DARIAH

Erhard Hinrichs* und Thorsten Trippel

CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften

DOI 10.1515/bfp-2017-0015

Zusammenfassung: Für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften stellt CLARIN eine Forschungsinfrastruktur bereit, die auf die hochgradig heterogenen Forschungsdaten in diesen Wissenschaftsbereichen angepasst ist. Mit Werkzeugen zum Auffinden, zur standardkonformen Aufbereitung und zur nachhaltigen Aufbewahrung von Daten sowie mit der Bereitstellung von virtuellen Forschungsumgebungen zur kollaborativen Erstellung und Auswertung von Forschungsdaten unterstützt CLARIN alle wesentlichen Aspekte des Datenmanagements und der Datenarchivierung. Diese CLARIN-Angebote werden durch Beratungs- und Schulungsmaßnahmen begleitet.

Schlüsselwörter: Forschungsdaten; Forschungsinfrastruktur; Geistes- und Sozialwissenschaften; sprachbasierte Forschung, CLARIN

CLARIN-D: A Research Infrastructure for Language Based Research in the Humanities and Social Sciences

Abstract: CLARIN provides a research infrastructure for language based research in the humanities and social sciences, adapted for the highly heterogeneous data used in the academic disciplines involved. CLARIN offers support for all essential aspects of data management and data archiving. It provides tools for accessing, preparation and depositing of data, and it makes available virtual research environments for the collaborative creation and analysis of research data. These CLARIN services are accompanied by helpdesk support and by dissemination activities.

***Kontaktperson: Prof. Erhard Hinrichs,**
erhard.hinrichs@uni-tuebingen.de
Thorsten Trippel, thorsten.trippel@uni-tuebingen.de

Keywords: Research data; research infrastructure; humanities and social sciences; language based research; CLARIN

Inhalt

1 Motivation

Geistes- und sozialwissenschaftliche Projekte basieren zunehmend auf empirisch erhobenen Daten, insbesondere in dem Bereich, der als *e-Humanities* oder *Digital Humanities* (DH) bezeichnet wird. Die Datenorientierung zeigt sich sowohl in quantitativen Studien, in denen Daten zur Verifizierung oder Falsifizierung von Hypothesen verwendet werden, als auch in qualitativen Studien, in denen Hypothesen durch die Betrachtung von Daten entwickelt und geschärft werden. Presner (2010) weist darauf hin, dass frühe DH-Projekte sich häufig mit der Digitalisierung von Daten und einer technologischen Grundstruktur beschäftigt haben, dass die Weiterentwicklung aber generativ sei und Arbeitsumgebungen bereitstelle, um digitales Wissen

zu erstellen, zu verwalten und damit umzugehen. Aus diesem neuen Umgang resultieren innovative Fragestellungen, die in den Digital Humanities prototypisch zu computergestützten Analyseansätzen führen.[1]

Mit der Verfügbarkeit von großen Beständen digitalisierter Daten sowie automatischen und maschinell unterstützten Analyseverfahren lassen sich wissenschaftliche Fragen und Hypothesen auf breiterer empirischer Basis bearbeiten, eröffnen sich neue methodische Zugänge bzw. entstehen gänzlich neue wissenschaftliche Fragestellungen. Diese Fragen gehen einher mit Diskussionen zum „digital turn“[2] und „empirical turn“[3] und ermöglichen eine Erweiterung des disziplinspezifischen Methodenspektrums. Fachnahe Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften greifen diese Methoden auf und haben die Aufgabe die Forschung in allen Phasen zu unterstützen – bei der Datenrecherche, der digitalen Bereitstellung von Daten, der Verknüpfung von Daten zu virtuellen Kollektionen, der Analyse von Daten mithilfe von interoperablen Softwaretools bis hin zur Speicherung und Archivierung von dabei entstehenden Forschungsdaten.

In den Geistes- und Sozialwissenschaften wird in den meisten Disziplinen sprach- und textbasiert gearbeitet, d. h. die Sprache selbst und die sprachlich markierten Phänomene sind Gegenstand der Forschung oder das primäre Medium, in dem die zu untersuchenden Gegenstände verfasst sind.[4] Sprachbasierte Untersuchungen mit Methoden der Digitalen Geisteswissenschaften schließen Studien zur Sprachevolution, des Sprachwandels und der Sprachvariation mit phylogenetischen und dialektometrischen Methoden in der Linguistik ebenso ein wie stylometrische und datenbasierte textanalytische Methoden in den Literatur-, Sozial- und Politikwissenschaften und inhaltsanalytische Verfahren zur thematischen Zuordnung und zur Entstehungsgeschichte von historischen Dokumenten. Dadurch lassen sich zum Beispiel mit automatischen Analyseverfahren generierte Listen signifikanter politischer Themen und deren Entwicklung in Zeitungen oder politischen Reden, Zusammenhänge zwischen in Texten erwähnten Personen sowie sprachlicher Wandel und Varietäten untersuchen.

Sprach- und textbasierte Forschungsdaten in den Geistes- und Sozialwissenschaften, besitzen die Besonderheit, dass sie häufig in iterativen Prozessen durch Forschende erstellt und fortlaufend qualitativ angereichert und quantitativ erweitert werden. Dazu werden sie zum Teil mit erheblichem Zeitaufwand von Experten händisch erstellt, zum Teil aber auch automatisiert annotiert, d. h. mit Anmerkungen und Analysen versehen. Diese Annotationen und zugrundeliegenden Daten dienen wiederum als Ausgangsdaten für weitere Analysen. Dieser Zyklus kann mehrfach durchlaufen werden, jeweils mit anderen Verfahren und Auswertungen, was die Komplexität der Daten erhöht. Diese Komplexität unterscheidet Forschungsdaten der Geistes- und Sozialwissenschaften von Forschungsdaten, wie sie etwa in den Naturwissenschaften als Messergebnisse anfallen oder von Objektwissenschaften, in denen Gegenstände möglichst detailgetreu abgebildet werden. Außerdem sind Forschungsdaten in den Geisteswissenschaften nicht statisch sondern dynamisch in dem Sinne, dass laufend weitere Daten in den Datensammlungen oder Annotationsebenen hinzukommen und zur weiteren Verarbeitung herangezogen werden. Infrastrukturen, die sich mit Sprachdaten und -werkzeugen befassen, haben diese Anforderungen zusätzlich zu betrachten und gehen damit durch ihre fachspezifische Ausrichtung über die Angebote von Bibliotheken und Rechenzentren hinaus.

2 CLARIN – ein Überblick

CLARIN – ein Akronym für *Common Language Resources and Technology Infrastructure* – ist eine Infrastruktur-Initiative für die Geistes- und Sozialwissenschaften, in denen sprachorientiert gearbeitet wird. Für die Forschenden, die mit Sprachdaten arbeiten, stellt CLARIN digitale Daten und Werkzeuge bereit. Sprachdaten in der Forschung sind neben der Komplexität sehr divers, z. B. Texte, Audio- und Videoaufnahmen, sowie multimodale Daten. In der CLARIN-Infrastruktur stehen spezialisierte Werkzeuge zur Verfügung, um Daten aufzufinden, zu erschließen, auszuwerten und zu annotieren, unabhängig davon, wo die Daten physikalisch vorliegen. Auch können Datensätze aufbewahrt und zusammengefasst werden, um die Datenbasis zu vergrößern oder komplexere Fragestellungen zu betrachten.

Als europäisches Vorhaben folgt CLARIN dem Aufbau und Verlauf der ESFRI-Roadmap, die von dem *European Strategy Forum for Research Infrastructures* erstellt worden ist.[5] Bestandteil dieser Roadmap sind für die Sozialwissen-

1 Vgl. z. B. Schaal und Kath (2014).

2 Siehe z. B. Berry (2011), Baum und Stäcker (2015).

3 Thiel (2012) zum Beispiel diskutierte am 24.07.2012 in der Frankfurter Allgemeinen Zeitung die empirische Wende in den Geisteswissenschaften.

4 Dies trifft natürlich nicht auf alle geistes- und sozialwissenschaftliche Disziplinen in gleicher Weise zu, v. a. nicht auf jene, bei denen Objekte und Artefakte, wie etwa in der Archäologie und der Kunstgeschichte, die primären Forschungsgegenstände ausmachen.

5 Hinrichs und Krauwer (2014).

schaften *CESSDA*, *European Social Survey* und *SHARE*, sowie CLARIN und DARIAH für die Geisteswissenschaften. CLARIN ist in Europa als *European Research Infrastructure Consortium* (ERIC) organisiert, einer besonderen Rechtsform für Forschungsinfrastrukturen unter europäischem Recht. Aktuell im September 2016 hat das CLARIN ERIC 18 Mitgliedsländer mit Bulgarien, Deutschland, Dänemark, Estland, Finnland, Griechenland, Italien, Lettland, Litauen, Niederlande, Norwegen, Österreich, Polen, Portugal, Schweden, Slowenien, Tschechien, Ungarn. Als Beobachter ist Großbritannien und als zwischenstaatliche Institution die Niederländische Sprachunion Mitglied des ERIC.

In Deutschland bestehen gegenwärtig neun CLARIN-Zentren (Stand September 2016), die sowohl an Universitäten als auch an außeruniversitären Forschungseinrichtungen beheimatet sind. Es handelt sich dabei um die CLARIN-Zentren am Institut für Deutsche Sprache (IDS) in Mannheim, an der Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) und am Max-Planck-Institut (MPI) für Psycholinguistik in Nijmegen. Hinzu kommen CLARIN-Zentren an den Universitäten von Hamburg, Leipzig, Stuttgart und Tübingen, sowie der Universität des Saarlandes und der Ludwig-Maximilians-Universität München. Die Koordination des CLARIN-Zentrenverbunds in Deutschland erfolgt durch das CLARIN-Zentrum an der Universität Tübingen.

Die CLARIN-Zentren arbeiten eng mit Facharbeitsgruppen (F-AGs) zusammen, die mit besonderem Augenmerk auf disziplinäre Anforderungen aus den Philologien, der Geschichtswissenschaft, Sozialwissenschaft und Sprachwissenschaft die Entwicklungen von CLARIN begleiten und die Daten und Dienste in ihrem Forschungskontext anwenden. Durch die fortlaufende Einbeziehung der Fachwissenschaften wird sichergestellt, dass die Angebote der CLARIN-Infrastruktur den tatsächlichen disziplinären Bedürfnissen entsprechen. In diesen Facharbeitsgruppen arbeiten gegenwärtig 200 Forschende aus ganz Deutschland mit, die im Bereich der e-Humanities aktiv sind. Zusätzliche technische Expertise und Dienste erhält CLARIN-D bedarfsorientiert durch die Zusammenarbeit mit wissenschaftlichen Großrechenzentren, der GWDG in Göttingen, dem JSC in Jülich und dem MPCDF in Garching.

Um die sprachorientierte Forschung in den Geistes- und Sozialwissenschaften bedarfsgerecht zu unterstützen, stellt CLARIN-D als Infrastruktur gemeinsame Dienste zur Verfügung, um Daten aufzufinden, auszuwerten und aufzubewahren. Dadurch werden auch die Bereiche der Geistes- und Sozialwissenschaften angesprochen, in denen Kompetenzen im Bereich der Technologie, Standards und rechtlichen Fragestellungen unterrepräsentiert sind. Neben den Diensten und Daten stellt CLARIN-D Forschenden Handreichungen von Experten zu technischen und juristischen Fragen bereit und bietet Kurse und Beispielverwendungen zu Werkzeugen und Daten an. Im Folgenden stellen wir dar, welche Angebote CLARIN-D bietet.

3 Auffinden von Forschungsdaten für die Geistes- und Sozialwissenschaften

Forschende, die Daten nachnutzen möchten, benötigen eine Suchfunktion, die vorhandene Daten auffindbar macht, Informationen darüber bereitstellt und beschreibt, unter welchen Bedingungen diese Daten nachgenutzt werden können. In CLARIN gibt es dafür zwei verschiedene Standardzugänge: eine Metadatensuche über strukturiert vorliegende Metadaten und eine föderierte Inhaltsrecherche über den Inhalt der Ressourcen selbst.

Die metadatenbasierte Suche wird in CLARIN mithilfe einer besonderen Suchmaschine durchgeführt, dem *Virtual Language Observatory* (VLO).[6] Vergleichbar mit Literatursuchen in Bibliotheken steht mit dem VLO ein Katalog für spezialisierte Forschungsdaten bereit, mit dessen Hilfe Forschende nach Ressourcen suchen können, die zu ihrem Anwendungszusammenhang passen, also statt nach Literatur in Bibliotheken suchen sie nach nachnutzbaren Daten. Die Suchmaschine wertet dazu die Metadaten aus, sodass eine Suche über Stichworte, Schlüsselbegriffe, Objektsprache, Genre, Lizenz, etc. möglich ist. Dazu verwendet das VLO sowohl eine Indizierung für eine Volltextsuche über die Metadaten, als auch die strukturierte Suche über Facetten, durch die gezielt über geschlossene Vokabulare und vorhandene Wertemengen gesucht werden kann. Darauf basierend können die zur jeweiligen Fragestellung passenden Ressourcen ausgewählt und von den aufbewahrenden Stellen bezogen werden. Im September 2016 weist das VLO über 900 000 Datensätze nach.

Eine Suche über den Inhalt von archivierten Daten ermöglicht eine Funktion, die als *Federated Content Search* (FCS)[7] bezeichnet wird. Die Suche kann auf eine Sprache oder einzelne Ressourcen eingeschränkt werden. Im Gegensatz zu den Metadaten, die vom VLO verwendet werden, sind beim FCS besonders auch die Rechte und Lizenzen zu berücksichtigen, die auf Sprachressourcen in Archiven liegen, z. B. Urheberrechte von Texten, Persönlichkeitsrechte etc. Da Lizenzen in der Regel für das Archiv

6 Vgl. z. B. van Uytvanck et al. (2012).
7 Vgl. Stehouwer et al. (2012).

gelten, das die Daten vorhält, wird kein zentraler Index angelegt, der eine Kopie der Daten enthält. Stattdessen wird die Suche verteilt – also föderiert – bei dem datenhaltenden Zentrum durchgeführt und die Suchergebnisse entsprechend der Lizenzbestimmungen ausgegeben. Sowohl das VLO als auch der FCS sind Kernbestandteile von CLARIN und werden auf europäischer Ebene angeboten und weiterentwickelt.

Ein besonderer Schwerpunkt bei der föderierten Suche liegt auf den Sprachressourcen, die von den CLARIN-Zentren selbst zur Verfügung gestellt werden. Dazu gehören insbesondere umfangreiche Datensammlungen in Form von National-, Akademie- oder Referenzkorpora zu den in CLARIN vertretenen Einzelsprachen. Der CLARIN-D-Zentrenverbund stellt mit dem Deutschen Referenzkorpus (DeReKo)[8] und mit dem Deutschen Text-Archiv (DTA)[9] die beiden größten gegenwärtig verfügbaren Korpora für das Gegenwartsdeutsche bzw. für den gesamten Zeitraum des Neuhochdeutschen zur Verfügung. Das Deutsche Referenzkorpus *DeReKo* am IDS Mannheim enthält mehr als 29 Milliarden Token[10] zur deutschen Gegenwartssprache. Das Deutsche Text Archiv ist ein Referenzkorpus mit Texten von 1600 bis 1900 und erlaubt damit sprachhistorische Querschnittsuntersuchungen auf der Grundlage von 142 351 149 Token aus 2 449 Werken.[11]

Linguistisch annotierte Korpora, sogenannte Baumbanken, liegen in CLARIN mit der Tübinger Baumbank des Deutschen/Zeitungskorpus (TüBa-D/Z)[12] und dem Stuttgarter Tiger-Korpus[13] als weitere Referenzkorpora vor. Neben dem Negra-Korpus[14] stellt das CLARIN-Zentrum an der Universität des Saarlandes auch fremdsprachige Korpora bereit.

Korpora beziehen sich nicht nur auf geschriebene Sprache. Das CLARIN-D-Zentrum am Bayerischen Archiv für Sprachsignale an der LMU München hat sich auf gegenwartsbezogene, gesprochene Sprache spezialisiert[15] und archiviert sowohl Audio- als auch multimodale Daten samt Annotation, die auch im Bereich der Sprachtechnologie eingesetzt werden. Mit dem DOBES-Archiv am MPI für Psycholinguistik[16] gibt es eine Referenzsammlung für den Bereich der Sprachdokumentation, in der sowohl audiovisuelle Daten als auch Texte in vielen verschiedenen, meist vom Aussterben bedrohten Sprachen vorhanden sind. Am CLARIN-Zentrum an der Universität Hamburg stehen zudem mehrsprachige und gebärdensprachliche Daten zur Verfügung.

Mit großen digitalen lexikalischen Ressourcen stellt CLARIN-D einen weiteren Datentypus zur Verfügung, der für die sprachbasierte Forschung von zentraler Bedeutung ist. Das Digitale Wörterbuch der Deutschen Sprache (DWDS)[17] wird am CLARIN-D-Zentrum an der BBAW in Berlin gepflegt und basiert auf umfangreichen Korpussammlungen, um die Verwendung von Wörtern zu repräsentieren. Beim Online-Thesaurus *GermaNet*[18] handelt es sich um das Referenzwortnetz für das Deutsche, in dem Wortbedeutungen als Netz lexikalischer Relationen abgebildet werden und das für Suchfunktion im DWDS eingesetzt wird. Das Leipziger CLARIN-D-Zentrum stellt mit dem Wortschatzprojekt eine weitere vielgenutzte Lexikonressource für das Deutsche zur Verfügung, die aus fortlaufend ergänzten Texten dynamisch korpusbasiert generiert werden.[19]

4 Aufbereitung und Aufbewahrung neu erstellter Datensammlungen

Wenn für die Behandlung einer Forschungsfrage oder Hypothese neue Datensammlungen angefertigt werden müssen, liegt das in der Regel daran, dass die passenden Datensätze nicht verfügbar sind, veraltet sind oder nicht nachgenutzt werden dürfen. Die Erstellung von Datensammlungen erfordert – abhängig von der Art der Daten und der erforderlichen Analyse – erhebliche Aufwendungen, gerade wenn Forschende die Daten zeitintensiv manuell aufbereiten müssen. Um den größtmöglichen Nutzen aus den Daten zu ziehen und um sie auch zusammen mit anderen Datensätzen verwenden zu können, setzen Werkzeuge gemeinsame Datenformate und Annotationen voraus. Dazu werden, wenn irgend möglich, vorhandene Standards und Normen verwendet, die neben den Verfahren in Datenmanagementplänen beschrieben werden, die

8 Kupietz et al. (2010).

9 Geyken et al. (2010).

10 Siehe Kupietz und Lüngen (2014). Die jeweils aktuelle Anzahl der Token ist auf den Webseiten des Korpus unter http://www.ids-mannheim.de/kl/projekte/korpora/ zu finden, für März 2016 wird die Größe mit 29 Mrd. Token angegeben.

11 Aktualisierte Zahlen zum DTA findet man unter http://www.deutschestextarchiv.de/, die Angaben hier beziehen sich auf September 2016.

12 Telljohann et al. (2004).

13 Brants et al. (2004).

14 Siehe Skut et al. (1997).

15 Siehe Schiel et al. (1997).

16 Drude et al. (2012).

17 Klein und Geyken (2010).

18 Hamp und Feldweg (1997), Henrich und Hinrichs (2010).

19 Siehe Quasthoff und Richter (2005) und http://wortschatz.uni-leipzig.de/.

den gesamten Lebenszyklus von Forschungsdaten abdecken. Dazu gehört die Erstellung, Aufbereitung, Archivierung und Bereitstellung zur Nachnutzung. Bibliotheken und Rechenzentren an wissenschaftlichen Einrichtungen und Universitäten unterstützen außerdem elektronische Publikationen über eigene Publikationsplattformen, allerdings können fachnahe Infrastrukturen für die Geistes- und Sozialwissenschaften gezielter die Besonderheiten der dort verwendeten Daten aufgreifen und die Aufbereitungs- und Aufbewahrungsprozesse forschungsnah begleiten.

4.1 Erstellung interoperabler Datensätze

Die Erstellung von eigenen Datensätzen in standardisierten und abgestimmten Formaten erlaubt die Verwendung von existierenden Werkzeugen und Methoden und ermöglicht eine Vergleichbarkeit von Ergebnissen. Daher ist es notwendig, bei der Erstellung die existierenden Standards zu kennen und zu verwenden. Dazu gehört auch die Auswahl von Datenformaten, z. B. einer angemessenen Variante nach den Empfehlungen der Text-Encoding-Initiative[20] oder anderer Normen, wie für Merkmalsstrukturen[21], für lexikalische Ressourcen[22], für Annotation[23] und für Transkriptionen gesprochener Sprache[24]. Da die Auswahl von Datenformaten von der Art der Daten und der Analyse abhängt, ist die Entscheidung für die technische Implementierung – die Serialisierung der Daten – weder eindeutig noch trivial, entscheidet aber darüber, wie die Daten nachgenutzt werden können. Die Forschenden in den Geistes- und Sozialwissenschaften haben in vielen Fällen nur begrenzte Erfahrungen bei der Auswahl von Datenformaten, Normen und Bestimmung der technischen Anforderungen. Daher wurde innerhalb von CLARIN-D ein Benutzerhandbuch erstellt, das die Erfahrungswerte bündelt und zugänglich macht.[25] Außerdem stehen an den CLARIN-Zentren und in den F-AGs disziplinnahe Experten zur Verfügung, die auf Anfrage, u. a. über einen eigenen von CLARIN-D eingerichteten elektronisch zugänglichen Helpdesk,[26] Kompetenzen einbringen können und sowohl national als auch auf europäischer Ebene zur Konsolidierung bei Fragen zu Datenformaten, Beschreibungen und Verfahren beitragen.

Ein zentrales Beispiel für nationale und internationale Konsolidierung ist die Festlegung eines Metadaten-Rahmens für die Beschreibung von Daten. In CLARIN wurde durch die sehr unterschiedlichen Datentypen die Bedeutung eines modularen Beschreibungsrahmens deutlich, in dem einige Beschreibungsebenen oder Datenkategorien zu Modulen zusammengefasst und somit auch für andere Datentypen wiederverwendet werden können. Die Datenkategorien selbst werden zentral definiert, kommen aber nur dann, wenn sie für einen Datentyp relevant sind, in der Beschreibung von Daten zum Einsatz. Zum Beispiel ist die Anzahl der Token für geschriebene Korpora relevant, die Aufnahmedauer für gesprochene Korpora, für lexikalische Ressourcen dagegen die Anzahl der Lexikoneinträge. Die zentrale Definition erlaubt dennoch eine Interoperabilität, so dass diese Kategorien in der Suche nach Daten zum Beispiel im VLO verwendet werden können. Innerhalb von CLARIN wird daher die Component Metadata Infrastructure (CMDI) nach ISO 24622-1 (2015), als gemeinsamer Metadatenrahmen verwendet. Die CLARIN-Zentren unterstützen Dritte bei der Verwendung für selbst erstellte Datensätze, sodass die Daten zur Nachnutzung zum Beispiel über das CLARIN-Netzwerk bereitgestellt werden können. Über auch von Bibliotheken verwendete Standardschnittstellen wie OAI-PMH (das *Open Archives Initiative Protocol for Metadata Harvesting*)[27] können die Metadaten in Bibliotheks- und Archiv-Katalogen nachgewiesen werden, wobei die Metadaten, die dort eingepflegt werden können, nicht die gleiche Beschreibungstiefe besitzen müssen.

4.2 Datenmanagement

Neben der Beschreibung der Daten selbst ist zur Nachnutzung von automatisierten und manuell erstellten Daten in anderen Forschungskontexten ein Datenmanagement erforderlich. Durch das Datenmanagement sind die Abläufe und Verfahren im Umgang mit Daten in jeder Phase eines Forschungsvorhabens bestimmt, d. h. von der Erhebung, Lokalisierung und Erwerb von Ausgangsdaten, über die Analyse und Archivierung bis zur Bereitstellung von Daten zur Nachnutzung. Ein strukturiertes Datenmanagement erlaubt es, dass Daten als Teil der wissenschaftlichen Leistung sichtbar werden, Forschungsergebnisse leichter verifiziert werden können und Daten in *enhanced publications* erscheinen und damit als Referenz dienen können, wobei OAI-PMH (2002–2015) die gesteigerte Sichtbarkeit der Da-

20 Siehe TEI (2016).
21 ISO 24610-1 (2006).
22 ISO 24613 (2008).
23 ISO 24612 (2012).
24 ISO 24624 (2016).
25 Herold und Lemnitzer (2012).
26 Lehmberg (2015).
27 Siehe OAI-PMH (2002–2015).

ten den Publikationscharakter von Forschungsdaten erhöht. Daneben dringen auch Drittmittelgeber auf eine Nachnutzbarkeit von Ausgangsdaten.

Um das Datenmanagement zu dokumentieren, muss von den Forschenden im Vorfeld der eigentlichen Untersuchungen ein Datenmanagementplan erstellt werden, in den die Maßnahmen und Festlegungen im Bereich des Datenmanagements spezifiziert werden, und in dem dargestellt wird, wo Daten in vertrauenswürdigen Archiven abgelegt werden sollen. Forschungsförderungsorganisationen wie etwa das Bundesministerium für Bildung und Forschung (BMBF) oder die Deutsche Forschungsgemeinschaft (DFG) gehen mehr und mehr dazu über, im Rahmen von Projektantragsprozessen einen Datenmanagementplan vorauszusetzen. Dies dient neben der Schaffung von Synergieeffekten auch der Qualitätssicherung im Bereich der datenbasierten Forschung, da Analysen reproduzierbar und somit überprüfbar werden (vgl. die Empfehlungen der DFG Kommission „Selbstkontrolle in der Wissenschaft“ und die DFG-Leitlinien zum Umgang mit Forschungsdaten im Literaturverzeichnis).

Um die Zusammenarbeit zwischen Antragstellenden von Forschungsprojekten und Forschungsinfrastrukturen zu vereinfachen, hat CLARIN-D mit DMPTY[28] ein Hilfsmittel zur interaktiven Erstellung von Datenmanagementplänen geschaffen. Dadurch soll frühzeitig in Projekten die Kooperation zwischen Forschenden und CLARIN-Zentren erreicht werden, um realistische Aufwandsabschätzungen für ein Projekt zu entwickeln und Daten standardkonform und interoperabel aufzubereiten. CLARIN-D unterstützt Forschende bei der Erstellung und Durchführung ihres Datenmanagementplans und sichert so die nachhaltige Verfügbarkeit der Forschungsdaten.

4.3 Aufbewahren von Ergebnissen bei einem CLARIN-D-Zentrum

Die nachhaltige Verfügbarkeit von Forschungsdaten, wie sie im Datenmanagementplan beschrieben werden muss, bezieht sich auf die mehrjährige Speicherung und Zugriffsmöglichkeit auf Daten. So sollen Forschungsdaten, die in Drittmittelprojekten entstehen, über mindestens zehn Jahre aufbewahrt werden (siehe z. B. die DFG-Leitlinien zum Umgang mit Forschungsdaten im Literaturverzeichnis). Um die Aufbewahrung von Daten über das Projektende hinaus zu gewährleisten, betreiben Forschungsinfrastrukturen Repositorien – Archive für Forschungsdaten, über die die Daten selbst zugänglich gemacht werden. Die Aufbewahrung von Forschungsdaten in besonders einschlägigen nationalen und internationalen Datenzentren ist ein wichtiger Faktor, um Sichtbarkeit zu erreichen. Die deutschen CLARIN-Zentren nutzen ihre Repositorien nicht nur, um eigene Daten nachhaltig zu speichern, sondern bieten die Datenübernahme auch für Sprachressourcen Dritter an. Für die Übernahme von Daten durch Repositorien werden vertragliche Rahmen definiert und beschrieben, welche Rechte zur Nutzung der Ressourcen vorliegen und unter welchen Bedingungen Daten weitergegeben werden dürfen.

Die Übernahme, Weitergabe und längerfristige Speicherung von Daten setzen voraus, dass Repositorien technisch so umgesetzt sind, dass sie verlässlich und langfristig operieren können. Auch im Datenmanagementplan muss glaubhaft gemacht werden, dass die Speicherung und Auffindbarkeit langfristig gewährleistet ist. Um die Verlässlichkeit und Sicherheit von Repositorien zu dokumentieren, werden die CLARIN-Zentren regelmäßig überprüft, wozu zwei verschiedene Modelle verwendet werden: Eine externe Evaluation im Rahmen des *Data Seal of Approval* (DSA)[29] dient der Überprüfung der Prozesse und Verfahren in den Repositorien. Ein CLARIN-internes Verfahren zur Zertifizierung von Zentren als sogenannte CLARIN-B-Zentren bewertet die Einhaltung der innerhalb der CLARIN-Initiative definierten Standards und Verfahren.

Neben technischen Fragen können rechtliche Fragen zu den Ressourcen eine hohe Komplexität aufweisen und zwar stärker als bei Druckpublikationen, für die akademische Bibliotheken Regelungen im Rahmen ihrer Publikationsinfrastrukturen bieten. Die höhere Komplexität resultiert daraus, dass Forschungsdaten im geistes- und sozialwissenschaftlichen Bereich in der Regel nicht nur die Rechte der Forschenden berühren, in deren Forschungsprozess Daten entstehen oder verwendet werden, sondern auch die Rechte weiterer juristischer oder natürlicher Personen, zum Beispiel von Verlagen und Autoren bei Textsammlungen, Sprechern bei Tonband und Videoaufnahmen etc. Daher werden diese Aspekte in einem Datenüberlassungsvertrag der Datengeber mit den Repositorien betrachtet, um dem Repositoriumsbetreiber eine rechtliche Grundlage zur Speicherung und Herausgabe von Daten zu geben. Auch bei diesen komplexen rechtlichen Fragen bietet CLARIN Unterstützung durch Handreichungen und Expertise.[30]

28 Trippel und Zinn (2015).

29 Sesink et al. (2010).

30 siehe z. B. Kamocki und Ketzan (2014), Kamocki et al. (2016).

Mit der Unterstützung beim Datenmanagement und mit dem Betrieb zertifizierter Repositorien bieten CLARIN-Zentren Lösungen für die sprachbasiert arbeitenden Geistes- und Sozialwissenschaften, die die fachlichen Besonderheiten und die Variationsbreite der Daten und betroffenen Rechte berücksichtigen.

5 Auswerten von Daten

Unabhängig davon, ob Daten nachgenutzt oder selbst erstellt werden, kommt der Auswertung der Daten eine zentrale Rolle im Forschungsprozess zu. Im Bereich der Sprachressourcen gehen verfügbare Analysewerkzeuge über reine Statistiksysteme hinaus, indem sie die Annotationen und Strukturen berücksichtigen. Diese Werkzeuge sind abhängig von der Annotation, den Dateiformaten und den wissenschaftlichen Modellen, die den Daten zugrunde liegen, und sind entsprechend variationsreich und eng mit den Daten verschränkt. Daneben können digitalisierte Daten manuell und automatisiert weiterverarbeitet werden. CLARIN-Zentren stellen dafür virtuelle Forschungsumgebungen mit webbasierten Werkzeugen sowohl für gesprochene als auch geschriebene Sprache zur Verfügung.

5.1 Abfragen von Daten

Die Abfrage von Referenzdatensätzen (vgl. Abschnitt 3), die deren Annotationen und Strukturen nutzt, geht über eine generische Volltextsuche und den FCS hinaus und erfordert, dass die Suchwerkzeuge auf den Datentyp angepasst sind. Daher unterhalten die CLARIN-Zentren mit Referenzdaten solche Suchwerkzeuge, auf die auch externe Forschende aus dem akademischen Bereich über das Web zugreifen können.

Für das DeReKo-Korpus steht mit *KorAP*[31] als Nachfolger des COSMAS II-Systems[32] am IDS in Mannheim eine Suchmöglichkeit bereit, mit der z. B. Kollokationen bei Anfragen berücksichtigt werden können. KorAP erlaubt es, auch Teilsammlungen zu untersuchen, etwa eingeschränkt nach Gattung, Genre und Quelle.

Für das DTA besteht neben einer Volltextsuche die Möglichkeit, zeitliche Einschränkungen für die Suche vorzunehmen und so sprachliche Entwicklungen zu untersuchen. Dazu steht mit *DiaCollo*[33] ein Werkzeug zur diachronen Kollokationsanalyse zur Verfügung, das an der BBAW entwickelt wurde. Ebenfalls an der BBAW wurde im Rahmen der Entwicklung des Digitalen Wörterbuchs der deutschen Sprache (DWDS) ein Konkordanzsystem entwickelt, das *DWDS/Dialing Concordance* (DCC).[34]

Syntaktische Strukturen von Baumbanken kann man mit *Tündra*[35] untersuchen und visualisieren. Weitere spezialisierte Werkzeuge gibt es auch für lexikalische Ressourcen, z. B. für das Leipziger Wortschatzprojekt und mit dem Wortauskunftssystem zur deutschen Sprache des DWDS.[36]

Werkzeuge wie Tündra eignen sich auch zur Untersuchung eigener Daten, wenn diese entsprechend aufbereitet vorliegen. Diese Verarbeitung kann mit CLARIN-Werkzeugen je nach Anforderung manuell oder automatisch erfolgen.

5.2 Annotation und Alignierung digitaler Sprachdaten

Die Auswertung von Forschungsdaten setzt in vielen Forschungsprojekten eine theoriegeleitete Datenannotation voraus. Der damit verbundene Arbeitsprozess kann durch den Einsatz von Softwarewerkzeugen zur manuellen oder (semi-)automatischen Annotation wesentlich erleichtert werden. Für die manuelle Annotation von textuellen Daten wurde im CLARIN-D-Verbund die Webapplikation *WebAnno* entwickelt.[37] Für die gesprochene Sprache stellt CLARIN die vielgenutzte Annotationssysteme *ELAN*[38] und *Exmaralda*[39] zur Verfügung, die im Umfeld der CLARIN-D-Zentren in Hamburg und am MPI entstanden sind. Sie erlauben eine Mehrebenenannotation gesprochener und multimodaler Daten und finden in der Erforschung von gesprochener Sprache weltweit breite Verwendung und werden speziell zur Dokumentation bedrohter Sprachen eingesetzt.

Automatische Analysewerkzeuge für geschriebene Sprache sind in CLARIN-D als Webservices realisiert, die mit Standard-Verfahren aufgerufen werden können. Technisch gesehen werden die Werkzeuge als sogenannte *RESTful Web Services* implementiert. Diese Verarbeitungsdienste werden abhängig von der gewünschten Analysetiefe nacheinander in sogenannten Prozessketten ausgeführt, in

31 Diewald et al. (2016), Bański et al. (2013).
32 Bodmer (1996).
33 Jurish et al. (2016).
34 Siehe Sokirko (2003).
35 Martens (2013).
36 Siehe unter http://dwds.de/.
37 Eckart de Castilho et al. (2014).
38 Sloetjes und Wittenburg (2008).
39 Schmidt (2012).

denen die Ausgabe eines Service die Eingabe für die nächste Annotationsebene bietet. Zur Verkettung der Webservices steht in CLARIN mit *WebLicht*[40] eine virtuelle Forschungsumgebung zur Verfügung, die Nutzende dabei unterstützt, Prozessketten zur Annotation selbst zu definieren oder auf vordefinierte Prozessketten zuzugreifen. Diese Verarbeitungsketten können sowohl auf nachgenutzte als auch auf eigene Daten angewendet werden. Durch WebLicht stehen für das Deutsche automatische Werkzeuge zur Wortformenerkennung, zur Lemmatisierung, zur Wortklassenbestimmung, zur syntaktischen Annotation und zur Eigennamenerkennung von Personen, Institutionen und geographischen Angaben zur Verfügung.

Neben Webservices für geschriebene Sprache stellt das CLARIN-Zentrum an der LMU Werkzeuge zur automatischen Verarbeitung gesprochener Sprache zur Verfügung. Diese Services dienen der automatischen phonetischen Transkription von Sprachaufnahmen (WebMINNI) und der Alignierung von Sprachaufnahmen mit deren Transkription (WebMAUS). Durch die Alignierung von Sprachsignalen mit deren Transkriptionen[41] können Sprachaufnahmen nach Einzelwörtern und Wortsequenzen gezielt durchsucht werden. Diese Funktionalität ist für computergestütztes Fremdsprachenlernen ebenso relevant wie für das Data Mining in großen Sprachkorpora, wie sie in den Sozial- und Politikwissenschaften verwendet werden.

In enger Absprache mit den CLARIN-D F-AGs wird das bestehende Angebot an Softwarewerkzeugen kontinuierlich ergänzt, um so die angebotenen Services an den aktuellen Stand der Wissenschaft anzupassen oder auf weitere Sprachen und Datentypen zu erweitern. So können Nutzende unterschiedlicher Fachrichtungen das Potential der Werkzeuge für ihre eigenen konkreten Fragestellungen anpassen und optimal ausschöpfen.

6 Zusammenfassung und Ausblick

CLARIN bietet als fachnahe Infrastruktur Daten und Dienste an, die über das hinausgehen, was generische Infrastrukturen wie Bibliotheken und Rechenzentren für die gesamte Wissenschaftslandschaft leisten können.

Vor dem Hintergrund der zunehmenden Verbreitung digitaler Arbeitstechniken, der Entwicklung neuer Fragestellungen, die digitale Infrastrukturen voraussetzen, und der Anforderungen von Drittmittelgebern hinsichtlich eines nachhaltigen Forschungsdatenmanagements ist zu erwarten, dass die Menge an digitalen Forschungsdaten und der Bedarf an deren Nachnutzung in den Geistes- und Sozialwissenschaften stetig zunehmen wird. Um diesen Bedarfen adäquat Rechnung tragen zu können, werden Forschungsinfrastrukturen wie CLARIN einen Prozess der Institutionalisierung vollziehen müssen und wird der gegenwärtige CLARIN-Zentrenverbund durch weitere fachnahe und thematisch einschlägige Datenzentren erweitert werden müssen.

7 Danksagung

Als verteilte Infrastruktur sind viele Einzelpersonen an CLARIN beteiligt, die an den CLARIN-Zentren die Infrastruktur betreiben, Ressourcen bearbeiten und bereitstellen, Expertise und Beratungsleitungen einbringen. Ohne sie wären CLARIN und dieser Artikel nicht möglich. Wir danken auch den Forschenden, die in den CLARIN-Fachgruppen mitarbeiten und wertvolle Rückmeldungen zu den Werkzeugen und Diensten von CLARIN geben, sowie Weiterentwicklungen initiieren und daran mithelfen.

Die CLARIN-Zentren werden in Deutschland derzeit durch BMBF gefördert, der Projektträger ist das Deutsche Zentrum für Luft- und Raumfahrt (DLR). Außerdem beteiligen sich die Institutionen, die CLARIN-D-Zentren beherbergen, und die jeweiligen Bundesländer an der Finanzierung.

Literaturverzeichnis

Bański, P.; Bingel, J.; Diewald, N.; Frick, E.; Hanl, M.; Kupietz, M.; Pezik, P.; Schnober, C.; Witt, A. (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference*, Poznań, Polen. Verfügbar unter http://korap.ids-mannheim.de/wp-content/uploads/2013/12/ltc-demo-126-banski.pdf.

Baum, C.; Stäcker, T. (2015): Methoden – Theorien – Projekte. In: *Zeitschrift für digitale Geisteswissenschaften.* (Sonderband 1: Grenzen und Möglichkeiten der Digital Humanities), 4–12. DOI 10.17175/sb001_023.

Berry, D. M. (2011): The Computational Turn: Thinking about the Digital Humanities. In: *Cultural Machine,* (12), 1–22. Verfügbar unter http://www.culturemachine.net/index.php/cm/article/view/440.

Bodmer, F. (1996): Aspekte der Abfragekomponente von COSMAS II. In: *LDV-INFO,* (8), 112–22.

Brants, S.; Dipper, S.; Eisenberg, P.; Hansen, S.; König, E.; Lezius, W.; Rohrer, C.; Smith, G.; Uszkoreit, H. (2004): TIGER: Linguistic Interpretation of a German Corpus. In: *Journal of Language and Computation,* (2), 597–620. DOI:10.1007/s11168-004-7431-3.

40 Siehe E. Hinrichs et al. (2010), M. Hinrichs et al. (2010).

41 Siehe Kisler et al. (2012).

Diewald, N.; Hanl, M.; Margaretha, E.; Bingel, J.; Kupietz, M.; Bański, P.; Witt, A. (2016): KorAP Architecture - Diving in the Deep Sea of Corpus Data. In: N. Calzolari et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenien, 3586–91. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2016/pdf/243_Paper.pdf.

Drude, S.; Trilsbeek, P.; Broeder, D. (2012): Language Documentation and Digital Humanities: The (DoBeS) Language Archive. In: J. C. Meister (Ed.): *Digital Humanities 2012*, Hamburg. Verfügbar unter http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/language-documentation-and-digital-humanities-the-dobes-language-archive.1.html.

Eckart de Castilho, R.; Biemann, C.; Gurevych, I.; Yimam, S. M. (2014): WebAnno: a flexible, web-based annotation tool for CLARIN. In: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Niederlande. Verfügbar unter https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf.

Geyken, A.; Haaf, S.; Jurish, B.; Schulz, M.; Steinmann, J.; Thomas, C.; Wiegand, F. (2010): Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In: S. Schomburg et al. (Eds.): *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, Köln, 157–61. Verfügbar unter http://fiz1.fh-potsdam.de/volltext/DigiWis/13472.pdf.

Hamp, B.; Feldweg, H. (1997): GermaNet – a Lexical-Semantic Net for German. In: *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spanien, 9–15. Verfügbar unter http://aclweb.org/anthology/W97-0802.

Henrich, V.; Hinrichs, E. (2010): Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, 456–64. Verfügbar unter http://www.aclweb.org/anthology/C10-1052.pdf.

Herold, A.; Lemnitzer, L. (2012): *CLARIN-D User Guide*. BBAW, Berlin. Verfügbar unter http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf.

Hinrichs, E.; Hinrichs, M.; Zastrow, T. (2010): WebLicht: Web-Based LRT Services for German. In: *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, Upsala, Schweden, 25–29. Verfügbar unter http://aclweb.org/anthology/P10-4005.

Hinrichs, E.; Krauwer, S. (2014): The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: N. Calzolari et al. (Eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavík, Island, 1525–31. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

Hinrichs, M.; Zastrow, T.; Hinrichs, E. (2010): WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In: N. Calzolari et al. (Eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 489–93. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2010/pdf/270_Paper.pdf.

ISO 24610-1 (2006): Language resource management – Feature structures – Part 1: Feature structure representation. Internationale Norm. Genf.

ISO 24612 (2012): Language resource management – Linguistic annotation framework (LAF). Internationale Norm. Genf.

ISO 24613 (2008): Language resource management - Lexical markup framework (LMF). Internationale Norm. Genf.

ISO 24622-1 (2015): Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. Internationale Norm. Genf.

ISO 24624 (2016): Language resource management – Transcription of spoken language. Internationale Norm. Genf.

Jurish, B.; Geyken, A.; Werneke, T. (2016): DiaCollo: diachronen Kollokationen auf der Spur. In: *DHd 2016: Modellierung - Vernetzung - Visualisierung*, Leipzig, 172–75. Verfügbar unter http://www.dhd2016.de/abstracts/vortr%C3 %A4ge-041.html.

Kamocki, P.; Ketzan, E. (2014): Creative Commons and Language Resources: General Issues and what's new in CC 4.0. In: *CLARIN Legal Issues Committee (CLIC)-White Paper Series*. Verfügbar unter https://www.clarin-d.de/images/legal/CLIC_white_paper_1.pdf.

Kamocki, P.; Stranák, P.; Sedlák, M. (2016): The Public License Selector: Making Open Licensing Easier. In: N. Calzolari et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenien, 2533–38. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2016/pdf/880_Paper.pdf.

Kisler, T.; Schiel, F.; Sloetjes, H. (2012): Signal processing via web services: the use case WebMAUS. In: J. C. Meister (Ed.): *Digital Humanities 2012 Conference Abstracts* (Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts), Hamburg, 30–34. Verfügbar unter https://www.clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf.

Klein, W.; Geyken, A. (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: *Lexikographica*, 79–93.

Kupietz, M.; Belica, C.; Keibel, H.; Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: N. Calzolari et al. (Eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta, 1848–54. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, M.; Lüngen, H. (2014): Recent Developments in DeReKo. In: N. Calzolari et al. (Eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavík, Island, 2378–85. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf.

Lehmberg, T. (2015): Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities. In: *Forschungsdaten in den Geisteswissenschaften (FORGE 2015)*, Hamburg, 25–27. Verfügbar unter https://www.gwiss.uni-hamburg.de/gwin/ueber-uns/forge2015/forge2015abstracts.pdf.

Martens, S. (2013): TüNDRA: A Web Application for Treebank Search and Visualization. In: S. Kübler et al. (Eds.): *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, Sofia, Bulgarien, 133–44.

OAI-PMH Version 2.0 (2002-2015): The Open Archives Initiative Protocol for Metadata Harvesting. Technische Spezifikation. Verfügbar unter http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Presner, T. (2010): Digital Humanities 2.0: A Report on Knowledge. OpenStax CNX. Verfügbar unter http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6.

Quasthoff, U.; Richter, M. (2005): Projekt Deutscher Wortschatz. In: *Babylonia,* (3), 33–35.

Schaal, S. G.; Kath, R. (2014): Zeit für einen Paradigmenwechsel in der politischen Theorie? In: A. Brodocz et al. (Eds.): *Die Verfassung des Politischen: Festschrift für Hans Vorländer*. 331–50. Springer Fachmedien Wiesbaden, Wiesbaden. DOI 10.1007/978-3-658-04784-9_20.

Schiel, F.; Draxler, C.; Tillmann, H. G. (1997): The Bavarian Archiv for Speech Signals: Resources for the Speech Community. In: *Proceedings of EUROSPEECH 1997*, Rhodos, Griechenland, 1687-1690.

Schmidt, T. (2012): EXMARaLDA and the FOLK tools — two toolsets for transcribing and annotating spoken language. In: N. Calzolari et al. (Eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Türkei, 236–40. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Sesink, L.; van Horik, R.; Harmsen, H. (2010): *Data Seal of Approval - Guidelines version 1.* Data Archiving and Network Services (DANS), Den Haag, Niederlande. Verfügbar unter http://www.datasealofapproval.org/media/filer_public/2013/09/27/guidelines_01-june-2010.pdf.

Skut, W.; Krenn, B.; Brants, T.; Uszkoreit, H. (1997): An Annotation Scheme for Free Word Order Languages. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing {ANLP}-97*, Washington, DC, USA.

Sloetjes, H.; Wittenburg, P. (2008): Annotation by Category: ELAN and ISO DCR. In: N. Calzolari et al. (Eds.): *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakesch, Morokko. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf.

Sokirko, A. (2003): A technical overview of DWDS/Dialing Concordance. In: *Computational linguistics and intellectual technologies*, Protwino; Russland. Verfügbar unter http://www.aot.ru/docs/OverviewOfConcordance.htm.

Stehouwer, H.; Ďurčo, M.; Auer, E.; Broeder, D. (2012): Federated Search: Towards a Common Search Infrastructure. In: N. Calzolari et al. (Eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Türkei, 3255–59. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf.

TEI P5 (2016): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Richtlinien. Verfügbar unter http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Telljohann, H.; Hinrichs, E.; Kübler, S. (2004): The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In: M. T. Lino et al. (Eds.): *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*, Lissabon, Portugal, 2229–35. Verfügbar unter http://lrec-conf.org/proceedings/lrec2004/pdf/135.pdf.

Thiel, T. (2012): Digital Humanities: Eine empirische Wende für die Geisteswissenschaften? In *Frankfurter Allgemeine Zeitung*, Frankfurt. Verfügbar unter http://www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html.

Trippel, T.; Zinn, C. (2015): DMPTY - A Wizard For Generating Data Management Plans. In: K. De Smedt (Ed.): *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*. 71–78. Linköping University Electronic Press, Linköping, Schweden. Verfügbar unter http://www.ep.liu.se/ecp/123/006/ecp15123006.pdf.

Van Uytvanck, D.; Stehouwer, H.; Lampen, L. (2012): Semantic metadata mapping in practice: The Virtual Language Observatory. In: N. Calzolari et al. (Eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Türkei. Verfügbar unter http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf.

Prof. Erhard Hinrichs
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Wilhelmstr. 19
D-72074 Tübingen
erhard.hinrichs@uni-tuebingen.de

Thorsten Trippel
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Wilhelmstr. 19
D-72074 Tübingen
thorsten.trippel@uni-tuebingen.de