Vernetzung statt Vereinheitlichung Digitale Forschungsinfrastrukturen in den Geisteswissenschaften

Hanna Hedeland, Daniel Jettka, Timm Lehmberg

Die Entwicklung der digitalen Infrastruktur am Hamburger Zentrum für Sprachkorpora (HZSK) kann als Beispiel für die Evolution individueller technischer Einzellösungen hin zu fachspezifischen virtuellen Arbeits- und Forschungsumgebungen, die im Rahmen supranationaler Forschungsinfrastrukturen für die digitalen Geisteswissenschaften miteinander vernetzt sind, angesehen werden. Im Fokus steht im konkreten Fall des HZSK die Sicherung der langfristigen Zugänglichkeit von Forschungsdaten (multimedialen Daten gesprochener Sprache) durch die Entwicklung einer virtuellen Forschungsumgebung, die einerseits an die zentrenbasierte Forschungsinfrastruktur CLARIN-D angebunden ist und andererseits fachspezifische Benutzerschnittstellen schafft.

The development of the digital infrastructure at the Hamburg Center for Language Corpora (Hamburger Zentrum für Sprachkorpora - HZSK) can be seen as an example for the evolution of individual technical solutions towards community-specific virtual workspaces and research environments that are interconnected in the context of supranational research infrastructures for the digital humanities. In the case of the HZSK the focus lies on the assurance of the long-term accessibility of research data (multimedial data of spoken language) by developing a virtual research platform, which on the one hand is connected to the center-based research infrastructure CLARIN-D, and on the other hand provides community-specific user interfaces.

Einleitung

Die Formulierung und Überprüfung von Hypothesen auf der Grundlage von Sammlungen sprachlicher Daten (Korpora) hat sich in den vergangenen zwanzig Jahren zu einem integralen Bestandteil empirisch fundierter Ansätze in den Sprachwissenschaften entwickelt.

Wesentlichen Einfluss hierauf hatten nicht zuletzt die umfassenden Möglichkeiten, die aus den neu entstandenen Methoden der digitalen Aufbereitung und Analyse sowie der Online-Zugänglichkeit sprachlicher Ressourcen erwuchsen. Erstmalig war es auch kleineren Projekten und Forschungsvorhaben möglich, unter vertretbarem Aufwand wertvolle Korpora, lexikalische Ressourcen, etc. zu schaffen, die zum Teil einzigartige Dokumente des realen Sprachgebrauchs (in Wort und Schrift) sind.

Kehrseite dieser Entwicklung war dabei eine zunehmende Vielfalt an (teilweise redundanten) Ressourcen, die mithilfe unterschiedlichster Werkzeuge (vom Textverarbeitungsprogramm bis zum linguistischen Annotationstool) und Datenformate seit den 1990er Jahren generiert wurden.

Dies führte schnell zu der Erkenntnis, dass, um einer Entstehung von „Datenfriedhöfen“ entgegenzuwirken[1], sowohl Standards als auch Strukturen geschaffen werden mussten, die eine nachhaltige Zugänglichkeit digitaler Ressourcen und Werkzeuge für künftige Forschungsvorhaben gewährleisten würden.

Der hier vorliegende Beitrag zeigt exemplarisch am Beispiel der Entstehung des Hamburger Zentrums für Sprachkorpora (HZSK) und seiner digitalen Infrastruktur die Entwicklung von individuellen Einzellösungen hin zur Entstehung virtueller Arbeits- und Forschungsumgebungen, die in supranationaler Forschungsinfrastrukturen eingebettet sind. Ein besonderes Augenmerk soll dabei auf der Erstellung und Analyse gesprochensprachlicher Datensammlungen liegen, die den thematischen Schwerpunkt des HZSK ausmachen.

Multimediale und multidimensionale Daten: Korpora gesprochener Sprache

Korpora gesprochener Sprache nehmen auf dem Feld der nachhaltigen Aufbereitung und Publikation linguistischer Ressourcen in vielerlei Hinsicht eine besondere Rolle ein, da sie infolge ihrer Multidimensionalität[2] besondere Fragestellungen bei der Standardisierung und Vernetzung aufwerfen. Sie sind damit ein besonders geeignetes Beispiel für die Notwendigkeit eines Paradigmenwechsel von der Standardisierung

1 Vgl. Schmidt, T./ Chiarcos, C./ Lehmberg, T./ Rehm, G./ Witt, A./ Hinrichs, E.: Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources, in: Tools and Standards: The State of the Art (Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation) Lansing, Michigan 2006.

2 Gemeint ist damit die Komplexität der in gesprochensprachlichen Ressourcen enthaltenen Informationsebenen und ihre Korrelation.

im Sinne einer Generalisierung von Datenstrukturen und -formaten hin zur Schaffung digitaler Infrastrukturen, die lokale, an spezifischen Bedarfen orientierte Lösungen integriert, und mit Hilfe von Infrastrukturlösungen in einem allgemeinen Kontext zugänglich macht.

Zunächst unterscheiden sich Korpora gesprochener Sprache naturgemäß von den übrigen, zumeist konzeptionell und medial schriftlichen, Ressourcen hinsichtlich des zugrundeliegenden Mediums, nämlich flüchtiger gesprochener Sprache, die in der Regel in Form von Audio- und Videoaufnahmen erhoben wird. Dabei handelt es sich zumeist um vergleichsweise große Datenmengen, die in guter Ton- und Bildqualität, (teilweise in gestreamter Form online) für alle Verarbeitungsschritte zur Verfügung stehen müssen. Sie bilden die Grundlage für Analysen, die sich neben den üblichen Beschreibungsebenen von Sprache auch auf genuin gesprochensprachliche Untersuchungsfelder wie Phonetik/Phonologie, Multimodalität, Diskurs- und Gesprächsanalysen etc. beziehen können.

Zu diesem Zweck wird das Mediensignal mit Hilfe von für die jeweilige Forschungsfrage geeigneten Konventionen in eine geschriebene Repräsentation überführt. Dieser Vorgang, die Transkription, resultiert in einem über Zeitstempel mit dem Mediensignal verknüpften (alignierten) Transkript.[3] Bereits die Verschriftlichung gesprochener Sprache ist stets interpretativer Natur und muss manuell aufgeführt werden, die gegebenenfalls zusätzlichen analyserelevanten Informationen, so genannte Annotationen, sind in ihrer Erstellung noch kosten- und zeitintensiver.

Transkripte und Mediendateien werden in der Regel mit sehr spezifischen Metadaten zu den beteiligten Sprechern, situativen und technisch-methodischen Kontexten angereichert, die sich ebenfalls signifikant von den zumeist linearen Metadatensets in schriftsprachlichen Korpora unterscheiden. Diese Metadatenkomponenten können wiederum in Relation zueinander stehen, zum Beispiel. um Rollen, die Sprecher in kommunikativen Situationen einnehmen, modellieren zu können und zu referenzieren, in welchen Mediendateien sie erfasst sind. Infolge dieser komplexen Strukturierung und ihres spezifischen Charakters ist diese Art von Metadaten nur bedingt standardisierbar[4]. Alle hier genannten Beschreibungsebenen müssen wiederum in Abhängigkeit des Forschungsschwerpunktes miteinander korreliert werden können.

Standardisierung und Interoperabilität: EXMARaLDA

Diese hohen Anforderungen führten in der jüngeren Vergangenheit zur Entstehung zahlreicher Werkzeuge der Datenaufbereitung. Zu nennen ist dabei bspw. die Entwicklung von EXMARaLDA[5] („Extensible Markup Language for Discourse Annotation“, einem System von Konzepten, nachhaltigen (XML-basierten) Datenformaten und Werkzeugen für die computergestützte Transkription und Annotation gesprochener Sprache, sowie für Erstellen und Auswerten von umfangreichen Korpora gesprochener Sprache. EXMARaLDA gehört mit über 10.000 Anwendern zu den weltweit meistgenutzten Systemen zur Aufbereitung gesprochensprachlicher Ressourcen. Seit 2011 wird die Entwicklung des Systems am HZSK in Zusammenarbeit mit dem Archiv für Gesprochenes Deutsch am Institut für Deutsche Sprache (IDS) in Mannheim weitergeführt[6].

Ein wesentlicher Grund für die weite Verbreitung des EXMARaLDA Systems ist neben seiner vergleichsweise intuitiven Bedienbarkeit, auch der hohe Grad

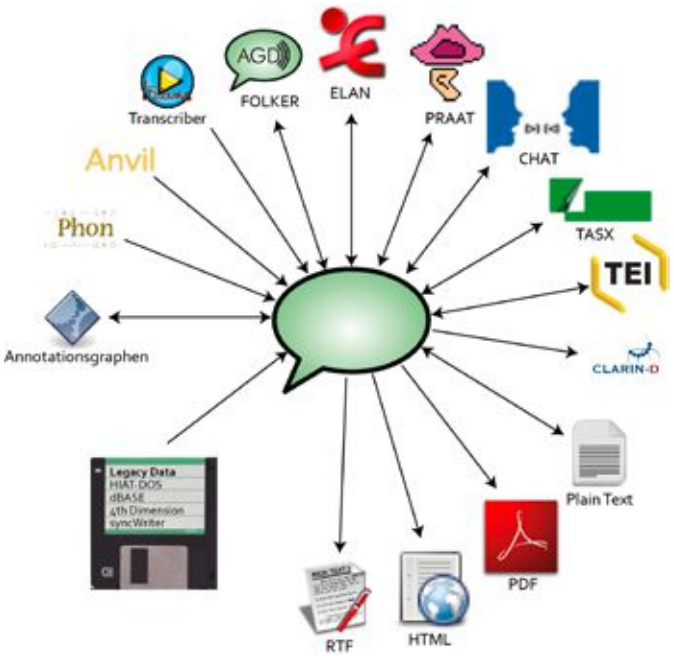

Abb. 1: Interoperabilität von EXMARaLDA mit Standards und Werkzeugen zur Datenaufbereitung

3 Für eine detaillierte Auseinandersetzung mit Methoden und Standards der Transkription siehe Schmidt, T.: Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a. M. 2005.

4 Wörner, K.: Finding the balance between strict defaults and total openness: Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager, in: SCHMIDT, T./ WÖRNER, K. (Hrsg.): Multilingual Corpora and Multilingual Corpus Analysis (Hamburg Studies in Multilingualism, 14). Amsterdam/ Philadelphia 2012, S. 383-400.

5 http://www.exmaralda.org

6 Schmidt, T./ Hedeland, H./ Lehmberg, T./ Wörner, K.: Multilingual Corpora at the Hamburg Centre for Language Corpora, in: Hedeland, H./ Schmidt, T./ Wörner, K. (Hrsg.): Proceedings of the GSCL conference Multilingual Resources and Multilingual Applications Hamburg (Arbeiten zur Mehrsprachigkeit / Working Papers in Multilingualism, Serie B, 96) Hamburg 2011, S. 227-233.

an Interoperabilität mit existierenden Standards und Werkzeugen der Datenaufbereitung, zum Beispiel den Transkriptionswerkzeugen Praat[7] und ELAN[8], dem Datenformat der Text-Encoding Initiative (TEI) sowie zahlreichen gängigen Textverarbeitungs- und (Online-) Publikationsformaten (vgl. Abb.1).

Im Folgenden soll ein kurzer Überblick über einige wichtige Standards der linguistischen Datenaufbereitung erfolgen[9]

Standardisierung und Institutionalisierung

Mit dem Fortschreiten der technischen Möglichkeiten in der Datenaufbereitung stellte sich, wie bereits erwähnt, vermehrt das Problem, dass linguistische Ressourcen aufgrund von uneinheitlicher und z.T. unvereinbarer Repräsentationen und Aufbereitungsmethoden zu großen Teilen nicht oder nur noch eingeschränkt zugänglich waren. Die Gründe hierfür lagen bisher zumeist in der Heterogenität der verwendeten Datenformate und Methoden, einer nicht-vorhandenen institutionellen Verstetigung von Einrichtungen zur Datenvorhaltung und -pflege sowie dem Fehlen institutionsübergreifender Strukturen zur Vernetzung derartiger Einrichtungen.

Diese Situation führte ab Beginn der 1990er Jahre zunächst zu international ausgerichteten Ansätzen zur Definition von Standards der Datenaufbereitung, die es ermöglichen sollten, Daten aus unterschiedlichen Formaten in generische sowie interoperable Datenformate zu überführen, sodass diese möglichst dauerhaft einer großen Nutzergemeinschaft zugänglich gemacht werden können. Zu nennen sind in Bezug auf Annotationsstandards die Arbeiten der Text Encoding Initiative (TEI)[10], der Expert Advisory Group on Language Engineering Standards (EAGLES)[11] und weiterer sich bis heute in kontinuierlicher Weiterentwicklung befindlicher Standards des Komitees ISO/TC 37/SC4 („Language resource management“)[12]. Bezogen auf die Standardisierung von Metadaten ist in diesem Zusammenhang auf die Vorarbeiten der Dublin Core Metadata Initiative (DCMI)[13], der darauf basierenden Erweiterung der Open Language Archives Community (OLAC)[14] sowie der ISLE Metadata Initiative (IMDI)[15] zu verweisen.

Gleichzeitig entstanden weltweit an zahlreichen Forschungs- und Lehreinrichtungen Archive, die die nachhaltige Zugänglichkeit linguistischer Daten - zumeist Korpora eines bestimmten Typs unter Verwendung eines bestimmten Datenstandards - zum Ziel hatten. Prominente internationale Beispiele aus dem Bereich gesprochensprachlicher Ressourcen sind das Archiv der OLAC (s.o.) und das Archiv des Forschungsprogramms Documentation of Endangered Languages (DOBES), welches mit dem Ende des DOBES-Programms in das von der Max-Planck-Gesellschaft, der Berlin Brandenburgischen Akademie der Wissenschaften und der Koninklijke Nederlandse Akademie van Wetenschappen gegründete Archiv TLA – The Language Archive – am Max-Planck-Institut für Psycholinguistik in Nijmegen überführt wurde.

Die Ressourcenlandschaft für gesprochensprachliche Korpora in Deutschland ist im Wesentlichen durch das Bayerische Archiv für Sprachsignale, das Institut für Deutsche Sprache und das Hamburger Zentrum für Sprachkorpora abgedeckt. Dabei ist das HZSK, das an der Universität Hamburg im Jahr 2010 als Zusammenschluss von Angehörigen verschiedener Fachbereiche und Einrichtungen gegründet wurde, die jüngste Einrichtung. Das Zentrum knüpft inhaltlich sowie personell an die Vorarbeiten des zentralen datenverarbeitenden Teilprojekts des Sonderforschungsbereichs 538 - Mehrsprachigkeit an, dessen Aufgabe ab dem Jahr 2000 zunächst darin bestand, Lösungen für die Aufbereitung und Analyse der in 20 empirisch arbeitenden Teilprojekten entstehenden digitalen Sprachdaten zu entwickeln[16].

Die hierin entstandene digitale Infrastruktur wurde in den anschließenden Forschungsprojekten CLARIN-D[17] und Etablierung eines Schwerpunkts ‘Mehrsprachigkeit und Gesprochene Sprache’ am Hamburger Zentrum für Sprachkorpora[18] weiterentwickelt. Die zentralen Anforderungen an die Entwicklung der virtuellen Forschungsumgebung ergeben sich im Wesentlichen aus den flgenden Zielen des HZSK[19]:

7 http://www.praat.org

8 https://tla.mpi.nl/tools/tla-tools/elan/

9 Für eine Übersicht siehe: LEHMERG, T./ WÖRNER, K: Annotation Standards, in: LÜDELING, A./ KYTÖ, M. (Hrsg.): Corpus Linguistics - An international handbook, Berlin 2008, S. 1484-501.

10 http://www.tei-c.org

11 http://www.ilc.cnr.it/EAGLES/home.html

12 http://www.iso.org/iso/standards_development/technical_committees/

13 http://dublincore.org/

14 http://www.language-archives.org/

15 http://www.mpi.nl/IMDI/

16 Vgl. Hedeland, H./ Lehmberg, T./ Schmidt, T./ Wörner, K.: Multilingual Corpora at the Hamburg Centre for Language Corpora, in: Ruhi, S./ Haugh, M./ Schmidt, T./ Wörner, K (Hrsg.): Best Practices for Spoken Corpora in Linguistic Research. Cambridge 2014, S. 208-225.

17 gefördert vom Bundesministerium für Bildung und Forschung

18 gefördert von der Deutschen Forschungsgemeinschaft im Programm Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)

19 vgl. Satzung für das Hamburger Zentrum für Sprachkorpora vom 26.03.2012:

1. Sicherung der Nachhaltigkeit, d.h. der langfristigen Verwendbarkeit und Verfügbarkeit empirischer digitaler Sprachdaten, die zu Forschungs und Lehrzwecken an der Universität Hamburg erstellt und genutzt wurden und werden
2. Entwicklung und Vermittlung von Methoden der computergestützten Datenerstellung, Datenhaltung und Datenanalyse in den Sprachwissenschaften und angrenzenden Disziplinen
3. Vernetzung der Universität Hamburg in der internationalen Sprachressourcen-Landschaft, d.h. insbesondere Integration der Universität Hamburg in bestehende und entstehende digitale Infrastrukturen

Infrastrukturelle Vernetzung von Institutionen und Archiven

Die vorab beschriebenen Standardisierungsbemühungen führten zwar zu einer Etablierung von Archiven und Ressourcen, es zeichnete sich jedoch bald, infolge der notwendigen institutionellen Einbindung, eine zunehmende Tendenz der Dezentralisierung der Ressourcenlandschaft ab. Um dem entgegenzuwirken, dabei den variierenden Anforderungen der Nutzergemeinschaft in den Geisteswissenschaften gerecht zu werden, und diese gleichsam miteinander zu vernetzen, wurde in jüngster Vergangenheit vor allem von Seiten von Fördereinrichtungen des Bundes und der Länder sowie auf europäischer Ebene die Entstehung digitaler Forschungsinfrastrukturen gefördert. Das zugrundeliegende Prinzip dabei ist, dauerhaft existierende Einrichtungen, die Ressourcen, Werkzeuge und Dienste vorhalten, in Forschungsinfrastrukturen derart miteinander zu vernetzen, dass dezentral ein standardisierter und ressourcenübergreifender Zugriff auf dieselben erfolgen kann. Dieser Prozess der Hinwendung zu einem infrastrukturellen Denken ist Teil einer generellen Tendenz, die aktuell in der paneuropäischen Forschungslandschaft zu beobachten ist und die von obersten Regierungsbehörden und Trägern mit initiiert wird.

So wird in der jüngst veröffentlichten Digitalen Agenda 2014 - 2017 der deutschen Bundesregierung die gezielte Förderung der “Vernetzung von Forschungsdatenbanken und Repositorien sowie virtuellen Forschungsumgebungen”[20] mithilfe “strategische[r] Projekte mit großer Hebelwirkung” hervorgehoben. Auf europäischer Ebene übernimmt das European Strategy Forum on Research Infrastructures (ESFRI)[21] der Europäischen Kommission die Identifizierung entstehender pan-

http://www.corpora.uni-hamburg.de/downloads/Satzung.pdf

20 http://www.bmwi.de/DE/Themen/Digitale-Welt/digitale-agenda.html und http://www.digitale-agenda.de/

21 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

europäischer Forschungsinfrastrukturen mit besonders hoher Relevanz für die Qualität der europäischen Forschung und erstellt auf dieser Basis eine Roadmap über die zu priorisierenden Infrastrukturen.

Die bereits erwähnte Common Language Resources and Technology Infrastructure (CLARIN)[22], die sprachliche Ressourcen und Werkzeuge nachhaltig der wissenschaftlichen Öffentlichkeit zur Verfügung stellen soll, wurde bereits 2006 auf diesem Roadmap aufgenommen und übernimmt somit diese Aufgabe im europäischen Raum. CLARIN erwarb 2012 als zweite paneuropäische Forschungsinfrastruktur den Status European Research Infrastructure Consortium (ERIC)[23], wodurch viele administrative und rechtliche Vorteile für den weiteren Ausbau und die Nachhaltigkeit der Forschungsinfrastruktur entstehen. Zu den Gründungsmitgliedern von CLARIN ERIC (Niederlande, Österreich, Tschechien, Dänemark, Estland, Deutschland, Bulgarien, Polen und der Niederländische Sprachunion) werden sich demnächst weitere nationale Konsortien, zum Beispiel Norwegen, anschließen, wodurch die Reichweite des Vorhabens stetig wachsen wird.

In Deutschland schloss CLARIN-D[24] 2011 an das dreijährige vorbereitende Projekt Deutsche Sprachressourcen-Infrastruktur (D-SPIN)[25] an und befindet sich somit bis 2016 in der Implementierungsphase, in der die Infrastruktur sowohl auf der Ebene der neun CLARIN-D-Zentren als auch auf nationaler und auch supranationaler Ebene aufgebaut wird. Beteiligt sind die folgenden Institutionen:

- Bayerisches Archiv für Sprachsignale, Ludwig-Maximilians-Universität München
- Berlin-Brandenburgische Akademie der Wissenschaften
- Institut für Deutsche Sprache, Mannheim
- Max Planck Institut für Psycholinguistik, Nijmegen
- Eberhard Karls Universität Tübingen, Seminar für Sprachwissenschaft
- Universität Hamburg, Hamburger Zentrum für Sprachkorpora
- Universität Leipzig, Institut für Informatik
- Universität des Saarlandes, Englische Sprach- und Übersetzungswissenschaft
- Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Die Vernetzung des HZSK als CLARIN-D-Zentrum geschieht einerseits innerhalb der Forschungsinfrastruktur auf nationaler und europäischer Ebene durch CLARIN-D bzw. CLARIN ERIC. Um den Austausch auch zwischen verwandten Forschungsinfrastrukturen auf dem ESFRI-Roadmap zu erleichtern und Synergiepotentiale zu entdecken, wurde mit dem Projekt Data Service Infrastructure for Social Sciences and Humanities (DASISH)[26] einen Überbau für die fünf Forschungsinfrastrukturen im Bereich Geistes- und Sozialwissenschaften geschaffen. Besonders hervorzuheben ist hier die Zusammenarbeit mit der Digital Research Infrastructure for the Arts and Humanities (DARIAH)[27], die in Österreich und den Niederlanden sogar zu einer Zusammenführung der beiden sich ergänzenden Forschungsinfrastrukturen geführt hat. Diese umfassende Vernetzung des Zentrums sichert die Integration und Kompatibilität der am HZSK vorgehaltenen Ressourcen und Werkzeuge in Bezug auf sich fortentwickelnde Standards und Best Practices für die Aufbereitung und Vorhaltung linguistischer Daten und manifestiert sich nicht zuletzt durch die gegenseitige Absicherung der Datenbestände der Zentren im CLARIN-D-Verbund.

Der supranationale Charakter ermöglicht zudem die Bewältigung nicht-technischer Anforderungen, die beispielsweise aus rechtlichen Fragestellungen der Zugänglichkeit von Ressourcen resultieren. So wird der Zugang zur europaweiten CLARIN Infrastruktur

22 http://www.clarin.eu/

23 http://ec.europa.eu/research/index.cfm?pg=newsalert&lg=en&year=2012&na=na-290212-1

24 http://de.clarin.eu/de/

25 http://weblicht.sfs.uni-tuebingen.de/

26 http://dasish.eu/

27 https://de.dariah.eu/

mit Hilfe einer einzigen Anmeldung (Single Sign-On) für Forschende aus allen Mitgliedsländern mittels der CLARIN Service Provider Federation (SPF) organisiert. Anstatt eine Vielzahl an Nutzungsvereinbarungen untereinander zu treffen, schließen sich dabei Forschungseinrichtungen der SPF an und erlauben damit Mitgliedern aller weiteren Institutionen der Federation den Zugriff auf Ressourcen und Dienste. Auch weitere rechtliche Aspekte wie Lizenztypen und Nutzungsvereinbarungen für die Weitergabe von Ressourcen werden zentrumsübergreifend ausgearbeitet. Jedes CLARIN-Zentrum durchläuft in Bezug auf die Umsetzung der CLARIN-Standards eine Zertifizierung[28], die auch externe Standards, wie etwa das Data Seal of Approval[29] erfordert.

Diese Rahmenbedingungen und Zielsetzungen wirken sich auch, wie im folgenden Abschnitt dargestellt wird, auf die Implementierung der lokalen Infrastruktur in an CLARIN-D angebundenen Forschungseinrichtungen aus.

28 Das HZSK ist beispielsweise ein CLARIN-Zentrum Typ B: hdl:1839/00-DOCS.CLARIN.EU-78

29 http://www.datasealofapproval.org/

Eine virtuelle Forschungsumgebung für die Arbeit mit Sprachkorpora

Mit der Entwicklung des EXMARaLDA-Systems während der Projektlaufzeit des SFB 538 (s.o.) wurde der Grundstein für eine virtuelle Forschungsumgebung für Forschungsprojekte, die Daten auf der Grundlage gesprochener Sprache erheben und auswerten, gelegt. Der Fokus lag dabei zunächst auf der Entwicklung und Bereitstellung von Desktop-Software, Formaten und Workflows für die Transkription und Analyse gesprochener Sprache im Rahmen individueller Forschungsprojekte. Infolge der Weiterentwicklung des WWW und mit dem Entstehen überregionaler Forschungsinfrastrukturen rückten die Publikation und Nachnutzung von Forschungsdaten sowie die Vernetzung mit anderen infrastrukturellen Einrichtungen und Diensten zunehmend in den Mittelpunkt.

Zentrales Merkmal der virtuellen Forschungsumgebung am HZSK war von Anfang an die Zugänglichkeit des sprachlichen Materials an der Universität Hamburg sowie für die internationale Forschergemeinschaft. Die Organisation der Forschungsdaten nach der Aufbereitung mit den Werkzeugen des EXMARaLDA-Systems erfolgte zunächst in einem struktu-

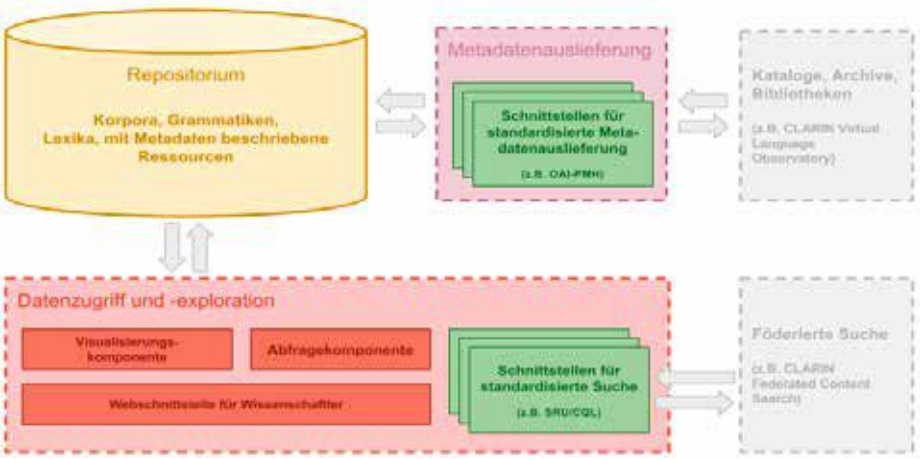

Abbildung 3: Basisinfrastruktur der virtuellen Forschungsumgebung

Abb. 4: Webschnittstelle des Virtual Language Observatory

rierten Dateisystem, aus dem heraus die Daten nach persönlicher Zugangsfreigabe passwortgeschützt über das WWW zugänglich gemacht wurden.

Die Entwicklungen im Bereich der Forschungsinfrastrukturen haben infolge dessen zu einer Weiterentwicklung dieses Systems geführt, zumal für die Vernetzung mit anderen Einrichtungen und der hierfür notwendigen Dienste (quasi-)standardisierte Lösungen zur Verfügung stehen, die deutliche Vorteile gegenüber eigens implementierten Insellösungen aufweisen.

Eine weitverbreitete und vielversprechende Lösung zur langfristigen Speicherung und Verfügbarmachung von Forschungsdaten ist die Flexible Extensible Digital Object Repository Architecture (Fedora)[30]. Fedora wird am HZSK als Grundlage des digitalen Repositoriums (vgl. Abb. 3) eingesetzt. Das System kann prinzipiell zur Speicherung beliebiger digitaler Daten und zur Beschreibung der Beziehungen zwischen digitalen Objekten, also einzelnen explizit zu definierenden Bestandteilen von Datensammlungen (im Fall von Korpora gesprochener Sprache bspw. Audio-/Videoaufnahmen, Transkriptionen, Metadaten oder auch Teilkorpora), mit Hilfe des Resource Description Framework (RDF)[31] genutzt werden. Ein entscheidender Vorteil von Fedora liegt damit in seinem hohen Grad an Flexibilität. Zwar erfordert das System einerseits eine explizite Modellierung der zu speichernden Daten und ihrer gegenseitigen Beziehungen, durch die dadurch erlangte Flexibilität wird jedoch unter anderem die nachhaltige Verfügbarkeit und Transferierbarkeit der Daten auf ein anderes Repositorien erleichtert. Dies kann zum Beispiel erforderlich werden, wenn das Fortbestehen des Repositoriums nicht mehr gewährleistet sein sollte oder andere Repositoriensysteme (angesichts unabsehbarer technischer Entwicklungen) genutzt werden müssen.

Die Möglichkeiten des webbasierten Zugriffs auf die Inhalte des Fedora-Repositoriums, der durch XACML[32] Policies und andere Authentifizierungsmechanismen kontrolliert werden kann, bietet Anknüpfungspunkte zur Erweiterung des Repositoriums um Schnittstellen zu externen Infrastruktureinrichtungen wie Datenzentren, Archiven und Bibliotheken, bspw. zum Zweck der Auslieferung von Metadaten, oder auch für den erweiterten Zugriff und die Exploration der Forschungsdaten (vgl. Kästen “Metadatenauslieferung” und “Datenzugriff- und exploration” in Abb. 3). Mit Hilfe eines OAI[33] Providers können zum Beispiel im Repositorium gespeicherte, standardisierte Metadaten, die wichtige Informationen zu den vorhandenen Forschungsdaten enthalten, über eine OAI-PMH-Schnittstelle ausgeliefert werden. Prinzipiell können auf diese Weise unterschiedlichste Metadatenformate zum Metadata Harvesting, und somit zur Aufnahme von Informationen in einschlägige Kataloge, bereitgestellt werden.

30 http://www.fedora-commons.org

31 http://www.w3.org/RDF/

32 eXtensible Access Control Markup Language, vgl. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

33 Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/

Abb. 5: Webschnittstelle der CLARIN-D Federated Content Search

Im Rahmen von CLARIN-D werden von den beteiligten Zentren, so auch vom HZSK, Metadaten im XML-Format der Component MetaData Infrastructure (CMDI)[34] ausgeliefert. Die Verwendung von CMDI erlaubt hierbei die flexible Spezifizierung von Metadatenkategorien, deren Semantik mit ISOcat[35]-Kategorien festgelegt werden kann, sodass sie interoperabel zu Beschreibungen anderer Sprachressourcen sind.

Im Rahmen der Anbindung des Repositoriums an die CLARIN-D-Infrastruktur werden die vorhandenen CMDI-Metadaten in regelmäßigen Abständen von dem Metadata Harvester des Virtual Language Observatory (VLO)[36] eingesammelt und über eine zentrale Webschnittstelle zusammen mit Informationen zu einer Vielzahl weiterer Ressourcen, die von den anderen CLARIN-D-Zentren und weiteren internationalen Forschungseinrichtungen zur Verfügung gestellt werden, zugänglich gemacht (vgl. Abb. 4).

Der CLARIN-D-Verbund gewährleistet nicht nur die Möglichkeit einer Langzeitarchivierung durch die mögliche Spiegelung von Ressourcen zwischen den verschiedenen Repositorien sowie die Auffindbarkeit der Daten durch Metadata Harvesting und die Benutzerschnittstelle des VLO, er erlaubt mit der Federated Content Search auch eine zentrale bzw. zentrenübergreifende Suche in allen verfügbaren Ressourcen. Die Implementierung eines Search/Retrieval via URL (SRU)[37]-Endpunkts ermöglicht in diesem Zusammenhang die Verarbeitung und Beantwortung von externen Suchanfragen, wie sie bspw. vom CLARIN Aggregator[38], dem Einstiegspunkt in die CLARIN Federated Content Search, an die Repositorien des CLARIN-D-Verbundes gesendet werden können (vgl. folgende Abb. 5).

Zum Zweck einer dauerhaften Kennzeichnung der Ressourcen werden Persistent Identifiers (PIDs) auf Basis des Handle Systems[39] verwendet, sodass eine eindeutige Identifikation und Zitierbarkeit der Quellen dauerhaft gewährleistet ist. Der Zugriff auf die Ressourcen erfolgt in der Regel webbasiert über das oben erwähnte Single Sign-on-Verfahren. Zu diesem Zweck werden Benutzerdaten von teilnehmenden Universitäten und Forschungseinrichtungen (Identity Providers) mit den zugangsbeschränkten Benutzerschnittstellen der Zentren (Service Providers) verknüpft und über das SAML[40]-Protokoll übertragen, sodass der föderierte Zugriff auf Forschungsdaten ermöglicht wird.

Neben der oben genannten Funktionalität der virtuellen Forschungsumgebung am HZSK, die in vielerlei Hinsicht durch die Vernetzung mit der CLARIN-D-Infrastruktur geprägt ist, beinhaltet sie weitere Merkmale bzw. Funktionen, die speziell auf den vorherrschenden Datentyp und die Expertise des Zentrums, nämlich der Verarbeitung gesprochensprachlicher Daten, zu-

34 vgl. Trippel, T./ Witt, A.: Standardizing metadata descriptions of language resources: The Common Metadata Initiative (CMDI), in: JANCSARY, J. (Hrsg.): Proceedings of KONVENS 2012 (SFLR 2012 workshop) Wien 2012, S. 495ff.; Hedeland, H./ Wörner, K.: Experiences and Problems creating a CMDI profile from an existing Metadata Schema, in: Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR (Proceedings of LREC-Workshop 2012) Istanbul 2012.; http://www.clarin.eu/content/component-metadata

35 http://www.isocat.org/

36 http://catalog.clarin.eu/vlo/

37 http://www.loc.gov/standards/sru/

38 http://weblicht.sfs.uni-tuebingen.de/Aggregator/

39 http://handle.net/

40 Security Assertion Markup Language, vgl. https://www.oasis-open.org/committees/security/

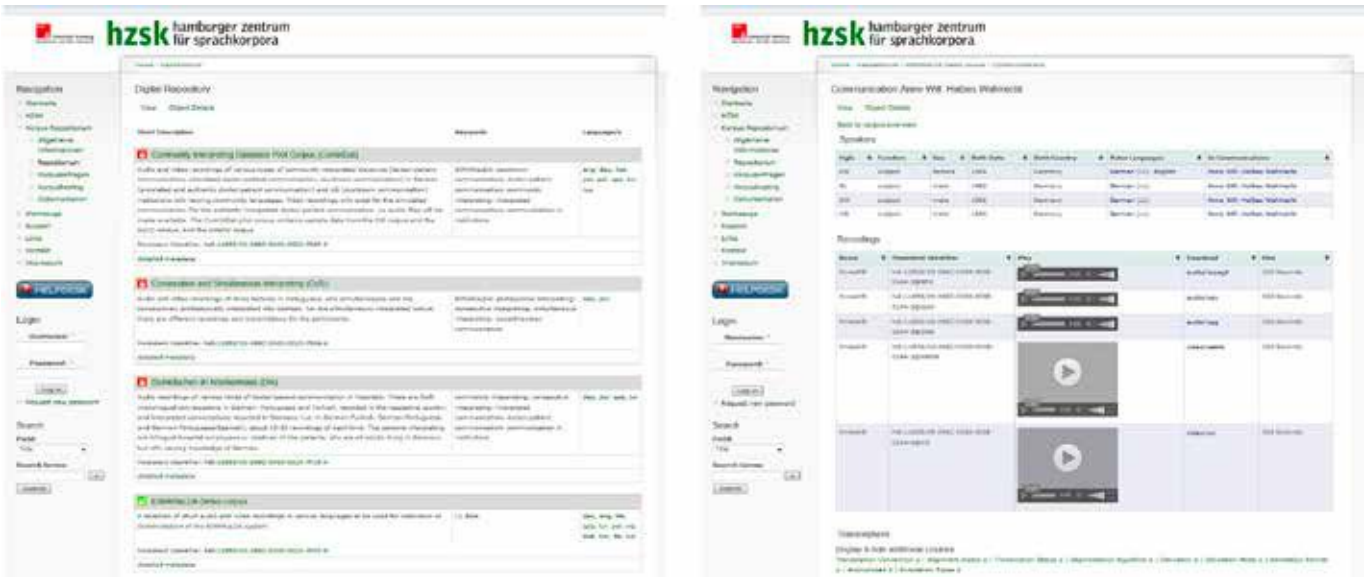

Abb. 6: Webfrontend des Repositoriums des HZSK

geschnitten sind. Die Leitgedanken der CLARIN-D-Infrastruktur beinhalten neben der Vernetzung von Ressourcen und Diensten u.a. die Offenheit für die individuelle Ausrichtung der Zentren, v.a. mit Blick auf die Bedürfnisse der Forschenden in den angesiedelten Forschungsprojekten und der Nutzer. Vor diesem Hintergrund wurde die Erweiterung des Repositoriums am HZSK auf eine webbasierte Architektur ausgerichtet, die es einerseits erlaubt eine Vielzahl von Nutzern mit individuellen Zugriffsrechten effizient zu verwalten, diesen aber andererseits umfangreiche und detaillierte Informationen zu den vorhandenen Ressourcen zur Verfügung stellt. Perspektivisch soll zudem die kollaborative Erstellung, Bearbeitung und Publikation von Sprachressourcen im Repositorium ermöglicht werden. Für diese Zwecke stellt das Software Framework Islandora[41], welches das Fedora System mit dem Content Management System Drupal[42] verbindet, eine geeignete Lösung dar. Islandora ist, wie auch Fedora und Drupal, Open Source Software und zielt auf die kollaborative Verwaltung und die Verfügbarmachung digitaler Daten ab. Es bietet eine interoperable, erweiterbare Grundlage für die Implementierung einer graphischen Benutzerschnittstelle für Fedora-Repositorien. Durch den möglichen Gebrauch bereits existierender und die Entwicklung neuer ‚Islandora Solution Packs' (z.Z. verfügbar sind bspw. Lösungen für Audiodaten, Daten im PDF-Format, Bilder und digitalisierte Bücher) stehen wiederverwendbare Module zur Verfügung, die von verschiedenen Repositorien verwendet werden können.

Die webbasierte, graphische Benutzerschnittstelle des Repositoriums am HZSK ist eine zentrale Komponente der virtuellen Forschungsumgebung und ist über die Webseite des HZSK[43] zugänglich. Durch die Verwendung von Islandora kann die Benutzerverwaltung und die Darstellung einzelner Bestandteile größtenteils aus dem Drupal-Backend heraus gesteuert werden. Die Anzeige der Forschungsdaten und ihrer Metadaten erfolgt durch den Gebrauch von XSLT-Stylesheets, mit deren Hilfe HTML-Ansichten für die spezifischen Inhalte der gespeicherten Ressourcen erzeugt werden können. Diese basieren im Wesentlichen auf den existierenden CMDI-Metadaten, und werden genutzt, um den Benutzern möglichst reichhaltige Informationen über die im Repositorium gespeicherten Ressourcen zur Verfügung zu stellen. Für den Export und die Anzeige von Transkriptionen gesprochener Sprache wurden RESTful Java Webservices implementiert, welche in die Benutzerschnittstelle integriert sind und on-the-fly die gewünschte Ansicht bzw. das gewünschte Exportformat liefern (vgl. Abb. 6). Diese unterstützen wiederum die Weiterverarbeitung der Daten mit verschiedenen Werkzeugen sowie die Auswertung aus unterschiedlichen fachwissenschaftlichen Perspektiven.

Perspektiven

Die hier beschriebene Entwicklung der digitalen Infrastruktur am Hamburger Zentrum für Sprachkorpora hat zu einer signifikanten qualitativen Verbesserung der Bedingungen für lokal angesiedelte Vorhaben an der Universität Hamburg geführt, deren Fokus auf der

41 http://islandora.ca/

42 https://www.drupal.org/

43 https://www.corpora.uni-hamburg.de/repository

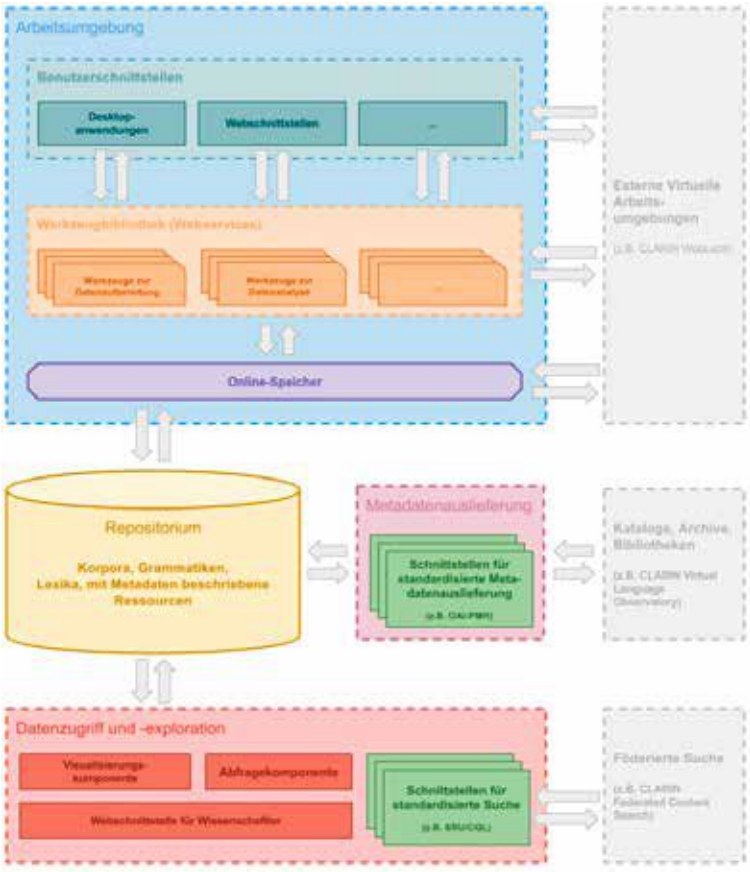

Abb. 7: Weiterentwicklung der virtuellen Forschungsumgebung am HZSK

Erstellung und empirischen Analyse sprachlicher Ressourcen liegt.

Die Anbindung des HZSK an CLARIN als supranationale Forschungsinfrastruktur kann dabei als Meilenstein im Prozess der Entwicklung einer vernetzten virtuellen Forschungsumgebung gewertet werden. Die damit verbundenen Erarbeitung und Umsetzung gemeinsamer technischer und administrativer Standards und Strukturen sind die erfolgreich geschaffenen Rahmenbedingungen für einen Ausbau und die Implementierung weiterer Funktionalitäten.

Das hier vorgestellte Repositorium für Korpora gesprochener Sprache am HZSK, welches zu diesem Zweck entwickelt wurde, setzt bereits einen großen Teil der gewünschten Funktionalitäten um. Damit die vorhandene digitale Infrastruktur hingegen als voll ausgebaute virtuelle Forschungsumgebung genutzt werden kann, in der Forschende sämtliche Aufgaben von der Erstellung, Kuration von Korpora gesprochener Sprache bis hin zur Analyse und Publikation der Forschungsdaten ausführen können, sind noch weitere Arbeitsschritte erforderlich.

Daher soll zukünftig die virtuelle Forschungsumgebung um Schnittstellen für die direkte und kollaborative Arbeit sowie weitere Aufbereitung der im Repositorium vorgehaltenen Daten erweitert werden. Geplant ist die Anbindung von Online-Speichersystemen (wie zum Beispiel ownCloud[44]) sowie die Schaffung einer Werkzeugbibliothek (unter Verwendung von RESTful Webservices) und weiterer webbasierten Benutzerschnittstellen (vgl. Abb. 7).

Bedingung für die Schaffung und den Erhalt digitaler Infrastrukturen ist dabei grundsätzlich eine (vor allen Dingen personelle) Verstetigung von Einrichtungen wie dem HZSK auch bzw. vor allem an Hochschulen. Bezüglich der aktuellsten Entwicklungen an der Universität Hamburg ist in diesem Zusammenhang die Einrichtung einer zentralen Informationsstruktur der Fakultät für Geisteswissenschaften im Rahmen ihrer eHumanities 2020+ Strategie zu nennen, mit der die Fakultät den Leitlinien und Empfehlungen der DFG, des BMBF, des Wissenschaftsrates der HRK und der Enquetekommission folgt und deren Ziel die lokale Verstetigung von Ressourcen sowie Einrichtungen in den Geisteswissenschaften ist.

Es ist mehr als wünschenswert, dass das hier beschriebene Zusammenspiel lokaler und supranationaler Infrastrukturen den Forschenden weltweit den notwendigen Zugang zu Werkzeugen, Diensten und Ressourcen ermöglichen wird. **I**

Autoren
Hanna Hedeland, Daniel Jettka, Timm Lehmberg
Hamburger Zentrum für Sprachkorpora
Max-Brauer-Allee
22765 Hamburg
timm.lehmberg@uni-hamburg.de

44 http://www.gwiss.uni-hamburg.de/de/service/ehumanities/