurden im Tutorial Informationen zu *TreeTagger* zusammengestellt, dem Tool, das dazu dient, einen Text mit Informationen zu den Sprachbestandteilen und den Lemmata zu versehen. Da *TreeTagger* aus Gründen des Copyright nicht direkt vom *DKPro Repository* aus genutzt werden darf, wird erklärt, wie man *TreeTagger* installieren und für die Nutzung konfigurieren kann.
Schließlich stellt sich die Frage nach dem Ausgabeformat, die im Tutorial behandelt wird. Dazu werden Hinweise gegeben, wie man die Ausgabedatei in den Programmiersprachen R und Python einbindet. Es folgen weitere Beispiele, etwa zu den Möglichkeiten bestimmte Textabschnitte, Sätze oder Satzbestandteile herausgefiltert werden können. Auch einzelne Sätze lassen sich nach spezifischen Bestandteilen (z.B. Nomen) filtern.
Das Tutorial stellt außerdem verschiedene Beispielfälle vor, die zeigen, wie Ergebnisse der *Pipeline* für fortgeschrittene Analyseprozesse genutzt werden können. So wird z.B. erklärt, inwiefern sich mithilfe des Output der *NLP Pipeline* und des *Stylo* Tools stilometrische Merkmale in einem bestimmten Textkorpus analysieren lassen. *Stylo* ist in R verfasst und verfügt über eine graphische Nutzeroberfläche, die mehrere Funktionen für die stilometrische Analyse bietet. Das Tutorial informiert in diesem Zusammenhang außerdem zu einem Beispielkorpus und erklärt die wesentlichen Schritte (*Preparing Descriptive Vocabulary and Part-of-Speech Tags*; *Using Stylo*), die die NutzerInnen während des Analyseprozesses nachvollziehen müssen. Weiterhin gibt es *Example Recipes* zum Topic Modeling in Python sowie zur stilometrischen Klassifizierung in Python. Auch zu diesen Beispielen werden alle relevanten Informationen (etwa zu Beispielkorpora, Datenvorbereitung und

Modellanpassung) vorgestellt, um den NutzerInnen die notwendigen Schritte für den eigenen Umgang zu erklären.
Das Tutorial bietet damit einen Leitfaden zur *NLP*-basierten Analyse literarischer Texte, der auch NutzerInnen mit wenigen Kenntnissen in Programmiersprachen in die Lage versetzen soll, eigenständig Analysen durchzuführen und die vorgestellten Beispielfälle auf eigene Projekte anzuwenden.

2.2 Use Case Biographien: Tutorial für das Erstellen von kontrollierten Vokabularen mithilfe des Labeling System

Use Case 2 *Biographien – Korrelationen zwischen Personen, Orten, Daten und Ereignissen*, der vom IEG und von dem Lehrstuhl für Medieninformatik an der Universität Bamberg bearbeitete wurde, untersuchte die Verbindungen von individuellen historischen Lebensläufen und Internationalitätskriterien. Die Grundlage dafür bildeten Daten aus Wikipedia und mehreren europäischen Nationalbiographien. Dabei wurden Frageraster und Kategorien für die Analyse serieller, digital verfügbarer Massendaten in der Biographieforschung entwickelt. Inhaltlich stand der Use Case Biographien in enger Verbindung mit dem Use Case des IEG in Cluster 4, in dessen Rahmen exemplarisch kontrollierte Vokabularen entwickelt und der Einsatz von Normdaten in der historisch arbeitenden Forschung erprobt wurde. Darüber hinaus wurde in diesem Zusammenhang auch die Nutzung des *Labeling System* demonstriert. Damit diente der clusterübergreifende Use Case Biographien und speziell das am IEG angesiedelte geschichtswissenschaftliche Forschungsprojekt *Cosmobilities – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiographien des 19. Jahrhunderts*, das den fachlichen Hintergrund für den geschichtswissenschaftlichen Teil des Use Case Biographien bildete, als Ausgangsbasis für die erstellten kontrollierten Vokabulare. Im *Cosmobilities*-Projekt wurden Biographien von bestimmten Personengruppen (z. B. Revolutionäre, Politiker, Unternehmer, Musiker) des 19. Jahrhunderts auf ihre räumliche und oftmals daraus resultierende nationale wie kulturelle Mobilität untersucht. Die Ausgangsthese des Projektes lautete, dass einzelne Persönlichkeiten und ihre Leistungen in Nationalbiographien oftmals im Namen einer Nation vereinnahmt werden, wodurch transnationale und transkulturelle Bezüge sowie kulturelle Austauschmomente unbeachtet blieben. Um die Mobilitätsprozesse, die die sozial verschiedenen Untersuchungsgruppen durchliefen, nachvollziehen zu können, wurden zwei Datensätze erhoben. Erstens sollten Orte und Ortstypen, an denen die Personen lebten, zweitens deren oftmals variierenden beruflichen Tätigkeiten, die sie im Laufe ihres Lebens ausübten und durch die sie einer bestimmten Gruppe zugehörig wurden, ausgewertet werden. Da die gruppenbiographische Herangehensweise des *Cosmobilities*-Projektes zwangsläufig Massendaten produzierte, war es notwendig, diese Daten in kontrollierten Vokabularen systematisch zu erfassen und aufzubereiten.

Diese kontrollierten Vokabulare für Tätigkeiten und Ortstypen wurden mithilfe der Webanwendung *Labeling System*[4] erstellt. Diese prototypische Anwendung ermöglicht es Usern, auf einfache Weise Begriffe zu importieren, *SKOS* (*Simple Knowledge Organization System Primer*) Vokabulare zu erstellen und die Termini der Vokabulare mit Begriffen eines oder mehrerer Referenzthesauri zu verlinken. Das *Labeling System* (*LS*) ist prinzipiell für alle Szenarien nutzbar, in denen projektspezifische kontrollierte Vokabulare benötigt werden. Für NutzerInnen, die selbst ein kontrolliertes Vokabular erstellen wollen, wurde ein entsprechendes Tutorial entwickelt, das alle wichtigen Schritte erklärt.

Das Tutorial[5] zum *LS* beschreibt zunächst, wie sich NutzerInnen registrieren und anmelden können. Nach der erfolgreichen Anmeldung im System ist es möglich, mit dem *LS* mehrere Projekte anzulegen, mehrere Vokabulare zu erstellen und bestimmten Projekten zuzuordnen sowie mehrere Labels zu erstellen und diese bestimmten Vokabularen zuzuordnen. Die User werden im Folgenden angeleitet, zunächst ein neues Projekt zu erstellen und erhalten Informationen zu den damit verbundenen Funktionen. Schließlich wird erklärt, wie man auf ähnliche Weise neue Vokabulare anlegen kann. Die Vokabulare weisen eine Reihe von Metadaten auf und können mit Projekten verbunden werden. Zudem lassen sich die Vokabulare im REST-Interface öffentlich machen und sind damit für andere NutzerInnen des *LS* sichtbar. Auch zu den Möglichkeiten Vokabulare als RDF- oder als CSV-Dateien zu exportieren (“Export as RDF” bzw. “Export as CSV”) wird informiert. Dazu ist es allerdings notwendig, dass dem Vokabular mindestens ein Label zugeordnet ist.

Das Erstellen und Zuweisen von Labels ist Inhalt der folgenden Rubrik. Anders als Projekte und Vokabulare lassen sich Labels auch nach der Erstellung noch verändern. Verknüpfungen von Labels und Vokabularen werden ebenfalls erklärt. Besonders interessant ist für NutzerInnen die Möglichkeit, mehrere Labels im *LS* mit Hilfe einer CSV-Datei einzufügen. Die dafür notwendige CSV-Datei muss zwingend einer bestimmten, vorgegebenen Struktur folgen, da die Datei ansonsten nicht im *Labeling System* hochgeladen werden kann. Das Tutorial gibt eine genaue Übersicht über alle erforderlichen wie optionalen Elemente und informiert die User, wo die Angaben zu finden sind. Um die Labels als CSV-Datei hochladen zu können, müssen die Inhalte der Excel-Datei schließlich in eine .txt-Datei eingefügt werden. Hierzu empfiehlt das Tutorial den Einsatz des Programms Notepad ++[6] und informiert über die weiteren Schritte. Wurde die entsprechende Datei schließlich im *LS* als korrekt akzeptiert, kann der Upload erfolgen.

Eine wichtige Funktion, die im Tutorial erklärt wird, betrifft das Verlinken von Labels mit Thesauri via SPARQL. Die Funktion ermöglicht das Verlinken von eigenen Labels mit bereits existierenden SKOS-Konzepten aus Thesauri anderer Anbieter, die im Internet vorhanden sind und die einen Zugang per SPARQL-Endpunkt bieten. Zu den im *LS* vorinstallierten Thesauri zählen etwa *Getty Thesaurus* oder *Data Culture*

4 http://labeling.i3mainz.hs-mainz.de

5 https://wiki.de.dariah.eu/download/attachments/26150828/LS%20Tutorial.pdf?version=1&modificationDate=1460132313878&api=v2

6 https://notepad-plus-plus.org/download

Thesaurus. Es wird beschrieben, wie die NutzerInnen Thesauri anderer Anbieter stichwortartig durchsuchen, auswählen und gefundene Konzepte schließlich mit einem eigenen Label verknüpfen können. Dazu zählen auch die Erklärungen der SKOS Zuordnungseigenschaften, die dazu dienen, die Beziehungen, in denen Konzepte und Labels zueinander stehen, genau zu definieren.
Zudem besteht die Möglichkeit, einen eigenen Referenzthesaurus im *LS* anzulegen, mit dessen Hilfe man die Labels des Vokabulars beschreiben kann. Dies ist dann notwendig, wenn die bereits existierenden, im System vorinstallierten Thesauri als Referenzen für ein erstelltes Vokabular nicht ausreichend sind. Das Tutorial zeigt alle notwendigen Schritte und erklärt den Vorgang der SPARQL Abfrage.
Im weiteren Verlauf wird zusätzlich über Parser informiert und die Funktion “Public Project Tree” vorgestellt. Letztere erlaubt es, die Struktur der vorhandenen Projekte, Vokabulare und Labels (Bezeichner) sowie der Konzepte, die jeweils mit ihnen verbunden sind, graphisch als Baumdiagramm (Tree) von allen Usern des *Labeling Systems* anzeigen zu lassen. Für jedes Projekt, Vokabular und Label wird das bevorzugte Label (prefLabel) in der Vorzugssprache (prefLang) angegeben. Das Tutorial bietet außerdem eine Übersicht zu den farbigen Markierungen, anhand derer man die Art der Beziehungen ablesen kann. Eine Übersicht zu den Farben und den entsprechenden Bedeutungen ist auch in den FAQs des *LS* enthalten.
Mit dieser Darstellung bietet das Tutorial allen interessierten NutzerInnen eine schrittweise Einführung in den Umgang mit dem *Labeling System*. Die User werden dadurch – unabhängig von technischen Vorkenntnissen und Erfahrungen mit dem Erstellen kontrollierter Vokabulare – in die Lage versetzt, für geschichtswissenschaftliche und andere Fragestellungen eigenständig kontrollierte Vokabulare zu konzipieren und technisch umzusetzen.

3. Fazit

Diese Bilanz hat aufgezeigt, welche Ziele bei der Bearbeitung der Use Cases als prototypische Anwendungsfälle für Big Data Methoden in den Geisteswissenschaften erreicht wurden. Die Lehrmittel, die in Use Case 1 und Use Case 2 erarbeitet wurden, stellen die jeweils genutzten Tools und Methoden umfassend vor und bieten interessierten Usern eine schrittweise Anleitung zum eigenständigen Arbeiten mit den Werkzeugen. Die Umsetzung der Tutorials zielte darauf, Wissen zu vermitteln und Lehr- und Lernprozesse zu unterstützen. Außerdem sollte es auf diese Weise gelingen, in einem eigenen Forschungsprozess die geplante Lehr- und Lernmittelsammlung zum Thema „Big Data Methoden in den Geisteswissenschaften“ zu unterstützen und damit die Forschung an DH-Methoden, Tools und Materialien sowie deren Förderung voranzutreiben. Neben der Entwicklung der vorgestellten Lehr- und Lernmittel in den Use Cases gilt es auch weiterhin, das in DARIAH I erstellte und gesammelte Material in Bezug auf das Thema *Data Science* sowie die in DARIAH II neu erarbeiteten Materialien langfristig in eine eigene Bibliographie zu “Big Data Methoden in den Humanities” zu integrieren. Beide Tutorials können

außerdem in der Bibliographie „Doing Digital Humanities“[7] sowie in den DARIAH Schulungsmaterialien[8] ergänzt werden. Zusätzlich könnte, wie in M 5.4.1 vorgeschlagen, die durch Cluster 5 bearbeitete erweiterte Bibliographie „Doing Digital Humanities“ mit Inhalten und Materialien aus DARIAH I und DARIAH II verbunden werden. Diese Übersicht belegt damit, welche Lehrmaterialien in Cluster 5 erarbeitet wurden und inwiefern diese für eine DARIAH Lehr- und Lernmaterialsammlung genutzt werden können.

4. Webseitenverzeichnis

Letzter Zugriff jeweils am 07.04.2016.

- Bibliographie „Doing Digital Humanities, https://de.dariah.eu/bibliographie; https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography/items/collectionKey/7MZHDTI5
- DARIAH Schulungsmaterialien, https://de.dariah.eu/schulungsmaterial
- Labeling System, http://labeling.i3mainz.hs-mainz.de/
- Labeling System: Tutorial, https://wiki.de.dariah.eu/download/attachments/26150828/LS%20Tutorial.pdf?version=1&modificationDate=1460132313878&api=v2
- Notepad ++, https://notepad-plus-plus.org/download
- Tutorial NLP, https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper/blob/master/doc/tutorial.adoc

7 https://de.dariah.eu/bibliographie; https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography/items/collectionKey/7MZHDTI5

8 https://de.dariah.eu/schulungsmaterial