Disziplinspezifisches Forschungsdatenmanagement

FDM-Bedarfserfassung in den Computational Literary Studies

Patrick Helling | Kerstin Jung | Steffen Pielström
Institut für Deutsche Philologie
Lehrstuhl für Computerphilologie und neuere deutsche Literaturgeschichte
Universität Würzburg

SPP 2207 Computational Literary Studies

FORGE Konferenz | Universität zu Köln | 08.-10. September 2021

Digitale Geisteswissenschaften | in a Nutshell

Computational LiteraryStudies

- **Computational Literary Studies** = Anwendung computergestützter, quantitativer Forschungsmethoden zur Analyse digitalisierter literarischer Texte zur Untersuchung literaturwissenschaftlicher Fragestellungen

Literaturwissenschaften

Computerlinguistik

Informatik

DFG Schwerpunktprogramm SPP 2207 „Computational Literary Studies“ (SPP CLS)

+ Zehn Teilprojekte (+ ein assoziiertes Projekt) an unterschiedlichen Standorten in Deutschland und der Schweiz
+ Ein koordinierendes Zentralprojekt

SPP CLS | die Teilprojekte

Computational LiteraryStudies

- **Zeta and Company – Measures of Distinctiveness for Computational Literary Studies**
  (Christof Schöch | Universität Trier)
- **Was ist wichtig? Schlüsselstellen in der Literatur**
  (Robert Jäschke & Steffen Martus | Humboldt-Universität zu Berlin)
- **Structuring Literature – Variants and Functions of Reflective Passages in Narrative Fiction**
  (Anke Holler, Caroline Sporleder & Benjamin Gittel |Georg-August-Universität Göttingen)
- **Relating the Unread – Network Models in Literary History**
  (Ulrik Brandes & Thomas Weitin | ETH Zürich/TU Darmstadt)
- **Quantitative Drama Analytics: Tracking Character Knowledge (Q:TRACK)**
  (Nils Reiter | Universität Stuttgart)
- **Evaluation Events in Narrative Theory**
  (Evelyn Gius & Chris Biemann TU Darmstadt/Universität Hamburg)
- **Emotions in Drama**
  (Christian Wolff & Katrin Dennerlein | Universität Regensburg/Universität Würzburg)
- **Computer-aided Analysis of Unreliability and Truth in Fiction – Interconnecting and Operationalizating Narratology (CAUTION)**
  (Jonas Kuhn & Janina Jacke | Universität Stuttgart/Universität Hamburg)
- **CHYLSA (Children‘s and Youth Literature Sentiment Analysis)***
  (Berenike Herrmann, Arthur Jacobs, Gerhard Lauer & Jana Lüdtke | Georg-August-Universität Göttingen/Freie Universität Berlin/Universität Basel)
- **The beginnings of modern poetry – Modeling literary history with text similarities**
  (Simone Winko & Fotis Jannidis | Georg-August-Universität Göttingen/Universität Würzburg)
- **Anomaly-based large-scale analysis of style and genre reflected in the use of stylistic devices in medieval literature**
  (Joachim Denzler & Sophie Marshall | Universität Jena)

SPP CLS | das Zentralprojekt

Computational LiteraryStudies

Aufgabenbereich A (Koordinierung)

- Betreuung aller Teilprojekte des SPP CLS
- Beförderung des inhaltlichen Austausches sowie der Kooperation sowohl innerhalb als auch außerhalb des Schwerpunktprogramms
- Organisation von zentralen Aktivitäten, wie bspw. General Meetings, fachspezifische Workshops, technische Schulungen, Veranstaltungen zur Nachwuchsförderung, Vernetzungstreffen etc.

Aufgabenbereich B (Forschungsdatenmanagement)

- aktive Beratung und Unterstützung der Teilprojekte bei Fragen des Forschungsdatenmanagements
- Entwicklung und Umsetzung einer passgenauen und bedarfsorientierten Datenstrategie für das gesamte Schwerpunktprogramm

SPP CLS | die Landschaftsvermessung

Computational LiteraryStudies

- Durchführung von leitfadengestützten Interviews mit Vertreter*innen aller Teilprojekte (virtuell)
- Gesprächsleitfaden = insgesamt 47 offene und nach Projektphasen gruppierte Fragen:

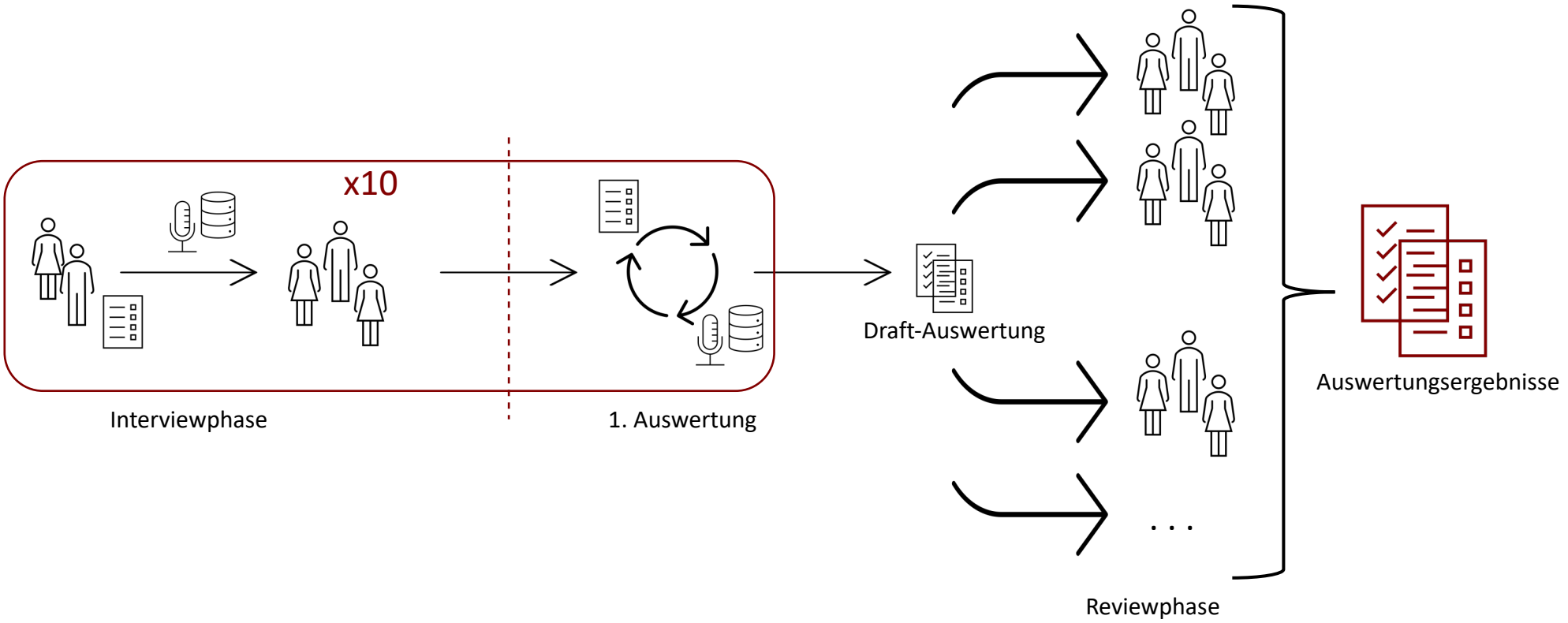
- Block A
  - Tägliches Arbeiten mit Daten im Projekt, u.a.
    - Nutzung von Tools, Programmiersprachen, virt. Umgebungen
    - Methoden und Analyseverfahren
  - Datenmanagement im Projekt, u.a.
    - Kollaboration und Zusammenarbeit
    - Backup- und Sicherungsstrategien, Datenaustausch
  - Entwicklung lebender Systemen, u.a.
    - Typ und Funktion von lebendem System
    - genutzte Technologie-Stacks

Block B

- Umgang mit Daten am Ende des Projekts, u.a.
  - bestehende Archivierungsstrategien (+ qualitative/quantitative Rahmenbedingungen)
  - bestehende Publikationsstrategien (+ qualitative/quantitative Rahmenbedingungen)
- Umgang mit lebenden Systemen am Ende des Projekts, u.a.
  - Strategien zum langfristigen Hosting und Betrieb lebender Systeme
  - Betreuung über die Projektphase hinaus

SPP CLS | die Vorgehensweise

Computational LiteraryStudies



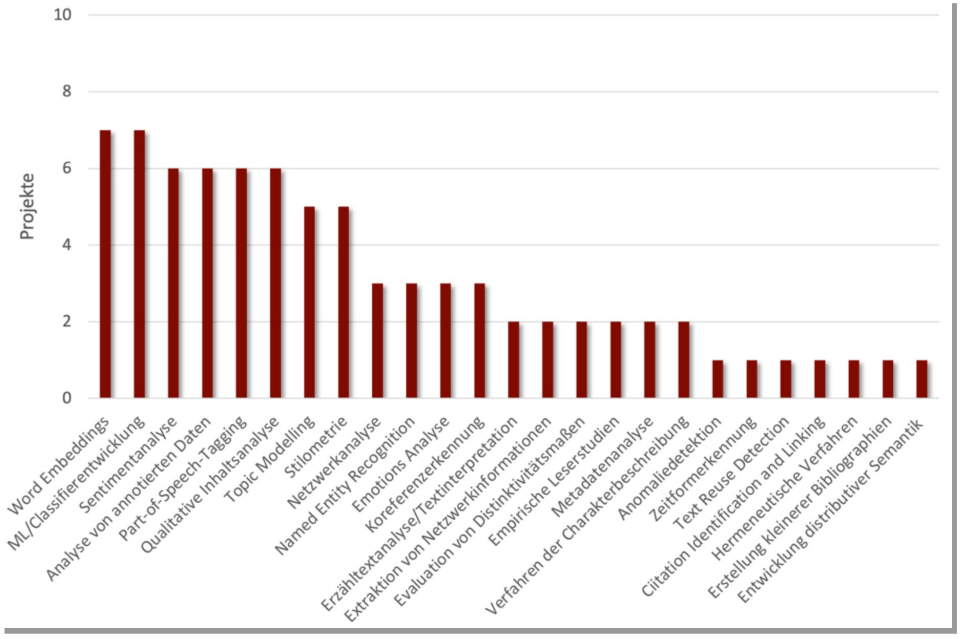
SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

Tägliche Arbeiten im Projekt

Tägliche Arbeiten	Projekte
Schreiben von Papern	8
Erstellung Annotationen	8
Datenbereinigung	8
Metadatenerstellung	7
Automatische Annotation	7
Aufbau Korpora	7
Erstellung Annotationsrichtlinien	6
Statistische Auswertungen	6
Datenkonvertierung	5
OCR	4
Datensichtung	4
Erstellung Trainingsdaten	4
Scannen von Text	3
Metadatensammlung	3
Schreiben von Dokus	3
Bewertung Verfahrensergebnis	3
Textauswahl	3
Metadatenanreicherung	2
Metadatenverarbeitung	2

Analyseverfahren und methodische Werkzeuge



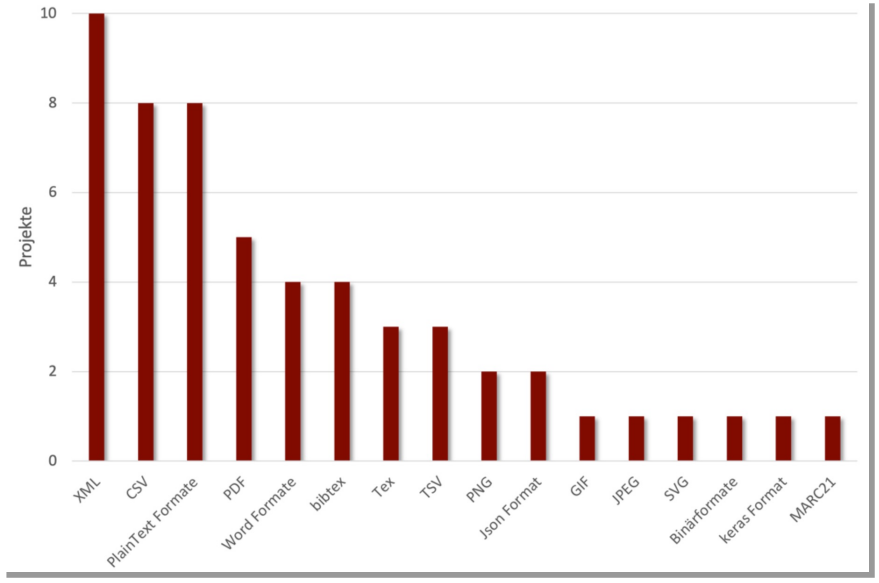
SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

Genutzte Datentypen innerhalb des SPP CLS

Datentypen	Projekte
Text	10
Softwarecode	9
Numerische Daten	6
Bilddaten	5
Bibliographische Daten	4
Wörterbücher/Listen	2
Interviewdaten	2
Netzwerkdaten	1

Genutzte Datenformate innerhalb des SPP CLS



SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

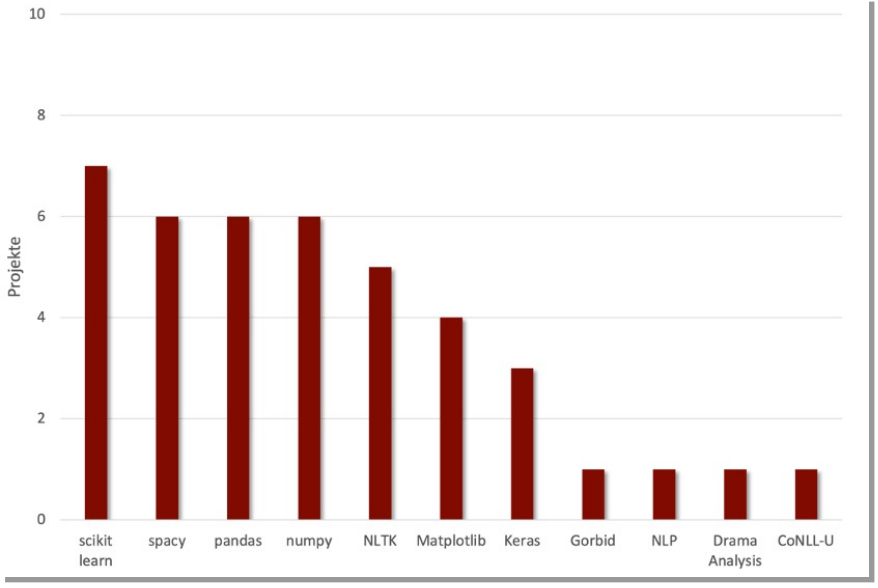
Programmier- und Skriptsprachen im SPP CLS

Programmier-/Skriptsprache	Projekte
Python	9
R	4
Shell-Skripte	4
Java	3
X-Technologien	2
JavaScript	2
HTML	2
CSS	2
SQL	1

SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

Genutzte Bibliotheken innerhalb des SPP CLS



Programmierumgebungen im SPP CLS

Programmierumgebungen	Projekte
Jupyter Notebooks	5
div. Editoren	5
oXygen	2
RStudio	2
Spyder IDE	1
Google Colab	1
Kaggle	1
Eclipse	1

SPP CLS | qualitative Beschreibung I

Computational LiteraryStudies

Analysen und Methoden

- Methodische Schwerpunkte/häufig angewendete Analyseverfahren konnten identifiziert werden
  (insb. Word Embeddings, ML/Classifierentwicklung, Sentiment Analyse, Analyse annotierter Daten, PoS-Tagging, qualitative Inhaltsanalyse)
- Dennoch konnte eine **starke Heterogenität der methodischen Landschaft** innerhalb der CLS sichtbar gemacht werden

➢ **Einrichtung methodenspezifischer Arbeitsgruppen**: Word Embeddings, Annotationen, Sentiment Analyse, Textähnlichkeiten

Tägliches Arbeiten mit Daten

- Interviews wurden kurz nach Beginn der Projekte geführt: Starke Häufigkeit vorbereitender, täglicher Arbeiten
- **Kernaufgaben/-tätigkeiten in den Literaturwissenschaften/CLS** wurden dennoch häufig erfasst
  (bspw. Erstellung von Annotationen/automatische Annotationen, Erstellung Annotationsrichtlinien, OCR etc.)

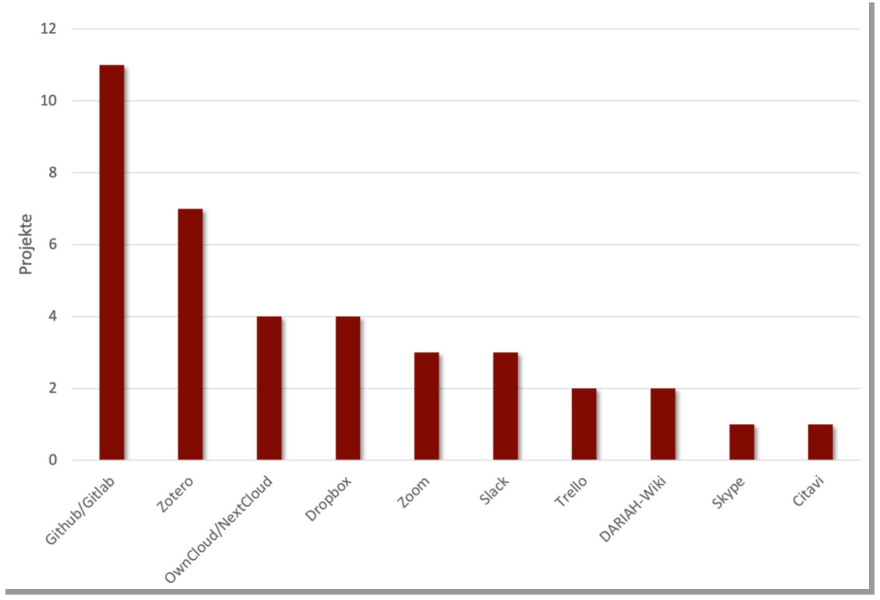
Daten, Software und Umgebungen

- Eindeutige **Dominanz von Text und XML** als allgemeiner Standard sowie **Software/Code** und dabei **Python** als meistgenutzte Programmiersprache
- Breites Spektrum an Datentypen und –formaten über **numerische Daten** und **CSV, PlainText und TSV** bis hin zu **Bilddaten** und **GIF, JPEG sowie SVG**
- Viele **unterschiedliche Bibliotheken und Programmierumgebungen**
  (bspw. scikit learn, spacy, pandas / Jupyter Notebooks, oXygen bzw. unterschiedliche Editoren)

SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

Genutzte Umgebungen innerhalb des SPP CLS



Annotationstools im SPP CLS

Annotationstool	Projekte
Catma	5
TreeTagger	1
CorefAnnotator	1
Sentiment Analyzer	1

SPP CLS | qualitative Beschreibung II

Computational LiteraryStudies

Arbeitsumgebung/Tools

- Zentrale Nutzung von min. **Github und/oder Gitlab** in allen Projekten
- **Bedarfe/Anforderungen** an Git-Systeme allerdings **häufig höher** als finanzielle Ausstattung ermöglicht
  (bspw. Anzahl Nutzende, Speicherkapazitäten, ggf. rechtlich konformer Speicherort, LFS)

➢ Einrichtung und Betreuung einer **zentralen Gitlab-Instanz** auf einer virtuellen Maschine am Rechenzentrum der Universität Würzburg
  - ➢ Sicherheit
    - ➢ organisatorische und technische **Betreuung/Verwaltung durch das Zentralprojekt** mit eigener Nutzer*innenverwaltung
    - ➢ **institutioneller Speicherort** (ink. Backups der gesamten VM)
  - ➢ Speicherkapazitäten
    - ➢ Speicherplatz im **hohen zweistelligen GB-Bereich**
    - ➢ Unterstützung von **Large File Storage (LFS)** ermöglicht Sicherung und Austausch großer Datenpakete und umfassenden Modellen

- **Nutzung vieler verschiedener Kommunikations- und Messangertools** innerhalb der einzelnen Projekte

➢ Einrichtung einer eigenen **Mattermost-Instanz via Gitlab—Server** an der Universität Würzburg als zentrales Angebot zur Kommunikation

Annotationstools

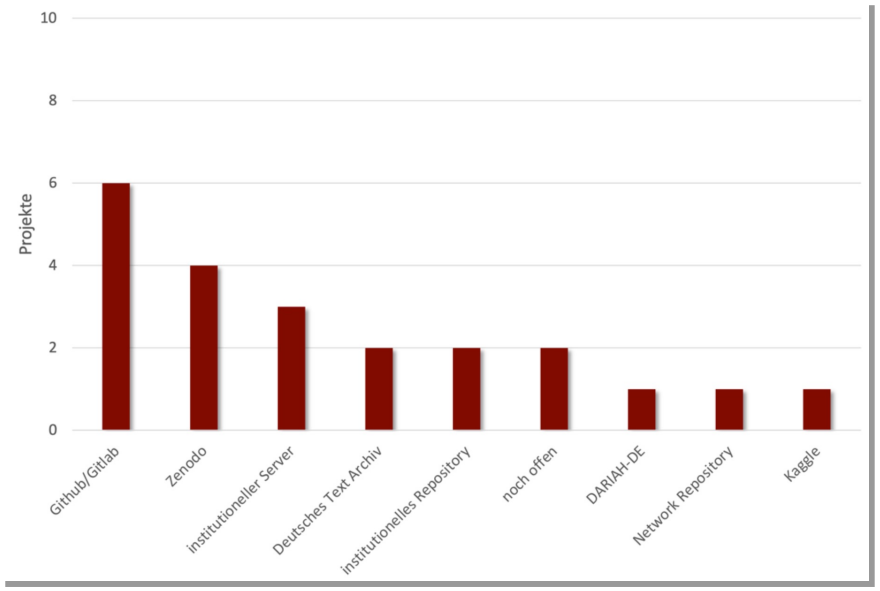
- Es konnte die **Nutzung des Annotationstools Catma** als Trend und somit **der Catma-Datenstandard** im Hinblick auf die Archivierung, Publikation und Nachnutzung von Annotationen im späteren Verlauf der Projekte identifiziert werden

SPP CLS | quantitative Ergebnisse

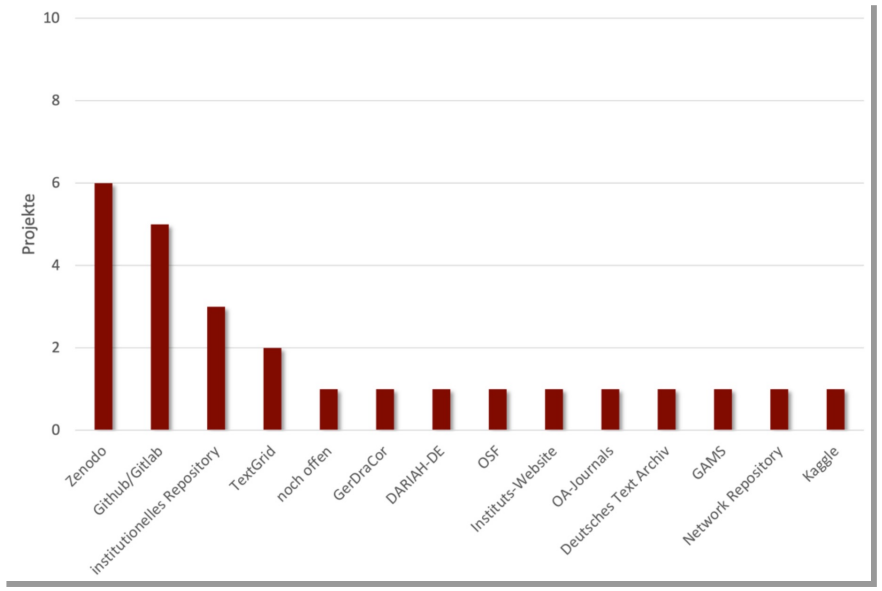
Computational LiteraryStudies

In der Evaluation

Archivierungsansätze im des SPP CLS



Publikationsansätze im SPP CLS



SPP CLS | qualitative Beschreibung III

Computational LiteraryStudies

In der Evaluation

Publikations- und Archivierungsstrategien

- vereinzelte **Nutzung von spezifischen Lösungen** (bspw. Deutsches Text Archiv, Network Repository, TextGrid, GerDraCor etc.)
- **keine fachspezifischen Angebotsstrukturen und –lösungen** in den Bereichen der Archivierung und der Publikation in der Breite
- Es gibt eine sehr **heterogene Daten- und Bedarfslandschaft** und entsprechend noch **blinde Flecken/Versorgungslücken** im fachspezifischen Forschungsdatenmanagement innerhalb der Computational Literary Studies

➢ **Handreichung** zum **Umgang mit Forschungsdaten** und zur Entsprechung der **guten wissenschaftlichen Praxis** innerhalb des Fachbereichs der CLS
➢ **FAIR in der Praxis** – Beitragsreihe zur Anwendung der FAIR-Prinzipien im Fachbereich der CLS

➢ Einrichtung und Verwaltung einer eigenen **Zenodo-Community für das SPP CLS**
  ➢ **Erfassung möglichst aller Publikationen** mit Bezug zum Schwerpunktprogramm
  ➢ Definition einer eigenen **Kuration- und Metadaten-Policy**

➢ Unterstützung bei der **Nutzung von Github/Gitlab** in Kombination **mit Zenodo** + ggf. **institutioneller Archivierungsstrategie**

SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

In der Evaluation

Geplante lebende Systeme im SPP CLS

Lebende Systeme	Projekte
Website	7
Tools/Anwendungen	3
Bibliotheken	1
Dashboard	1

SPP CLS | quantitative Ergebnisse

Computational LiteraryStudies

In der Evaluation

Funktionsumfang der lebenden Systeme im SPP CLS

LS Funktionsumfang	Projekte
Zugangsschicht zu Daten	8
Nachnutzung Ergebnisse	5
Visualisierungen	3
Projektpräsentation	2
Testumgebung für Verfahren	1
Textmarkierungen	1
Annotation	1
Analyse	1

SPP CLS | qualitative Beschreibung IV

Computational LiteraryStudies

In der Evaluation

Umgang mit lebenden Systemen

- Erfassung eines **Status Quo** bei der Entwicklung von lebenden Systemen innerhalb der CLS
  (insb. Websites und kleinere Tools)
- **Heterogenität der Systeme** und ihrer Funktionen **bedarf möglichst spezifischer Lösungsansätze**
- große Herausforderungen sind noch nicht gelöst
  (bspw. langfristiges Hosting, Betreuung und Kuration, nachhaltiger Dauerbetrieb)

- Allgemeine Beratung und Hinweise auf Best Practices
  - Nutzung **offener, gut dokumentierte**r und **weit verbreiteter Technologie-Stacks**
  - nach Möglichkeit: **Reduktion dynamischer** und **wartungsintensiver Visualisierungs- und Präsentationsschichten**
  - **Veröffentlichung** gut dokumentierten **Quellcodes** unter möglichst offenen Lizenzen

- Mögliche Strategie: Publikation via Github und Zenodo
  - Websites: **Statische HTML-Version** via **Github** + finaler Zustand persistent auf **Zenodo**
  - Tools/Anwendungen: **Dokumentierter Quellcode** oder **Docker Images** via **Github** + finaler Zustand persistent auf **Zenodo**

SPP CLS | Ausblick

Computational LiteraryStudies

Die Landschaftsvermess

- Qualitative und quantit ngs- und Datenlandschaft innerhalb des SPP CLS
  - Ableitung von Be Trends und Standards (bspw. in Bezug auf genutzte Tools, Umgebunge

Gemeinsame Datenst

- Entwicklung von **Bes** ergebnissen innerhalb des SPP CLS sowie einer **umfass**
  - Aktueller Sta Publikationsstrategien sowie des Status lebender Sys

  ➢ Unterstütz **s- und Publikationsstrategie**

  ➢ **Integratio**

  Ziel:

  ➢ Erfassung, Sammlung und persistente Verfügba en, Archivierungen und Nachnutzbarmachung via die statische Website des SPP CLS (G

Publications
…
Presentations
…
Dissemination
…
Annotation guidelines
…
Tools and plattforms
…
Datasets and analyses

SPP CLS | Ausblick

Computational LiteraryStudies

Bisherige Bilanz

- **institutionell**
  lokal betriebene, generische Lösungsstrategie für die **Kollaboration innerhalb der Projekte** → Gitlab und Mattermost
- **global ausgerichtet, generisch**
  Lösungsstrategie für die **Publikation** von Forschungsergebnissen → SPP CLS Zenodo Community
  Lösungsstrategie für die **Publikation** von Forschungsdaten/Software → Github + SPP CLS Zenodo Community
- **(teil)-spezifisch**
  Unterstützung der Projekte im **Umgang mit Forschungsdaten** → FAIR in der Praxis Beiträge, Handreichung zur guten wissenschaftlichen Praxis
- **fachspezifisch**
  vereinzelte Lösungen innerhalb einzelner Projekte

  (bspw. GerDraCor, Deutsches Textarchiv, Network Repository)

Aufgrund des Mangels an tatsächlich dezidiert fachspezifischer Angebotsstrukturen und Servicestrategien für die **Computational Literary Studies** in ihrer Breite bedarf es für eine Datenstrategie (zur Zeit) an Lösungen, die sowohl **lokal als auch global** und **sowohl generisch als auch (fach)-spezifisch** ausgerichtet und angesiedelt sind!

SPP CLS | Vielen Dank für Ihre Aufmerksamkeit

Computational LiteraryStudies

SPP 2207 Computational Literary Studies

E-Mail: cls-fdm-coordination@clariah.de
Twitter: @spp_cls

Patrick Helling | Kerstin Jung | Steffen Pielström
Institut für Deutsche Philologie
Lehrstuhl für Computerphilologie und neuere deutsche Literaturgeschichte
Universität Würzburg