Text+: Digitale Forschung auf der Grundlage von Text- und Sprachdaten bereichern

Barth, Florian
barth@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Genêt, Philippe
p.genet@dnb.de
Deutsche Nationalbibliothek, Deutschland
ORCID: 0009-0001-5095-8052

Körner, Erik
koerner@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0002-5639-6177

Leinen, Peter
p.leinen@dnb.de
Deutsche Nationalbibliothek, Deutschland
ORCID: 0000-0002-3014-000X

Weimer, Lukas
weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0001-6919-3646

Witt, Andreas
witt@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Mannheim, Deutschland
ORCID: 0000-0002-0299-5713

Ausgangslage und Zielsetzung

Infrastrukturen für die Wissenschaft können nur dann ihr Ziel erfüllen, wenn diese zusammen mit Forschenden und mit Blick auf deren Bedarfe konzipiert, aufgebaut und betrieben werden. Um diesem Anspruch gerecht zu werden, soll in diesem Workshop der dazu notwendige Austausch stattfinden.

Die Vision der Nationalen Forschungsdateninfrastruktur (NFDI) ist es, mit Daten als gemeinsames Gut, organisiert durch die Wissenschaft, die Grundlage für exzellente Forschung zu legen (NFDI, 2024). Dazu sollen die Nutzungsmöglichkeiten von Daten verbessert sowie Rahmenbedingungen für deren rechtskonforme, interoperable und nachhaltige Verwendung geschaffen werden, gesäumt von Schulungen, um Data Literacy zu stärken (vgl. NFDI, 2024). Das deckt sich mit Mission und Vision des Konsortiums Text+, das den Zugriff auf sprach- und textbasierte Forschungsdaten wissenschaftsgeleitet und bedarfsorientiert in all seinen diversen Communitys ertüchtigen und die Digital Literacy stärken will (Text+, 2024). Innerhalb von Text+ fokussiert sich die Task Area *Collections* auf sprach- und textbasierte Sammlungen und nimmt dabei alle Facetten einer FAIRen Verfügbarmachung in den Blick, wie deren Auffindbarkeit, ethische und rechtliche Fragestellungen, Angebote zur Datenablage oder Softwareservices zu deren (Weiter-)Bearbeitung.

Im Zuge dieser Bedarfsorientierung hatte Text+ im Vorfeld der Antragstellung einen Call for User Stories (Bertino, 2020) in seine Fachcommunitys lanciert, um infrastrukturelle Bedarfe zu adressieren und diese als Desiderata in den Antrag zu integrieren. In den User Stories wurde v.a. der Bedarf an interoperablen und nachnutzbaren Daten und Datenformaten laut sowie überhaupt die Zugänglichkeit zu diesen. Aber auch Unterstützung bei der Erstellung von Daten, bei Annotation und der Verlinkung von Daten wurde erbeten. Die User Stories kamen dabei aus allen Disziplinen der text- und sprachbasiert arbeitenden Geisteswissenschaften, v.a. aber aus den Sprach-, Literatur- sowie den Geschichtswissenschaften (zu Auswertungen vgl. Rißler-Pipka et al., 2021; Calvo Tello et al., 2021).

Im Workshop sollen konkrete Forschungsfragen, die durch die Nutzung der digitalen Infrastruktur Text+ besser beantwortet werden können, im Zentrum stehen. Einerseits müssen offen gebliebene Bedarfe aus User Stories in ihrer Aktualität überprüft und ggf. an veränderte Gegebenheiten angepasst werden. Andererseits hat sich die Forschungslandschaft seit dem Call for User Stories in gewohnt hoher Geschwindigkeit und zum Teil erheblich weiterentwickelt (aktuelle Beispiele sind sicher maschinelles Lernen und Large Language Models).

In unterschiedlichen Settings (Vorträge, interaktive Elemente) möchten wir daher die aktuellen Angebote von Text + präsentieren, an den Bedarfen der Wissenschaft spiegeln und auf diese Weise ein Feedback zu den derzeitigen Diensten und Services von Text+ erhalten. Hands-on-Demonstrationen vermitteln Forschenden einen praxisnahen Überblick über die Bandbreite der Angebote und sind eingeladen, direktes Feedback zu formulieren. Anhand konkreter Forschungsfragen wird zudem die Passung der Werkzeuge evaluiert. Hierbei werden die oben genannten User Stories ebenso wie neu formulierte Anforderungen einbezogen. Darüber hinaus werden wir im Workshop Elemente integrieren, die es uns erlauben, ein weitergehendes Verständnis der Anforderungen der Wissenschaft zu dokumentieren. Beide Aspekte dienen als Grundlage zur Weiterentwicklung des Portfolios von Text+.

Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370

Angaben zum Format

Der Workshop ist als ganztägige Veranstaltung geplant und enthält sowohl Vorträge als auch interaktive Elemente. In der ersten Session (90 Minuten) erfahren die Teilnehmenden zunächst, welche Entwicklungen und Angebote Text+ in der bisherigen Projektlaufzeit hervorgebracht hat. Anschließend stellen fünf Forschende ihre Forschungsfragen näher vor und erläutern, inwiefern das aktuell verfügbare Portfolio von Text+ ihre daraus resultierenden Bedarfe erfüllt bzw. wo sie noch Desiderate erkennen.

Der zweite Block (90 Minuten) besteht aus einer Postersession, in der weitere Anforderungen – darunter sowohl einige aus den User Stories von 2020 als auch aktualisierte oder gänzlich neu eingereichte (siehe Call for Contributions) – präsentiert werden. Hier steht der Austausch der Teilnehmenden, der Forschenden und der Workshop-Ausrichtenden von Text+ im Vordergrund. Ziel ist es, die Bandbreite der Forschungsfragen und Problemstellungen auszuloten, für die Text+ hilfreiche Tools und Angebote zur Verfügung stellen will.

Mit diesem Tableau vielfältiger Anforderungen im Hintergrund gehen die Teilnehmenden in den dritten Block des Workshops (90 Minuten). Dieser ist als hands-on-Session konzipiert, bei der an mehreren Stationen verschiedene Angebote von Text+ getestet werden können – insbesondere sind das die Instrumente des Text+ Data Space (z.B. Registry und Federated Content Search), NLP-Tools wie die MONAPipe, die LLM-Services der GWDG, die GND-Agentur mit ihrem Austausch- und Speicherformat entityXML und das Angebot an Forschungsprojekte der Übernahme, Archivierung und Verfügbarmachung von Forschungsdaten in den Text+ Zentren.

In der abschließenden Plenumsdiskussion (vierter Block, 60 Minuten) ziehen alle Beteiligten des Workshops gemeinsam Bilanz: Wo unterstützen die Angebote von Text+ die Forschung bereits effektiv? Wo besteht Optimierungsbedarf? Und wo sind noch Lücken im Portfolio?

Zielpublikum und Vorbereitung

Der Workshop richtet sich an Forschende aller Karrierestufen aus den Digital Humanities, die mit text- und sprachbasierten Daten arbeiten und diese analysieren, durchsuchen und nachhaltig zur Verfügung stellen möchten.

Eine inhaltliche Vorbereitung ist nicht notwendig. Allerdings sind alle Teilnehmenden eingeladen, Bedarfe aus der eigenen wissenschaftlichen Arbeit sowie Anregungen zur Erweiterung des Portfolios in den Workshop einzubringen. Einzelheiten dazu sind dem Call for Contributions zu entnehmen.

Benötigte technische Ausstattung

Der Raum, in dem der Workshop stattfindet, muss ausreichend groß sein, Tische und Bestuhlung müssen im Laufe des Tages mehrfach umgebaut werden können. Es werden für den Vortrags-Block ein Beamer mit Leinwand benötigt sowie ausreichend Tische und Stühle. Für die Postersession sind 15 Stellwände erforderlich. Im hands-on-Teil werden fünf größere Monitore, Stromanschlüsse für diese und mehrere Laptops benötigt. Zur Durchführung des gesamten Workshops ist zudem eine stabile WLAN-Verbindung notwendig. Die technische Ausstattung beschreibt ein Idealszenario; sollten einzelne Elemente nicht zur Verfügung stehen, werden sich die Workshop-Ausrichtenden um einfachere Lösungen bemühen.

Kontaktdaten und Forschungsinteressen der Beteiligten

Florian Barth ist wissenschaftlicher Mitarbeiter in der Abteilung Forschung und Entwicklung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen. Er ist verantwortlich für die Implementierung von Textverarbeitungspipelines sowie die Integration von Daten, Repository-Diensten und Sprachmodellen in die NFDI Text+.
Kontakt: barth@sub.uni-goettingen.de, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14, 37073 Göttingen

Philippe Genêt ist Theater-, Film- und Medienwissenschaftler und Amerikanist. Er arbeitet in der Deutschen Nationalbibliothek (DNB) und koordiniert dort alle Aktivitäten der DNB in der NFDI. Im Konsortium Text+ koordiniert er die Datendomäne Collections.
Kontakt: p.genet@dnb.de, Deutsche Nationalbibliothek, Informationsinfrastruktur, Adickesallee 1, 60322 Frankfurt am Main

Erik Körner leitet die Aktivitäten der AG FCS im NFDI Konsortium Text+. Er ist auch Entwickler und Maintainer für die FCS in CLARIN-ERIC.
Kontakt: koerner@saw-leipzig.de, Sächsische Akademie der Wissenschaften zu Leipzig, Karl-Tauchnitz-Straße 1, 04107 Leipzig

Peter Leinen ist Leiter des Fachbereichs Informationsinfrastruktur der Deutschen Nationalbibliothek, vertritt diese in Text+ sowie der NFDI. Besonderer Fokus seiner Arbeit liegt in diesem Zusammenhang auf der Bereitstellung urheberrechtlich geschützter Daten.
Kontakt: p.leinen@dnb.de, Deutsche Nationalbibliothek, Informationsinfrastruktur, Adickesallee 1, 60322 Frankfurt am Main

Lukas Weimer arbeitet an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen im Koordinationsteam des NFDI-Konsortiums Text+ sowie im Office des Verbunds Base4NFDI. Von Haus aus Literaturwissenschaftler ist sein besonderes Interesse, Forschenden den

Nutzen von Forschungsinfrastrukturen zu vermitteln und diese bedarfsorientiert zu gestalten.
Kontakt: weimer@sub.uni-goettingen.de, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14, 37073 Göttingen

Andreas Witt leitet im Leibniz-Institut für Deutsche Sprache die Abteilung Digitale Sprachwissenschaft. Diese Position ist verbunden mit einer Professur für Computational Humanities an der Universität Mannheim. Andreas Witt ist der National Coordinator Germany im CLARIN ERIC und Sprecher des NFDI-Konsortiums Text+.
Kontakt: witt@ids-mannheim.de, Leibniz-Institut für Deutsche Sprache, R5, 6-13, 68161 Mannheim

Bibliographie

Bertino, Andrea. 2020. „Call for User Stories.“ In *Text+*. *DHd-Blog*. https://dhd-blog.org/?p=14043. (zugegriffen: 24. Juli 2024)

Calvo Tello, José, Nanette Rißler-Pipka, Raisa Barthauer, Stefan Buddenbohm, Sonja Friedrichs und Lukas Weimer. 2021. „Text+ User Stories Data.“ In *DARIAH-DE Repository*. http://dx.doi.org/10.20375/0000-000E-67ED-4.

Nationale Forschungsinfrastruktur (NFDI). 2024. „Die Nationale Forschungsdateninfrastruktur.“ https://www.nfdi.de/verein/. (zugegriffen: 24. Juli 2024)

Rißler-Pipka, Nanette, Raisa Barthauer, Stefan Buddenbohm, José Calvo Tello, Sonja Friedrichs, Lukas Weimer. 2021. „Community Involvement in Research Infrastructures: The User Story Call for Text+ (1.0.0).“ Zenodo. https://doi.org/10.5281/zenodo.5384085.

Text+. 2024. „Text+.“ https://text-plus.org/. (zugegriffen: 24. Juli 2024)