Wunsch und Wirklichkeit – Forschungsinfrastrukturen in den Computational Literary Studies: interdisziplinär, modular, vernetzt?

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg

Helling, Patrick
patrick.helling@uni-koeln.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg

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

Kababgi, Daniel
daniel.kababgi@stud-mail.uni-wuerzburg.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg

Einleitung

Die Computational Literary Studies (CLS) befinden sich als Disziplin zwischen der Literaturwissenschaft, der Informatik und der Computerlinguistik. Sowohl aus einer theoretischen als auch einer methodischen Perspektive spielen unterschiedliche Aspekte aus allen angrenzenden Fachbereichen in den CLS eine Rolle. So kommen in den CLS bspw. computergestützte Verfahren wie maschinelles Lernen und Annotationen mit literaturwissenschaftlichen Fragestellungen zum Einsatz und sorgen für eine starke heterogene Prägung des Forschungsfeldes. Entsprechend ergibt sich auch eine diverse Landschaft an genutzten Datentypen und -formaten sowie lebender Systeme, bspw. Software, Tools, Visualisierungen und Plattformen, die es im Sinne der FAIR Prinzipien (Wilkinson et al. 2016) zu managen gilt. Ebenso kombinieren sich Konventionen der unterschiedlichen, fachlichen Teildisziplinen der CLS in der Nutzung von Technologien, Infrastrukturen und der Publikation von Ergebnissen und Daten. Hieraus ergibt sich eine Heterogenität des Forschungsfeldes sowie des spezifischen Forschungsdatenmanagements, wie sie sich auch grundsätzlich in den Geisteswissenschaften allgemein darstellt (Pempe 2012).

Ausgangslage

Das DFG Schwerpunktprogramm 2207 „Computational Literary Studies“ (SPP CLS)[1] bildet mit 10 geförderten Einzelprojekten sowie einem assoziierten Projekt seit 2020 einen Teil der deutschsprachigen CLS-Community. Im Rahmen des Zentralprojekts des Programms werden die einzelnen Teilprojekte individuell und projektübergreifend beim fachspezifischen Forschungsdatenmanagement (FDM) unterstützt.

Zu diesem Zweck wurde unter anderem eine Landschaftsvermessung vorgenommen, bei der in drei durch einen Leitfaden (Helling et al. 2020) gestützten Interviewrunden und einer Reviewrunde die Teilprojekte zum Umgang mit Forschungsdaten und lebenden Systemen, aber auch zu disziplinspezifischen Methoden und alltäglicher Projektarbeit befragt wurden. Zentrale Ziele der Landschaftsvermessung im SPP CLS sind die Entwicklung und Umsetzung einer möglichst fachspezifischen FDM-Strategie für das SPP CLS, sowie die Entwicklung einer Handreichung zu Best Practices und eines Anforderungsprofils für relevante/benötigte Infrastrukturen in den CLS.

Identifizierte Herausforderungen im FDM für die CLS

Neben einer grundsätzlichen Heterogenität in Bezug auf Datenformate und lebende Systeme in den CLS, für deren langfristige Sicherung und Verfügbarmachung kaum fachspezifische Lösungen existieren,[2] konnte im Rahmen der Landschaftsvermessung insbesondere auch die zentrale Herausforderung kollaborativer Arbeit identifiziert werden. Viele Fragestellungen der CLS werden, wie in vielen anderen Fachdisziplinen auch, mit Hilfe diverser Methoden von interdisziplinären Teams aus Forschenden, gegebenenfalls an verschiedenen Standorten, bearbeitet. Dabei ist es wichtig gemeinsam an Daten und Dokumenten zu arbeiten, teilweise sogar gleichzeitig auf denselben Dateien.

Um die interdisziplinäre Zusammenarbeit zu fördern ist es daher unabdingbar, dass gemeinsam nutzbare Infrastrukturelemente verfügbar und leicht zugänglich sind. Institutionell aufgesetzte Versionskontrollsysteme, die den Zugang von Institutions-externen Forschenden nur über restringierte Gastzugänge zulassen, können dabei ebenso Hürden schaffen, wie verschiedene Vorgaben bezüglich der Nutzung von kommerziellen Angeboten oder proprietären Formaten.

Das Vorgehen und diese bisherigen Zwischenergebnisse der Landschaftsvermessung sowie pragmatische

Lösungsstrategien zum Umgang mit Forschungsdaten in den CLS, wie bspw. der Betrieb einer gemeinsamen Gitlab-Instanz, wurden bereits mit den Communities der Digital Humanities (Helling et al. 2022a; Helling et al. 2022b) und des geisteswissenschaftlichen Forschungsdatenmanagements (Helling et al. 2021) diskutiert.

Ziele des vorgeschlagenen Workshops

Der Workshop soll eine oft implizit angenommene Ebene beleuchten, die im alltäglichen Umgang mit Forschungsdaten regelmäßig für kleine oder größere Ärgernisse sorgt oder sogar bestimmte Vorgehensweisen verhindert: Gemeinsames Arbeiten auf interaktiven Plattformen, Datenaustausch, unterschiedliche Datenformate, fehlende fachspezifische Infrastrukturangebote für die Publikation und Archivierung von Forschungsergebnissen sowie nicht mehr verfügbare oder lauffähige lebende Systeme wie Werkzeuge und Plattformen – dieser Ist-Zustand führt unter Umständen an entscheidenden Stellen zu pragmatisch- technischen Entscheidungen. Wir möchten die Community einladen, Erfahrungen aus ihrem Forschungsalltag zu teilen und Hürden aufzuzeigen, um dann gemeinsam eine Vision zu entwickeln, was wir benötigen um Wunsch und Wirklichkeit in Bezug auf Forschungsinfrastrukturen für die CLS in Einklang zu bringen, damit die technisch unterstützende Ebene ihre Rolle erfüllt und nicht zum Verhinderer wird.

Entsprechend möchten wir mit unserem Workshop die Ergebnisse der FDM-Landschaftsvermessung im SPP CLS als Ausgangspunkt nehmen und die damit verknüpften Fragestellungen mit der breiteren CLS-Community diskutieren, um das bisher entwickelte FDM-Anforderungsprofil der CLS um bisher ungesehene Aspekte ebenso zu erweitern wie die Konturen der identifizierten Best Practices zu schärfen. Der Workshop soll als offenes Forum verstanden werden, in dem die CLS-Community einerseits konkrete Bedarfe und Herausforderungen im FDM adressiert und an einem spezifischen, praxis- und community-getriebenen FDM-Bedarfsprofil arbeitet. Andererseits soll der Workshop auch auf operativer Ebene einen konstruktiven Austausch zwischen Fachwissenschaftler*innen der CLS und Datenmanager*innen ermöglichen.

Vor dem Hintergrund einer Ausgangslage mit Datentypen, Formaten und Methoden die - bedingt durch die Diversität der spezifischen Forschungsfragen - hochgradig heterogen ist, sollen unter anderem folgende Fragen in den Fokus genommen werden:

- Was benötigen Forschende der Computational Literary Studies für die tägliche Arbeit mit Forschungsdaten?
- Wie gelingt die Zusammenarbeit über Fach- und Institutionsgrenzen hinweg?
- Welcher Angebote bedarf es für die Sicherung, den Zugang, die Reproduzierbarkeit und Nachnutzbarkeit von CLS Forschungsergebnissen?

Dabei soll der Blick nicht nur auf disziplinspezifischen Werkzeugen und Infrastrukturen, wie sie z.B. über Initiativen wie DARIAH-DE[3] und CLARIAH-DE[4] zur Verfügung gestellt werden, liegen, sondern auch auf der disziplinspezifischen Nutzung von generischer Infrastruktur wie bspw. dem Forschungsdatenrepositorium Zenodo[5] und der Softwareentwicklungs- und Versionskontrollplattform GitHub[6].

Ein besonderes Augenmerk soll in diesem Zusammenhang auf der Unabhängigkeit von kommerziellen / proprietären Infrastrukturen liegen:

- Gibt es Zusammenhänge zwischen erzeugten Datenformaten und -strukturen und genutzten, proprietären Systemen?
- Welche Bedingungen verhindern möglicherweise die Nutzung spezifischer FDM-Lösungen, bspw. aufgrund von rechtlichen und finanziellen Hürden oder mangelnder Nachhaltigkeit?

Dabei ist es ein Anliegen des Workshops die Erfahrungen der Forschenden der CLS zu nutzen um strukturell wie anekdotisch den Ist-Zustand im Bezug zu den Ergebnissen aus dem Schwerpunktprogramm zu kartografieren und dabei Wunsch und Wirklichkeit einer interdisziplinären, modularen und vernetzten Infrastruktur in Beziehung zu setzen. Nicht zuletzt soll es um die Aussicht gehen, was von den digitalen Erzeugnissen der CLS die Chance hat auch in mehr als zehn Jahren noch nachvollziehbar zu sein.

Entsprechend möchten wir alle CLS-Community-Mitglieder und Interessierte einladen mit uns eine Bedarfsskizze für die Vision einer fachspezifischen und für alle zugänglichen Forschungsinfrastrukturlandschaft anzufertigen, die

- Zusammenarbeit über Institutions- und (Bundes-)Ländergrenzen ermöglicht,
- Nachnutzbarkeit, Zugänglichkeit und Reproduzierbarkeit unterstützt sowie
- (Langzeit)Archivierung in den Blick nimmt.

Ablauf des Workshops

Der halbtägige Workshop wird in drei Teile gegliedert sein (siehe Tab. 1), die durch zwei 15-minütige Pausen strukturiert werden. Im ersten Teil führen wir in Thema und Begriffe ein und berichten über Erfahrungen und Ergebnisse aus dem Forschungsdatenmanagement im DFG Schwerpunktprogramm „Computational Literary Studies". Dieser Teil endet mit einer kurzen Onlineumfrage, in der der bisherige Umgang mit Methoden des Forschungsdatenmanagements sowie typische Problemfälle der Teilnehmenden abgefragt werden. Ähnlich dem Format der CRETA-Werkstatt (Reiter et al. 2020) werden wir im zweiten Teil Thementische zur Archivierungsinfrastruktur, Arbeitsinfrastruktur und lebenden Systemen anbieten, an denen in vor Ort gebildeten Gruppen Erfahrungen, Herausforderungen, Lösungen sowie Visionen und Wünsche formuliert und diskutiert werden können. Dabei wird jeder Tisch von einer*m der Workshop-Organisator*innen begleitet, um im dritten Teil des Workshops Umfrage und Ergebnisse der Thementische gemeinsam auszuwerten und Wunsch und Wirklichkeit

in einem gemeinsamen Anforderungsprofil zu beschreiben, das wir im Anschluss an den Workshop über Zenodo veröffentlichen werden.

Tabelle 1: Zeitplan des Workshops.

	Dauer	Inhalt
Teil 1		
	0-30 Min.	Begrüßung und Einführung in das Thema / den Workshop
	30-45 Min.	Durchführung Onlineumfrage
	45-60 Min.	Kaffeepause
Teil 2		
	60-150 Min.	Durchführung Thementische (jeweils 30 Min.)
	150-165 Min.	Kaffeepause
Teil 3		
	165-240 Min.	Zusammenführung der Ergebnisse: Formulierung eines gemeinsamen Anforderungsprofils

Neben dem unmittelbaren Bezug zum Forschungsdatenmanagement in den Computational Literary Studies und der Erweiterung der Ergebnisse aus der Landschaftsvermessung im SPP CLS soll der Workshop grundsätzlich zur Sichtbarkeit der FDM-Bedarfe der CLS-Community in Infrastrukturinitiativen wie dem NFDI-Konsortium Text+[7] und dem EU-geförderten CLS INFRA[8] Projekt beitragen.

Adressat*innen des Workshops

Der Workshop richtet sich an etablierte und potentielle Mitglieder der CLS-Community und Interessierte, die Erfahrungen auf ähnlichen Gebieten, mit interdisziplinären Methoden zur Untersuchung von Textgrundlagen haben und sich für die Methoden und Fragestellungen der CLS interessieren. Darüber hinaus möchten wir explizit auch Expert*innen im Bereich des geisteswissenschaftlichen Forschungsdatenmanagements einladen am Workshop teilzunehmen. Die maximale Teilnehmendenzahl beträgt 20. Bei größerem Interesse können die interaktiven Teile des Workshops ggf. in zwei bis drei Iterationen durchgeführt und für die Teilgruppen mit der Onlineumfrage verschachtelt werden.

Als technische Ausstattung wird vor Ort der Zugang zu Strom, stabilem Internet und einem Projektor mit Leinwand/Projektionsfläche benötigt. Um sich an der Online-Umfrage beteiligen zu können, sollten die Teilnehmenden über ein digitales Endgerät verfügen.

Organisator*innen des Workshops

Kerstin Jung (Conceptualization, Writing – original draft) promovierte in der Computerlinguistik zum Thema der aufgabenbezogenen Kombination von automatisch erstellten Syntaxanalysen. Sie arbeitet im Zentralprojekt des SPP CLS zur disziplinspezifischen Unterstützung des FDM und bringt Erfahrung aus verschiedenen Infrastrukturprojekten und der Koordination kollaborativer Annotationsvorhaben ein. Ihre Forschungsinteressen liegen im Bereich der Nachhaltigkeit von Sprachressourcen und Abläufen sowie Metadaten- und Annotationsformaten.

Steffen Pielström (Conceptualization, Writing – review & editing) ist promovierter Biologe und arbeitet seit fast 10 Jahren im Bereich der Evaluation, Entwicklung und Vermittlung von quantitativen Methoden für die computergestützte Textanalyse in den Geisteswissenschaften. Er hat an verschiedenen Infrastrukturprojekten für die Digital Humanities mitgewirkt und ist zur Zeit im Zentralprojekt des SPP CLS tätig.

Patrick Helling (Conceptualization, Writing – review & editing) ist Medienwissenschaftler und Medieninformatiker. Er arbeitet im Zentralprojekt des SPP CLS und ist für die Entwicklung einer umfassenden FDM-Strategie für das Schwerpunktprogramm zuständig. Darüber hinaus ist er bereits seit 2017 am Data Center for the Humanities (DCH) an der Universität zu Köln tätig und dort Teil des FDM-Beratungsteams. Patrick Helling verfügt über Expertise im geisteswissenschaftlichen Forschungsdatenmanagement. Im Rahmen seiner Promotion arbeitet er an der Entwicklung eines formalen Beschreibungsmodells für das Management von Forschungsdaten.

Daniel Kababgi (Writing – review & editing) ist Masterstudent an der Universität Würzburg für Digital Humanities und Germanistik. Sein Studienschwerpunkt liegt auf NLP und dessen Anwendung innerhalb der Literaturwissenschaften. Das Hauptaugenmerk liegt auf der distanzierten Betrachtung der Literatur des 18. und 19. Jahrhunderts im Bezug auf die literarischen Epochen der Aufklärung und der Romantik.

Fußnoten

1. https://dfg-spp-cls.github.io/ (letzter Zugriff: 02. August 2022).
2. Im Gegensatz zu anderen Fachbereichen und Disziplinen, wie bspw. die Lingustik (siehe u. a. Blumtritt und Rau 2018) oder die Medienwissenschaften (siehe u. a. Matuszkiewicz 2022).
3. https://de.dariah.eu/ (letzter Zugriff: 02. August 2022).
4. https://www.clariah.de/ (letzter Zugriff: 02. August 2022).
5. https://zenodo.org/ (letzter Zugriff: 02. August 2022).
6. https://github.com/ (letzter Zugriff: 02. August 2022).
7. https://www.text-plus.org/ (letzter Zugriff: 02. August 2022).
8. https://clsinfra.io/ (letzter Zugriff: 02. August 2022).

Bibliographie

Blumtritt, Jonathan und Felix Rau. 2018. " *Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten."* In *DHd2018: Kritik der digitalen Vernunft.* https://doi.org/10.5281/zenodo.4622314.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2021. "Disziplinspezifisches Forschungsdatenmanagement. FDM-Bedarfserfassung in den Computational Literary Studies." In *FORGE 2021 - Forschungsdaten in den Geisteswissenschaften: MAPPING THE LANDSCAPE - Geisteswissenschaftliches Forschungsdatenma-*

nagement zwischen lokalen und globalen, generischen und spezifischen Lösungen. Konferenzabstracts. 83-95. https://doi.org/10.5281/zenodo.5379629.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2022a. "Making Research Data FAIR. Seriously? Reflections on Research Data Management in the Computational Literary Studies." In *Digital Humanities 2022 Conference Abstracts*, 230-233.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2022b. "Pragmatisches Forschungsdatenmanagement - Qualitative und Quantitative Analyse der Bedarfslandschaft in den Computational Literary Studies". In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, 193-199. https://doi.org/10.5281/zenodo.6328021.

Helling, Patrick, Kerstin Jung, Nils Reiter und Steffen Pielström. 2020. "Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies." Zenodo. https://doi.org/10.5281/zenodo.4269639.

Matuszkiewicz, Kai. 2022. " *Forschungsdaten in den Medienwissenschaften: Eine Auswertung von qualitativen Interviews zur Bedarfsermittlung für die Gestaltung eines medienwissenschaftlichen Forschungsdatenrepositoriums."* In *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern 2/2022,* 1-14. https://doi.org/10.17192/bfdm.2022.2.8433.

Pempe, Wolfgang. 2012. *"Geisteswissenschaften."* In *Langzeitarchivierung von Forschungsdaten: eine Bestandsaufnahme*, 137–60. *Boizenburg: vwh, Verlag Werner Hülsbusch.*

Reiter, Nils, Gerhard Kremer, Kerstin Jung, Jansi Pagel, Axel Pichler und Benjamin Krautter. 2020. "Reaching out: Interdisziplinäre Kommunikation und Dissemination: Ein CRETA-Erfahrungsbericht" In *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt hg. von Nils Reiter, Axel Pichler und Jonas Kuhn, 467-484. Berlin, Boston: De Gruyter.* https://doi.org/10.1515/9783110693973-019.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data 3. Article number: 160018.* https://doi.org/10.1038/sdata.2016.18.