Originalveröffentlichung in: Sprachreport Jg. 38 (2022) Nr. 1, S. 1-7.
DOI: https://doi.org/10.14618/sr-1-2022_trip

Thorsten Trippel

MIT *TEXT+* FORSCHUNGSDATEN DIGITAL VERNETZEN – EIN FALL FÜR DIE SPRACHWISSENSCHAFT?

Der Autor ist wissenschaftlicher Mitarbeiter in der Abteilung „Digitale Sprachwissenschaft“ am Leibniz-Institut für Deutsche Sprache, Mannheim.

Abhängig davon, mit wem man spricht, erhält man unterschiedliche Einschätzungen dazu, was denn Forschungsdaten sind. Dies reicht von Messwerten für die Klimaforschung bis zu Positionsbestimmungen von Objekten in der Astronomie, von Umfragedaten in der Demoskopie bis zu Gensequenzen in der Biotechnologie. Konsens besteht meistens darin, dass diese Forschungsdaten digital vorliegen sollten, am besten noch für alle Forschenden frei zugänglich und leicht auffindbar. Daneben sollten sie für neue Forschungsfragen in den Disziplinen, aus denen sie kommen, einfach nachzunutzen sein. Aber auf die Frage nach Forschungsdaten in der Sprachwissenschaft und anderen geisteswissenschaftlichen Disziplinen sowie deren Digitalisierung und Vernetzung stößt man oft zunächst auf weitere Fragen, zum Beispiel, ob man den anderen Forschungsgebieten folgen will, wie durch Forschungsdaten ein Mehrwert für die Geisteswissenschaften erreicht wird, und vor allem auf die Frage, was Forschungsdaten in der Sprachwissenschaft überhaupt sind. Dabei gibt es gerade bezogen auf Sprachdaten eine lange Tradition der Weitergabe und Nachnutzung von Forschungsdaten. Diese Tradition umfasst auch die Vernetzung sowie Nutzung digitaler Werkzeuge zur Erschließung und Verknüpfung der Forschungsdaten. Derartige Verfahren sind in der Sprachwissenschaft so normal, dass es sogar Nutzenden, die Daten verwenden, meist nicht einmal auffällt, welche Infrastrukturen daran beteiligt sind. Im Rahmen der Nationalen Forschungsdateninfrastruktur, kurz oft NFDI genannt, wurde zum 1. Oktober 2021 mit dem Verbundprojekt Text+ (gesprochen /tɛkstˈplʊs/) eine Vernetzung bestehender Angebote aus verschiedenen geistes-, kultur- und sozialwissenschaftlichen Disziplinen eingerichtet, die besonders für die Sprachwissenschaft neue Möglichkeiten für die Erforschung neuer Forschungsfragen und eine Verknüpfung bestehender Erkenntnisse eröffnen wird. Die Angebote beziehen sich auf Forschungsdaten und methodische Werkzeuge, die für die Forschung mit diesen Daten benötigt werden. Das Leibniz-Institut für Deutsche Sprache ist an diesem Verbund zentral als antragstellende Einrichtung beteiligt.

Forschungsdaten in der Sprachwissenschaft – Grundlage der empirischen Forschung

Forschungsdaten bilden die Grundlage empirischer Forschung. Aufgrund von Daten, die vorliegen oder zielbezogen gesammelt werden, können Hypothesen gebildet und überprüft werden; daneben werden Aussagen mit Beispielen illustriert und Theorien erläutert.

In der Sprachwissenschaft werden sehr unterschiedliche Typen von Forschungsdaten verwendet. Korpora sind ein Beispiel einer großen Klasse von Forschungsdaten. In Sammlungen von Texten werden Wortverwendungen oder grammatikalische Strukturen untersucht, Diskursstrukturen betrachtet und die Auflösung von Koreferenzen erforscht, in Sprachaufnahmen regionale Aussprachevarianten, prosodische Muster oder das Lautinventar analysiert. Beispielsätze aus Korpora können auch in Wörterbüchern und anderen lexikalischen Ressourcen genutzt werden, um die Verwendung von Lemmata zu exemplifizieren. Auch lexikalische Ressourcen, die zur semantischen Erschließung von Texten, zur Erforschung von semantischen Klassen und als Grundlage in der Sprachdidaktik verwendet werden, stellen eine Klasse von Forschungsdaten in der Sprachwissenschaft dar. Informationen aus diesen Ressourcen werden eingesetzt, um Testmaterial für Experimente zu erstellen, z. B. für Lesezeitexperimente in der Psycholinguistik. Die Ergebnisse solcher Experimente sind hierbei Forschungsdaten, die nicht unmittelbar nach Sprache aussehen.

Die genannten Daten sind nur einige schlaglichtartige Beispiele. Im Bereich der Sprachwissenschaft, die nicht ausschließlich auf Introspektion setzt, sind Forschungsdaten allgegenwärtig und werden im Rahmen der üblichen Forschungsmethoden eingesetzt. Der allgegenwärtige Einsatz

und die Differenzierung der Datentypen als Korpus, lexikalische Ressource, Experimente, etc. führen sogar dazu, dass Forschenden nicht immer bewusst ist, dass die eingesetzten Ressourcen an anderer Stelle als Forschungsdaten bezeichnet werden. Sie bilden jedoch eine Grundlage der Forschung und damit auch eine Grundlage von Forschungsergebnissen, die das Wissen in der Sprachwissenschaft erweitert.

Sicherung guter wissenschaftlicher Praxis – Vorgabe der Forschungsförderung

Wissenschaftliche Ergebnisse beruhen darauf, dass zugrundeliegende Daten mittels einer bestimmten Methode analysiert und die daraus gewonnenen Schlüsse im Zusammenhang mit Vorarbeiten dargestellt werden. Ein Anspruch an wissenschaftliche Qualität besteht in der Reproduzierbarkeit der Ergebnisse, d. h. dass der Einsatz vergleichbarer Methoden auf vergleichbaren Ausgangsdaten zu vergleichbaren Schlüssen führt bzw. die gleichen Ausgangsdaten, die mit den gleichen Methoden bearbeitet werden, zu den gleichen Ergebnissen führen. Gutachtende von Publikationen folgen der Argumentation von Forschenden, um zu erkennen, ob diese schlüssig ist und gleichzeitig zu den Daten und Methoden passt. In einigen Bereichen – nicht zuletzt auch in der akademischen Ausbildung – werden Untersuchungen reproduziert, indem z. B. Korpusabfragen aus Publikationen wiederholt und die publizierten Ergebnisse für Lernende als Gold-Standard verwendet werden. Die Publikation der Ausgangsdaten soll dabei nicht nur die Reproduzierbarkeit gewährleisten, sondern auch dazu beitragen, dass andere Forschungsfragen mit den gleichen Ausgangsdaten bearbeitet werden können. Im Bereich der Korpora ist diese Nachnutzung offensichtlich, aber auch bei anderen Klassen sprachwissenschaftlicher Forschungsdaten ist die Nutzung in neuen Kontexten üblich.

Forschungsförderungsorganisationen wie die Deutsche Forschungsgemeinschaft (DFG) haben sich damit beschäftigt, wie die Qualität der von ihnen finanzierten Forschung gesichert werden kann. In den *Leitlinien zur Sicherung guter wissenschaftlicher Praxis* der DFG werden grundlegende Anforderungen zur Qualitätssicherung an die wissenschaftliche Praxis beschrieben, zu denen auch gehört, dass neben der Darstellung der durchgeführten Forschungsschritte auch die Ausgangsdaten archiviert werden.[1]

Nachnutzung von Forschungsdaten durch FAIRe Daten und Verfahren

Ein Schlüssel zur Reproduzierbarkeit und Nachnutzung wird in der Verfügbarmachung von Forschungsdaten gesehen, die auf den sogenannten FAIR-Prinzipien beruhen (siehe Wilkinson et al. 2016). FAIR ist dabei ein Akronym der englischen Wörter Findable (auffindbar), Accessible (zugänglich), Interoperable (interoperabel) und Re-usable (nachnutzbar). Unter *Auffindbarkeit* wird dabei verstanden, dass es Möglichkeiten geben muss, die Daten zu finden. Das Auffinden von Daten erfolgt durch aussagekräftige Beschreibungen (sogenannte Metadaten) und Verzeichnisse für Forschungsdaten oder durch ein eindeutiges Identifikationsmerkmal des Datensatzes, einem persistenten Identifikator. Für den *Zugang* zu den Daten soll ein Verfahren festgelegt sein, das angibt, wie man zu den Daten gelangen kann, wenn man den Identifikator eines Datensatzes kennt. Die Daten werden *interoperabel*, wenn sie in einem Format vorliegen, das so beschrieben ist, dass eine Nachnutzung ermöglicht wird. Neben entsprechenden Datenformaten gehört zur *Nachnutzbarkeit* auch, dass die Rechte zur Nutzung bekannt und dokumentiert sind, wozu auch die Provenienz der Daten gehört. Auf dieser Grundlage wird es möglich, Ausgangsdaten mit geeigneten anderen Methoden zu bearbeiten.

FAIRe Forschungsdaten und Verfahren in Text+

Die Grundlagen zur Sicherung der guten wissenschaftlichen Praxis sind unabhängig vom eingesetzten Medium. Kataloge mit Beschreibungen von Forschungsdaten, z. B. Bibliothekskataloge, können in Papierform vom Grundsatz her genauso verwendet werden wie Wörterbücher als Forschungsdaten. Mit der allgemeinen Verfügbarkeit digitaler Formate ist die Umsetzung der FAIR-Prinzipien für Forschungsdaten allerdings einfacher geworden. Auch die Bereitstellung und Dokumentation von Verfahren oder Methoden wird einfacher, wenn sie in Form von Software festgelegt und beschrieben werden, die als Werkzeug für die Bearbeitung der Daten eingesetzt wird. Für Forschende aus dem sprachwissenschaftlichen Bereich und anderen geisteswissenschaftlichen Disziplinen, die mit Sprach- und Textdaten arbeiten, entwickeln die Partner von Text+ auf Grundlage einer gemeinsamen Forschungsdatenmanagementstrategie eine Infrastruktur, die Forschende beim FAIRen bereitstellen von Forschungsdaten und Werkzeugen unterstützt. In Text+ entsteht durch die unterschiedlichen Disziplinen und disziplinären Traditionen sowie die gewachsenen lokalen Struk-

turen bei den Beteiligten die Infrastruktur, verschiedenste Sprach- und Textdaten sowie Werkzeuge zusammen zu betrachten und zu verwalten.

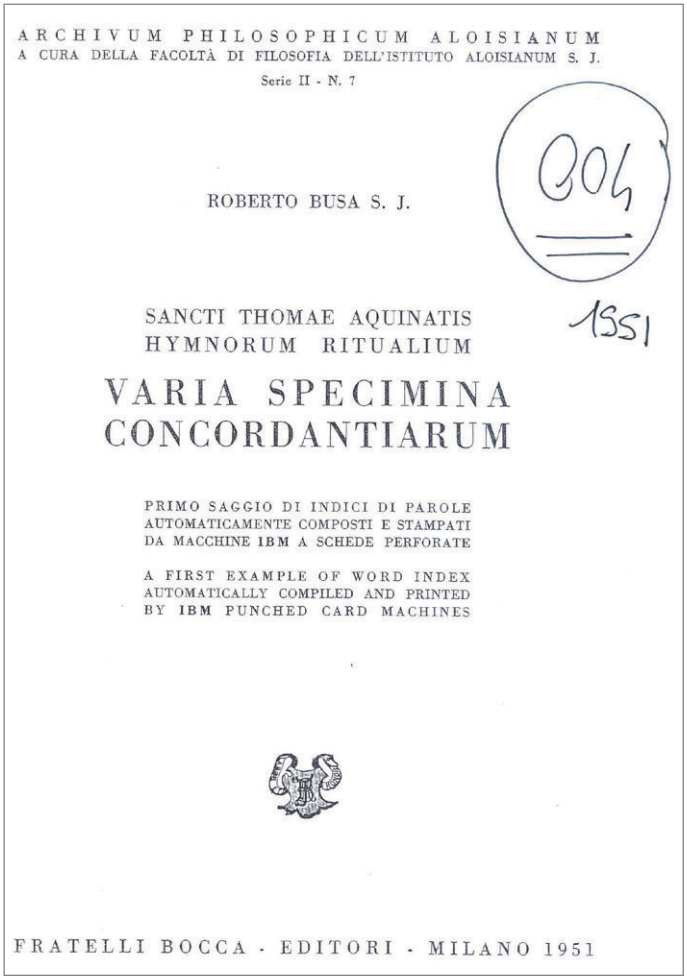
Digitale Sprach- und Textdaten

In der Sprachwissenschaft und weiteren geisteswissenschaftlichen Disziplinen werden Sprach- und Textdaten als Grundlage empirischer Forschung verwendet. Die Katalogisierung der Sprach- und Textdaten erfolgt nicht mehr auf Karteikarten, sondern in digitaler Form. Auch die Daten selbst werden digital verarbeitet, was die Weiterverarbeitung durch Software ermöglicht. Die Digitalisierung von Sprach- und Textdaten hat dabei eine lange Tradition und lässt sich auf einen sehr frühen Einsatz von Computern bereits in den 1950er Jahren zurückführen, Text+ kann also auf eine lange Tradition zurückblicken.

Am Anfang war Thomas von Aquin – und der Jesuit Roberto Busa

Das Gesamtwerk von Thomas von Aquin ist aus philosophischer, theologischer und sprachlicher Perspektive für verschiedenste Forschende von Interesse. Der italienische Jesuit Roberto Busa beschäftigte sich bereits in den 1940er Jahren mit diesem Gesamtwerk und plante die Erstellung einer Konkordanz. Vor dem Hintergrund der ersten Computer veröffentlichte er 1951 erste Ergebnisse seiner Arbeiten zur Digitalisierung der Schriften von Thomas von Aquin (siehe Busa 1951, siehe auch Abb. 1). Ursprünglich auf Lochkarten wurde das Gesamtwerk von Thomas von Aquin später digitalisiert und durchsuchbar gemacht. In diesem Zusammenhang wurden viele Grundlagen der heutigen Computerlinguistik und der Digital Humanities gelegt.[2] Auch für das Deutsche gibt es Sammlungen historischer Texte, etwa das Deutsche Textarchiv,[3] das an der Berlin Brandenburgischen Akademie beheimatet ist.

Neben diesen historischen, digitalisierten Texten ist auch die Übersetzungswissenschaft auf digitale Daten angewiesen. So erfordert die maschinelle Übersetzung die Digitalisierung von Daten, bevor Computer sie übersetzen können. Neben automatischen Übersetzungssystemen gibt es aber auch andere Werkzeuge zur Übersetzung, die auf digitalen Daten aufsetzen. Moderne, in der Übersetzungsindustrie eingesetzte Werkzeuge wie Translation Memory-Systeme sind z. B. digitale Nachfolger von Übersetzungsdateien, mit denen Übersetzende wiederkehrende Phrasen und Sätze


ARCHIVUM PHILOSOPHICUM ALOISIANUM
A CURA DELLA FACOLTÀ DI FILOSOFIA DELL'ISTITUTO ALOISIANUM S. J.
Serie II - N. 7

ROBERTO BUSA S. J.

SANCTI THOMAE AQUINATIS
HYMNORUM RITUALIUM
VARIA SPECIMINA
CONCORDANTIARUM

PRIMO SAGGIO DI INDICI DI PAROLE
AUTOMATICAMENTE COMPOSTI E STAMPATI
DA MACCHINE IBM A SCHEDE PERFORATE

A FIRST EXAMPLE OF WORD INDEX
AUTOMATICALLY COMPILED AND PRINTED
BY IBM PUNCHED CARD MACHINES

FRATELLI BOCCA · EDITORI · MILANO 1951

Abb. 1: Titelblatt der wahrscheinlich ersten Veröffentlichung zu digitalen geisteswissenschaftlichen Forschungsdaten von Roberto Busa (1951); die Publikation war zweisprachig, Englisch und Italienisch, und bezog sich auf lateinische Texte. Zum Einsatz kam ein Rechner von IBM, der mit Lochkarten arbeitete. Das Faksimile wurde zur Verfügung gestellt unter der Creative Commons CC-BY-NC Lizenz durch das CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Mailand, Italien. Das Original befindet sich im „Busa Archive" der Università Cattolica del Sacro Cuore, Mailand.

einheitlich übersetzt haben. Terminologieverwaltungssysteme dagegen erlauben ein schnelles Nachschlagen von Begriffen zur Vereinheitlichung einer Übersetzung und stellen eine besondere Klasse lexikalischer Ressourcen dar.

Die Suche nach Wörtern und Kontexten in Konkordanzen und der Verweis auf die ursprünglichen Texte ist eine Voraussetzung für neue Ansätze zur Textanalyse wie dem von Moretti (2000) beschriebenem *Distant Reading*, das in verschiedenen geisteswissenschaftlichen Disziplinen verwendet wird. Die Korpuslinguistik basiert auch auf der Verarbeitung von digitalisierter Sprache, um Wörter und Strukturen einer Sprache zu analysieren, einschließlich sprachlicher Entwicklungen und regionaler oder stilistischer Unterschiede. Anders als zu Zeiten von Roberto Busa ist es bei neu entstehenden Textdaten in der Regel nicht mehr notwendig, sie zu digitalisieren, da sie bereits am Computer entstehen und elektronisch bereitgestellt werden.

Abb. 2: Ineinandergreifende Datendomänen in Text+

Von Anfang an digital: neue Arten sprachwissenschaftlicher Daten

Sprach- und textuelle Daten sind in digitaler Form bei vielen universitären und außeruniversitären Forschungseinrichtungen, Bibliotheken und Archiven vorhanden. Sammlungen und Korpora wie das Deutsche Referenzkorpus[4] des Leibniz-Instituts für Deutsche Sprache werden kontinuierlich ausgebaut, indem neu verfügbare digitale Daten ergänzt werden und die Verfügbarkeit auch der aktuellen Rechtslage angepasst wird (siehe z. B. Kupietz et al. 2018). Dadurch, dass neue Daten bereits digital bereitgestellt werden, z. B. über entsprechende Lizenzen mit Verlagen, kann die Sammlung weiter wachsen und ergänzt werden, ohne einen Digitalisierungsschritt vornehmen zu müssen.

Die Digitalisierung ermöglicht auch Forschungsdaten, die in gedruckter Form nicht leicht darzustellen wären. Ein Beispiel dafür ist das GermaNet (siehe Hamp / Feldweg 1997; Henrich / Hinrichs 2010), ein lexikalisch-semantisches Wortnetz für das Deutsche, das nicht wie ein traditionelles Wörterbuch organisiert ist, sondern ein Netzwerk aus Bedeutungen, sogenannten Synsets, aufspannt und deren semantische Beziehungen modelliert. Ohne zugehörige Werkzeuge wie dem GermaNet Rover[5] können die semantischen Beziehungen nicht umfassend dargestellt werden. Forschungsdaten, die von Anfang an digital vorliegen und wie Hypertexte nicht ausschließlich für die gedruckte Darstellung erstellt wurden, ermöglichen neue Forschungsfragen und Verknüpfungen. Wenn beispielsweise das GermaNet mit digitalen Texten verknüpft wird, ergeben sich neue Möglichkeiten der Erschließung von Texten, und auch die Texte selbst können miteinander verknüpft werden.

Text+ baut daher eine auf Sprach- und Textdaten ausgerichtete Forschungsdateninfrastruktur auf, die sich zunächst auf digitale Sammlungen, lexikalische Ressourcen und Editionen konzentriert. Diese Datendomänen haben eine lange Tradition in der geisteswissenschaftlichen Forschung. Sie sind mit ausgereiften methodologischen Paradigmen verknüpft, die jeweils charakteristische, aber auch bereichsübergreifende Praktiken der Datenerzeugung, -nutzung, -analyse, -vernetzung und -kuratierung erfordern. Dadurch wird eine breite Palette von Fachdisziplinen adressiert, z. B. die Linguistik, Literaturwissenschaft, Philologien auch der sogenannten „Kleinen Fächer“, Philosophie sowie sprach- und textbasierte Forschung in den Sozialwissenschaften und der Politikwissenschaft. Die Datendomänen werden zusammen und mit ihren Querverbindungen betrachtet (siehe Abb. 2).

Das Plus in Text+

Forschungsdaten in der Sprachwissenschaft und in angrenzenden geisteswissenschaftlichen Disziplinen beschränken sich nicht auf Texte, wie sie von Verlagen in Büchern und Artikeln veröffentlicht werden. Auch abseits von Texten gibt es einschlägige Forschungsdaten, wie Aufnahmen gesprochener oder gebärdeter Sprache. Auch Experimente, z. B. in der Phonetik mit Messungen zur Lauterzeugung im Vokaltrakt oder psycholinguistische Experimente zum Sprachverstehen sind für Forschende in der Sprachwissenschaft relevant. Der Name Text+ soll vermitteln, dass sich diese Initiative auf sprach- und textbasierte digitale Forschungsdaten konzentriert. Die Daten sind aber gleichzeitig heterogen: Sie beziehen sich auf unterschiedliche Sprachräume (auch über Europa hinaus) und Modalitäten von Sprache und Schriftsystemen. Das +-Zeichen weist auf die Offenheit für diese verschiedenen Datentypen hin und darauf, dass neben den Daten auch weitere Angebote und Werkzeuge Teil der Initiative sind. Die Aufteilung gemäß unterschiedlicher Datentypen erlaubt dabei eine Strukturierung der Arbeiten.

Sprach- und textbasierte **Sammlungen** enthalten Zusammenstellungen geschriebener, gesprochener oder gebärdeter Sprache und Texte. Außerdem sind sprach- und textbezogene Experimental- oder Messdaten adressiert, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden. Beispiele für entsprechende Daten sind dabei Textsammlungen, mono- und multimodale Aufnahmen beispielsweise von spontaner und formaler Sprache, Sensordaten (z. B. EEG, Eyetracking, Artikulographie), Befragungen, Reaktionszeitexperimente etc.

Unter **lexikalischen Ressourcen** werden Daten verstanden, die die Verwendung von Wörtern in Sätzen, Texten und multimodaler Kommunikation beschreiben, darunter: Wörterbücher, Enzyklopädien, Normdaten, terminologische Datenbanken, Ontologien, Wortlisten, Wortkarten und linguistische Atlanten, Übersetzungswörterbücher (für menschliche oder maschinelle Übersetzung) etc.

Kritische Repräsentationen historischer Dokumente werden als **Editionen** bezeichnet. Editionen sind durch die zuverlässige methodengeleitete Bewahrung, Präsentation und Kommentierung aller Arten von Texten in verschiedenen Sprachen und Schriftsystemen charakterisiert. Zu den editorischen Modellen gehören dokumentarische oder diplomatische Editionen, Editionen zur Entstehungsgeschichte von Dokumenten und historisch-kritische Editionen.

Werkzeuge zur Erstellung, Bearbeitung, Analyse und Archivierung in allen Datendomänen sind ebenso Teil von Text+. Um mit diesen Programmen kollaborativ arbeiten zu können, müssen einige technische Voraussetzungen erfüllt sein: Es muss technisch gelöst sein, was passiert, wenn verschiedene Nutzende auch gleichzeitig mit den gleichen Werkzeugen arbeiten, wie die Werkzeuge zusammenarbeiten und die Nutzenden einen einfachen Zugang zu den Werkzeugen finden können. Um dies zu ermöglichen, sind verschiedene Komponenten der Werkzeuge aufeinander abgestimmt, teilweise werden auch gleiche Komponenten verwendet. Daher werden diese Bestandteile in einem übergreifenden Bereich für den Betrieb und die gemeinsame Infrastruktur zusammen weiterentwickelt.

Verknüpfung von Forschungsdaten, Werkzeugen und Forschenden

Forschungsinfrastrukturen dienen der FAIRen Bereitstellung und Verbreitung von Forschungsdaten, damit sie – auch außerhalb von Ausdrucken in Anhängen von Forschungsarbeiten – zur Reproduktion von Ergebnissen genutzt werden können. Da Forschungsdaten in der Sprachwissenschaft und verwandten Disziplinen häufig sehr spezifisch sind – sowohl in Bezug auf die Datenformate als auch die Inhalte –, erfordert die Nachnutzbarkeit auch die Bereitstellung von Software, die mit diesen Daten umgehen kann. Die Partner von Text+ haben bereits eine Vielzahl von Werkzeugen entwickelt und bereitgestellt, die in Text+ nun zusammengeführt werden können. Viele dieser Werkzeuge stehen als Anwendungen im World Wide Web zur Verfügung, einige als Webservices, also als Programme, die von anderen Programmen über ähnliche Technologien wie Webseiten aufgerufen werden können. Auf diese Weise können Werkzeuge, die bei einem Partner installiert sind, auch von externen Nutzenden verwendet werden. Diese ortsverteilte oder föderierte Architektur ist sowohl für Werkzeuge als auch für Daten innerhalb von Text+ zentral.

Ortsverteilte Datenhaltung und Werkzeuge

Viele Forschungsdaten in den Geisteswissenschaften berühren Rechte Dritter: Sie wurden über Jahre und Jahrzehnte von Forschenden zusammengetragen, einzelne Institutionen haben von Verlagen Lizenzen für Daten erworben. Darüber hinaus können personenbezogene Daten enthalten sein, d. h. die Daten enthalten Informationen, die Rückschlüsse auf die involvierten Personen zulassen, z. B. Sprachaufnahmen mit nicht verfremdeten Stimmen, Texte mit Bezug zu Personen oder Situationen oder Informationen, die im Rahmen von Experimenten erfasst wurden.

Diese Daten können häufig nicht an andere Institutionen weitergereicht werden. Ein Grund dafür kann darin bestehen, dass diejenigen, deren Rechte betroffen sind, einer Weitergabe an andere nicht zugestimmt haben, z. B. weil zum Zeitpunkt der Erhebung die Frage der Weitergabe sich nicht gestellt hat oder es überhaupt nicht absehbar war, dass die Daten auch über mehrere Jahre in Forschungsprozessen verwendbar sind. Auch die Software für die Verarbeitung von Daten kann nicht immer an andere Orte verschoben werden, wenn z. B. die Installation aufwendig ist, die Software kontinuierlich von Forschenden weiterentwickelt wird, die Lizenzen es nicht zulassen oder die technischen Anforderungen spezielle Umgebungen erfordern.

In Text+ arbeiten 30 Institutionen zusammen, um ein Netzwerk von Datenhaltungseinrichtungen – Repositorien – zu bilden und Schnittstellen zu schaffen, durch die auf die Daten mit gemeinsamen Verfahren zugegriffen werden kann. Die Beteiligten stammen aus den Kreisen der Hochschulen, wissenschaftlichen Bibliotheken, Datenzentren der Digital Humanities, außeruniversitären Forschungseinrichtungen der Max-Planck-Gesellschaft und der Leibniz-Gemeinschaft oder sie sind Mitglieder der Union der deutschen Akademien der Wissenschaften. Text+ umfasst außerdem führende Rechenzentren, die robuste und persistente Infrastrukturdienste für eine distribuierte Forschungsdateninfrastruktur bis hin zur Langzeitarchivierung absichern. Durch vielfältige Kooperationen, zum Beispiel mit Europäischen Forschungsdateninfrastruktur Konsortien (ERICs), ist Text+ auch international vernetzt und angebunden.

Die Forschungsdaten an den beteiligten Institutionen werden seit Jahren gesammelt und stehen für die sprachwissenschaftliche Forschung und andere Disziplinen zur Verfü-

gung. Daher wurden bei vielen der Partner auch spezialisierte Werkzeuge entwickelt – in Abhängigkeit auch von der jeweiligen Spezialisierung auf bestimmte Typen von Daten. Forschende, die nach Daten und Werkzeugen suchen, stehen bei ortsverteilten Angeboten vor der schwierigen Frage, wo sie welche Daten und Werkzeuge finden. Aus diesem Grund ist es in einer verteilten Infrastruktur wie Text+ essenziell, Möglichkeiten zu bieten, an einer Stelle zu erfahren, welche Daten und Werkzeuge zur Verfügung stehen und wie man darauf Zugriff erlangen kann. Solche Nachweissysteme dienen als Katalog für Daten und Werkzeuge. Text+ wird daher die Möglichkeit schaffen, die Angebote der verschiedenen Partner über gemeinsame Suchfunktionen zu finden und, wo möglich, auf die Inhalte zuzugreifen. An den Stellen, wo ein Zugriff nur für Forschende akademischer Institutionen möglich ist, gibt es vielfach bereits die Möglichkeit, sich über die Zugänge an Heimatinstitutionen für diese Angebote einzuloggen und sie dort – unabhängig vom Aufenthaltsort – zu nutzen. Nutzende, die Hilfestellungen bei der Verwendung von Daten und Werkzeugen benötigen, können zusätzlich Kontakt zu anderen Forschenden erhalten. Text+ vernetzt damit nicht nur Daten und Werkzeuge, sondern bietet auch eine Grundlage für die Zusammenarbeit von Forschenden.

Beratung und Vernetzung der Forschenden

Komplexe Software und Daten sind vielfach nicht selbsterklärend. Forschende, die noch keine Erfahrung mit diesen Werkzeugen und Daten haben, erhalten in Text+ Beratung und Unterstützung. Ein niederschwelliges Angebot wird durch die Einrichtung eines Helpdesks geschaffen. Über ein Webformular können dabei Anfragen zu den Angeboten von Text+ gestellt werden, die von Fachleuten aus dem Verbund beantwortet werden.

Neben der Beantwortung von konkreten Fragen zur Verwendung von Angeboten von Text+ werden auch weiterreichende Schulungen und Kurse erstellt. Einige Materialien sind für die Nachnutzung im Rahmen der akademischen Lehre auch an weiteren Einrichtungen vorgesehen, aber Kurse und Angebote richten sich insbesondere auch an fortgeschrittene Studierende und Promovierende, die so Grundlagen des Forschungsdatenmanagements, der Arbeit mit Daten und dem Umgang mit den entsprechenden Softwarewerkzeugen erlernen können.

Um die Daten und Dienste möglichst allgemein verwendbar zu machen, sind die Partner von Text+ auch im Bereich der Standardisierung engagiert. Von Normen für Sprachressourcen im Deutschen Institut für Normung (DIN) bis zu fachspezifischen Institutionen wie der Text Encoding Initiative sind die Partner von Text+ daran beteiligt, durch die Standardisierung von Schnittstellen, Verfahren und Formaten eine möglichst große Interoperabilität von individuellen Forschungsfragen und -methoden zu unterstützen.

Weitere Entwicklungen bei der digitalen Vernetzung von Forschungsdaten – nicht nur für die Sprachwissenschaft

Die Interoperabilität von Forschungsdaten, Forschungssoftware und Forschenden beschränkt sich nicht nur auf die Sprachwissenschaft. Die verwendeten Methoden reichen von traditionellen geisteswissenschaftlichen Fragestellungen bis zu aktuellen Fragen der Forschung im Bereich des maschinellen Lernens. Daher kooperiert Text+ im Rahmen der NFDI mit anderen Disziplinen. Die NFDI umfasst ein breites Portfolio von der Genetik, Physik, den Material- und Sozialwissenschaften bis zu weiteren geisteswissenschaftlichen Disziplinen.

Nachweissysteme, Zitationsverfahren für Forschungsdaten und Datenhaltungssysteme, wie sie in Text+ verwendet werden, sind zu großen Teilen unabhängig von konkreten Disziplinen, weshalb die Vernetzung auch mit anderen Konsortien einen Austausch von Erfahrungen und Entwicklungen ermöglicht. Gleichzeitig werden die sprachwissenschaftlichen Angebote zum Beispiel für die Erschließung von Texten für andere Disziplinen nutzbar. Entscheidend sind aber auch konkrete geisteswissenschaftliche Forschungsfragen: Text+ hat über 100 Einsatzszenarien zusammengestellt,[6] in denen Forschende die Infrastruktur im Rahmen konkreter Forschungsaufgaben nutzen möchten. Damit vernetzt Text+ Forschungsdaten digital nicht nur für die Sprachwissenschaft, bietet aber gerade auch für die Sprachwissenschaft die notwendigen Daten und Werkzeuge für neue und zusätzliche Forschungsfragen.

Um neue Forschungsfragen innerhalb einer Infrastruktur mit erfassen zu können, enthält Text+ verschiedene dynamische Aspekte und ist von vornherein wissenschaftsgeleitet aufgebaut. Dazu kann der Arbeitsplan durch Mitbestim-

mungsgremien aus Außenstehenden mit beeinflusst werden. Dies geschieht in wissenschaftlichen Koordinationskomitees für alle Datenbereiche, als Scientific Coordination Committees bezeichnet, in die Fachverbände und -verbünde Forschende entsenden können. In diesen Komitees werden die Arbeiten des Verbundes priorisiert und evaluiert, um auf neue Forschungsfragen, Entwicklungen und Anforderungen aus der Forschung reagieren zu können.

Forschungsdaten haben in der Sprachwissenschaft eine lange Tradition und werden seit langem dynamisch weiterentwickelt. Die Beteiligung der Sprachwissenschaft an der geistes-, kultur- und sozialwissenschaftlichen Forschungsinfrastruktur Text+ betont die Bedeutung des komplexen Feldes der Forschungsdaten auch für die Sprachwissenschaft.

Weitere Informationen zu Text+ sind unter <www.text-plus.org> erhältlich. ■

Literatur

Busa, Roberto (1951): Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate. (= Archivum philosophicum Aloisianum 2, 7). Milano: Fratelli.

Deutsche Forschungsgemeinschaft (DFG) (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis. <www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf> (Stand: 9.12.2021).

Hamp, Birgit / Feldweg, Helmut (1997): GermaNet – a lexical-semantic net for German. Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid.

Henrich, Verena / Hinrichs, Erhard (2010): GernEdiT – the Germanet editing tool. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Malta, S. 2228-2235.

Kupietz, Marc / Lüngen, Harald / Kamocki, Paweł / Witt, Andreas (2018): The German reference corpus DeReKo: new developments – new opportunities. In: Calzolari, Nicoletta / Choukri, Khalid / Cieri, Christopher et al. (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA), S. 4353-4360. <www.lrec-conf.org/proceedings/lrec2018/pdf/737.pdf> (Stand: 9.12.2021).

Moretti, Franco (2000): Conjectures on world literature. In: New Left Review 1. <https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature.pdf> (Stand: 9.12. 2021).

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand J. et al. (2016): Addendum: The FAIR Guiding Principles for scientific data management and stewardship. In: Sci Data 3, 160018. <https://doi.org/10.1038/sdata.2016.18> (Stand: 9.12.2021).

Winter, Thomas N. (1999): Roberto Busa, S. J., and the Invention of the Machine-Generated Concordance. Faculty Publications. In: Classics and Religious Studies Department 70. University of Nebraska – Lincoln. <https://digitalcommons.unl.edu/classicsfacpub/70> (Stand: 9.12.2021).

Anmerkungen

1 Siehe DFG (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis, Leitlinie 7 „Phasenübergreifende Qualitätssicherung“ (S. 14), Leitlinie 11 „Methoden und Standards“ (S. 17), Leitlinie 12 „Dokumentation“ (S. 17-18), Leitlinie 13 „Herstellung von öffentlichem Zugang zu Forschungsergebnissen“ (S. 18-19) und Leitlinie 17 „Archivierung“ (S. 22).

2 In seinem lesenswerten Aufsatz beschreibt Winter (1999) das Projekt von Roberto Busa in der Retrospektive.

3 Siehe <www.deutschestextarchiv.de/> (Stand: 9.12.2021).

4 Siehe <www.ids-mannheim.de/digspra/kl/projekte/korpora/> (Stand: 9.12.2021).

5 Siehe <https://weblicht.sfs.uni-tuebingen.de/rover/> (Stand: 9.12.2021); die Nutzung erfordert das Login über eine universitäre oder außeruniversitäre Forschungsinstitution.

6 Siehe <www.text-plus.org/forschungsdaten/user-stories/> (Stand: 9.12.2021). ■