GitMA oder CATMA für Fortgeschrittene

Projektdaten via Git abrufen und mittels Python-Bibliothek weiterverarbeiten

Schumacher, Mareike
schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Vauth, Michael
vauth@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Gerstorfer, Dominik
gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Meister, Malte
meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Dieser CATMA-6-Workshop richtet sich an fortgeschrittene CATMA User*innen mit Vorkenntnissen in digitaler Annotation, die im Rahmen der eigenen Arbeit oder von Forschungsprojekten mit größeren Mengen von Annotationsdaten operieren (wollen). Im Zentrum steht die Weiterverarbeitung und Analyse von Annotationsdaten. Wie greife ich über Git auf meine CATMA-Annotationsdaten zu? Wie erstelle ich individuelle, interaktive Visualisierungen meiner Annotationsdaten? Wie berechne ich die Übereinstimmung zwischen mehreren Annotator*innen? Diese und ähnliche Fragen werden während des Workshops beantwortet.

CATMA (Gius et al. 2021) ist eine webbasierte, kollaborative Textannotations- und Analyse-Plattform, die seit 2008 an der Universität Hamburg und im Rahmen des DFG-geförderten Projektes forTEXT seit 2020 an der Technischen Universität Darmstadt entwickelt wird.[1] Hauptzielgruppe sind traditionell-analog arbeitende Geisteswissenschaftler*innen, die über eine intuitiv bedienbare GUI Texte annotieren und analysieren können. Mit dem Release von CATMA 6 im Jahr 2019 wurde für die Plattform ein auf Git basierendes Backend eingeführt. Für zahlreiche Projekte, die bereits auf sehr fortgeschrittenem Niveau CATMA nutzen, und Interessierte aus der Digital-Humanities-Community mit Erfahrung in der Nutzung von Git und Grundkenntnissen in Python eröffnet sich dadurch eine Reihe neuer Funktionen, die es in bisherigen CATMA-Versionen nicht gab. Einige dieser Funktionen werden im Laufe dieses Ganztagesworkshops vorgestellt und vermittelt.

Der Workshop bietet:

- kurze Einführung in die Nutzung von CATMA über das graphische Userinterface
- Kennenlernen der Datenstrukturen des Backends
- Zugriff auf das Backend mit Git
- Weiterverarbeitung der Daten mit Hilfe eines zur Verfügung gestellten Python-Packages

Annotation in CATMA 6 – projektorientiert, gemeinsam, vielfältig

Eine der wichtigsten Neuerungen von CATMA 6 gegenüber früheren Versionen ist die Umstellung auf eine projektzentrierte Nutzungsarchitektur. Am Beginn der Arbeit mit CATMA steht das Anlegen eines Projektes mit beliebig vielen Dokumenten, die analysiert werden sollen, und beliebig vielen Team-Mitgliedern, die daran arbeiten wollen. Zur Annotation können eigene Taxonomien entworfen oder auf der Plattform *fortext.net* bereitgestellte Ressourcen genutzt werden. Die Annotationskategorien können frei gestaltet werden und jede Passage im Text kann frei damit annotiert werden. Einzelne und Merfachannotationen, einander überlagernde oder überlappende Annotationen oder sogar widersprüchliche Annotationen – in CATMA ist durch die Speicherung der Daten als Standoff-Markup vieles möglich. Eine weitere Neuerung im Funktionsumfang ist die Möglichkeit, Textstellen und Annotationen zu kommentieren. Offene Fragen, nicht zu Ende gedachte Interpretationsansätze oder auch der Austausch mit den anderen Team-Mitgliedern können über die Kommentarfunktion in den Annotationsprozess integriert werden. Sowohl Annotationen als auch Kommentare können über die Analyse-Funktionen von CATMA durchsucht, in tabellarische Form gebracht oder visualisiert werden. Der Umfang dessen, was über die CATMA-GUI umgesetzt werden kann, ist also recht groß. Und doch macht die Einführung des auf Git basierenden Backends das Tool für die Digital-Humanities-Community erst richtig interessant. Der undogmatische Zugang, der bisher nur zu Annotationen und Annotationstaxonomien ermöglicht wurde, erstreckt sich nun bis zu den Annotationsdaten und der Weiterverarbeitung derselben (siehe Abbildung 1). Dieser neue Teil des CATMA-Workflows wird in diesem Workshop vermittelt werden.

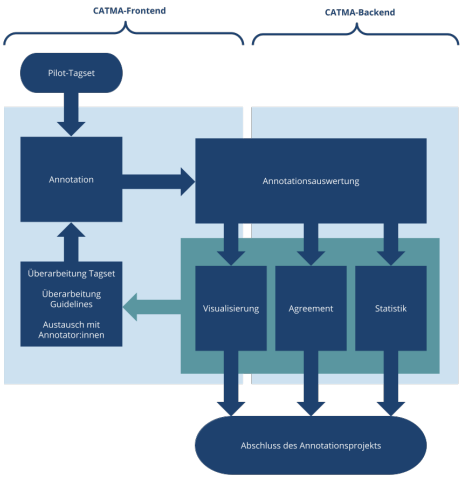

Abb. 1: Im Workshop vermittelter Workflow zur Annotationsauswertung und -überarbeitung mit dem CATMA-Backend

Standards und Best Practices nicht aus den Augen verlieren mit GitMA

Niedrigschwelligkeit und Nähe zu traditionell-analogen Methoden der Geisteswissenschaften sind nach wie vor wichtige Grundsätze, die in CATMA implementiert sind. Doch mit zunehmender Verbreitung des Tools in den digitalen Geisteswissenschaften sind neben der Möglichkeit zu hermeneutisch-vielfältiger Textanalyse auch die Einhaltung von Best Practices und Standards, die innerhalb der Digital-Humanities-Community entwickelt wurden, von Bedeutung. Eine Verschmelzung von CATMA und Git zu “GitMA” ermöglicht beides. Dabei bleibt der Annotationsprozess selbst völlig frei gestaltbar. Die resultierenden Daten aber können zum Beispiel nach der Übereinstimmung der Annotierenden untereinander ausgewertet werden. Es ist möglich eine der Annotationen als ‘Silver Annotation’ festzulegen und die anderen daran zu messen. Das festgestellte Disagreement kann zur Grundlage eines Disagreement-Tagsets werden, das über das Backend auch wieder ins Frontend der CATMA-GUI zurückgespielt werden kann (siehe Abb. 1). Dasselbe gilt für die nicht übereinstimmend annotierten Passagen, welche wiederum selbst durch Annotationen dargestellt/hervorgehoben werden können. So ergibt sich ein harmonischer Workflow vom Frontend zum Backend und zurück, der in Zukunft auch die Erstellung von Goldannotationen unterstützen wird.

Die GitMA-Funktionalitäten werden im Rahmen dieses Workshops erstmals einem Fachpublikum vorgestellt. Neben der Vermittlung von Nutzungskompetenzen möchten wir darum auch eine kritische Diskussion anregen. Feedback zu Idee und Umsetzung der CATMA-Backend-Nutzung sind uns überaus willkommen!

Format und Ablauf des Workshops

Der Workshop wird als ganztägiges hands-on Tutorial angeboten, das an einem oder an zwei aufeinander folgenden (halben) Tagen stattfinden kann.

Ablauf:

Teil 1

1. CATMA Backend (45 Minuten)
2. kurze Einführung in das CATMA-Frontend
3. Struktur: Tagsets, Documents, Annotation Collections
4. Annotationsrepräsentation: JSON-Files
5. Zugriff auf Annotationsdaten über Git (45 Minuten)
6. wie clone ich ein CATMA Project?
7. wie update ich ein CATMA Project, um neue Annotationen zu laden?

Pause

1. Zugriff auf ein CATMA Project mit Python (45 Minuten)
2. Installation des Packages
3. Laden eines Projects
4. Zugriff auf Annotation Collections, Dokumente und Tagsets

Teil 2

1. Annotationsauswertungen (90 Minuten)
2. Visualisierungen zum Annotationsfortschritt und zur Exploration von Annotationen (Plotly)
3. IAA Auswertung von zwei Annotation Collections des gleichen Dokuments (15 Minuten)
4. weiterführende Auswertungen mit Pandas

Pause

1. Unterstützung der Goldannotation (75 Minuten)
2. Festlegung der Silver Annotations
3. Umgang mit Annotationsspannen
4. Automatische Erstellung eines Disagreement Tagsets
5. Darstellung von Disagreement als Annotationen
6. Manuelle Überabeitung von automatischen Goldannotationen
7. Diskussion und Feedback (60 Minuten)

Zielgruppe:

Nutzer*innen, die Annotationen mit CATMA in Forschungsprojekten oder Lehrsituationen managen, sowie alle, die einen schnellen Workflow zwischen Annotation bzw. Annotationsbearbeitung und Annotationsauswertung benötigen.

Zahl der möglichen Teilnehmer*innen:

30

Technische Voraussetzungen:

Die benötigten Vorinstallationen von Git, Anaconda und Plotly können durch die Bereitstellung eines Docker-Image vermieden werden. Die Teilnehmer*innen sollten die Installation von Docker selbst auf einem eigenen Laptop (Touch Devices werden nicht unterstützt), den sie zum Workshop mitbringen, möglichst schon erledigt haben. Für die Durchführung des Workshops benötigen wir außerdem einen Beamer.

Zur Vorbereitung sollten Teilnehmer*innen außerdem schon einen CATMA-Account erstellt (unter *https://app.catma.de/catma/*) und sich mit der CATMA-Nutzung bekannt gemacht haben (z.B. mithilfe von der forTEXT-Lerneinheit zu CATMA 6: *Manuelle Annotation mit CATMA*). Wenn eigene CATMA-Annotationsdaten vorhanden sind, können diese während des Workshops analysiert werden. Für Teilnehmende, die nicht an eigenen Daten arbeiten möchten, stellen wir ein Demo-Projekt zusammen, dem man während des Workshops beitreten kann.

Benötigte Vorkenntnisse:

Die Teilnehmer*innen sollten über grundlegende Kenntnisse der Kommandozeile, Git und Python sowie Jupyter verfügen.

Beitragende

Michael Vauth, M.Ed.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Michael Vauth promoviert über “Zur Annotation intradiegetischen Erzählens. Binnenerzählungen im literarischen Werk Heinrich von Kleists” an der Technischen Universität Darmstadt. Er

ist wissenschaftlicher Mitarbeiter im Forschungsprojekt EvENT (Evaluating Events in Narrative Theory) an der Technischen Universität Darmstadt. Zuvor hat er an der Technischen Universität Hamburg im Projekt hermA (Automatisierte Modellierung hermeneutischer Prozesse - Der Einsatz von Annotationen für sozial- und geisteswissenschaftliche Analysen im Gesundheitsbereich) gearbeitet. Er beschäftigt sich insbesondere mit der digitalen Narratologie und der Methodik der Netzwerkanalyse.

Dominik Gerstorfer, M.A.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Dominik Gerstorfer promoviert über "Philosophische Fragen der Digital Humanities" an der Universität Stuttgart. Derzeit ist er im DFG-Projekt forTEXT tätig, zuvor war er im Digital-Humanities-Projekt CRETA in Stuttgart beschäftigt. Dominik hat an der Universität Tübingen Philosophie, Politikwissenschaften und Soziologie (M.A.) studiert. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftstheorie, formale Methoden und Argumentationsanalyse. Im Rahmen von forTEXT beschäftigt sich Dominik u.a. mit Intertextualität, Ontologien und der Entwicklung von Kategoriensystemen.

Malte Meister, B.Sc.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Malte Meister hat 2009 sein Informatik-Diplom (B.Sc.) in Kapstadt erworben. Im Rahmen des Abschlussprojekts für sein Diplom wurde er beauftragt, das Text-Annotations und -Analysetool CATMA, für die Universität Hamburg zu erstellen. Bis Anfang 2010 wirkte er im Team an CATMA mit, bevor er sich auf seine Karriere in der freien Wirtschaft konzentrierte. Nach mehr als zehn Jahren Berufserfahrung als Softwareentwickler und Teamleiter entschied er sich, wieder in die CATMA-Entwicklung einzusteigen. Er ist seit 2021 technischer Mitarbeiter an der TU Darmstadt und beschäftigt sich dort im Rahmen von forTEXT hauptsächlich mit dem Betrieb und der Weiterentwicklung von CATMA und den damit verbundenen Systemen.

Mareike Schumacher, M.A.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Mareike Schumacher koordiniert das DFG-Projekt forTEXT (*https://fortext.net*), in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die traditionelleren Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Sie promoviert als digitale Literaturwissenschaftlerin über Orte und Räume im Roman, beschäftigt sich besonders mit den Methoden des *distant reading* (u. a. *Named Entity Recognition* oder Stilometrie), der Digital-Humanities-Theorie und der Verbindung von digitalen Methoden und theoriebasierter Literatur- und kulturwissenschaftlicher Forschung.

Fußnoten

1. CATMA (Computer Assisted Text Markup and Analysis) erscheint zum Beispiel im *TAPoR Toolverzeichnis* , sowie in „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften“ (Frey-Endres & Simon 2021).

Bibliographie

Frey-Endres, Marcel / Simon, Tobias (2021): „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und3 Netzwerkanalyse in den Geisteswissenschaften“. In: *Digital Philology | Working Papers in Digital Philology 02|2021*. Darmstadt: TUPrints. URL: *https://tuprints.ulb.tu-darmstadt.de/17850/1/Digital_Philology__Working_Papers_in_Digital_Philology_vol002.pdf* [letzter Zugriff 24. November 2021]

Gius, Evelyn / Meister, Jan Christoph / Meister, Malte / Petris, Marco / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Gerstorfer, Dominik / Flüh, Marie / Horstmann, Jan (2021): CATMA 6 (Version 6.3). Zenodo. DOI: 10.5281/zenodo.1470118. URL: *https://catma.de/* [letzter Zugriff 24. November 2021]