CLARIN-D und Forschungsinfrastrukturen

Thorsten Trippel, Dieter van Uytvanck

In den Geistes- und Sozialwissenschaften wird im Kontext der Digital Humanities von Forschungsinfrastrukturen gesprochen. Anders als in ingenieurwissenschaftlichen Bereichen oder in naturwissenschaftlichen Labors geht es bei diesen Einrichtungen nicht um die Aufstellung technischer Geräte, sondern um ein komplexes Zusammenspiel von Forschungsdaten, Archiven und Programmen zur Analyse der Daten. Im Bereich der europäischen Forschung im Rahmen des European Strategy Forum on Research Infrastructures (ESFRI) werden diejenigen Einrichtungen als Forschungsinfrakstrukturen bezeichnet, die Ressourcen oder Dienste anbieten, die von Wissenschaftlern zur Durchführung Ihrer Forschung verwendet werden (siehe etwa die ESFRI-Broschüre der Europäischen Kommission, S. 4[1]). Bereits bei der Vorstellung der ESFRI-Roadmap 2006 wurde die Funktion von Forschungsinfrastrukturen folgendermaßen beschrieben:

> *This definition of Research Infrastructures, including the associated human resources, covers major equipment or sets of instruments, as well as knowledge-containing resources such as collections, archives and databases. Research Infrastructures may be "single-sited", "distributed", or "virtual" (the service being provided electronically). They often require structured information systems related to data management, enabling information and communication. These include technology-based infrastructures such as grid, computing, software and middleware.*
> (ESFRI Roadmap, 2006, S. 14[2])

Sprachbasierte Ressourcen, seien es Text-, gesprochene Sprache oder multimodale Daten stellen besondere Anforderungen an eine Infrastruktur. Diese kann nicht nur ein spezialisiertes Archiv für Forschungsdaten sein, sondern sie muss die speziellen Anforderungen, insbesondere rechtliche und ethische Richtlinien, berücksichtigen. Daher ist die CLARIN Infrastruktur selbst Ort und Werkzeug, um Forschung zu ermöglichen und komplexe Suchfunktionen und weitere Funktionen zur Inhaltserschließung anzubieten und Analysen auf den Ressourcen vorzunehmen.

CLARIN-D ist das nationale Projekt, das zur europäischen Forschungsinfrastruktur CLARIN gehört, die als European Research Infrastructure Consortium (ERIC) organisiert ist. Nachdem bislang primär die *Implementation* der grundlegenden Infrastrukturelemente Vordergrund stand und diese in Teilen bereits verfügbar sind, z.B. Repositorien, Vernetzung der Suchfunktionen über Repositorienkataloge hinweg, Dienste zur Erschließung von Ressourcen über Webservices usw., liegt nun der Fokus auf der *Anwendung* in den Geistes- und Sozialwissenschaften.

Für Forschungsinfrastrukturen im Bereich der eHumanities sind Datenrepositorien als Grundlage für wissenschaftliche Untersuchungen unerlässlich. Sie müssen ergänzt werden durch fachspezifische Werkzeuge zur Suche, Analyse und Visualisierung der Daten. In seinem Aufsatz „Controversies around the Digital Humanities: An Agenda"[3] weist Thaller darauf hin, dass die Möglichkeiten der Analyse bisher mit der Zunahme verfügbarer Ressourcen nicht Schritt halten. Die Analyse innerhalb der Infrastruktur von CLARIN begegnet dieser Kritik durch eine skalierbare Services und durch einen Mehrwert durch die Kombination der verteilt eingesetzten Werkzeuge und Ressourcen. CLARIN geht auf die Geistes- und Sozialwissenschaften zu, die sprachbasierte Daten verwenden, um passende Werkzeuge anzubieten und gegebenenfalls anzupassen.

In dem Vortrag stellen wir die Kernkomponenten der Infrastruktur vor und wie Forschende die Infrastruktur für ihre Forschungsfragestellung einsetzen können.

---

[1] http://ec.europa.eu/research/infrastructures/pdf/esfri_brochure_0113.pdf
[2] http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf
[3] http://www.cceh.uni-koeln.de/files/ThallerIntroWahn.pdf

Schritte zur Integration einer Ressource in CLARIN-D

Thomas Eckart, Volker Boehlke, Jörg Knappen

In den Digital Humanities werden kontinuierlich neue Daten und Verfahren erstellt und aufbereitet welche in der jeweiligen Community einen hohen Stellenwert haben, aber auch für Fachanwender aus verschiedenen anderen Bereichen eine hohe Relevanz besitzen. Dabei treten erfahrungsgemäß wiederholt vergleichbare Probleme auf, welche häufig aus Zeitmangel gar nicht oder nur rudimentär adressiert werden. Hierzu zählen:

- die langfristige Archivierung und Wiederverwendbarkeit von Ressourcen und Werkzeugen
- die Bereitstellung von Metadaten und die Sichtbarkeit dieser in einem zentralen Metadatenkatalog
- die Bereitstellung von Methoden welche das sichere Zitieren einer Ressource und deren einzelner Bestandteile ermöglichen
- der feingranulare Zugriff auf Ressourcen mit dem Ziel der Wiederverwendbarkeit
- die Anbindung an ressourcen- und anbieterübergreifende Anwendungen, welche effizientes Arbeiten mit einer Vielzahl digitaler Ressourcen ermöglicht (Arbeit mit verschiedenen zugriffsgeschützten Ressourcen, Speicherung und Austausch umfangreicher Zwischenergebnisse, ...)
- die Bereitstellung von Dokumentationen sowie von Lehr- und Lernmaterialien zu digitalen Ressourcen und Werkzeugen

In diesem Vortrag, dem zweiten der Sektion, wird daher die Integration von Ressourcen in die CLARIN-D Infrastruktur thematisiert. Dabei werden die hierfür obligatorischen und optionalen Schritte zur Einbindung von Daten und Verfahren definiert und detailliert beschrieben. Die Basis einer jeden Integration von Ressourcen in CLARIN-D bildet dabei die Erzeugung geeigneter Metadaten im CMDI-Format und deren Bereitstellung über eine standardisierte Web-Schnittstelle. Auf dieser Basis bauen diverse Anwendungen (wie zum Beispiel der Metadaten-Katalog Virtual Language Observatory VLO) auf. Ein weiterer obligatorischer Bestandteil ist die Sicherstellung der Zitierbarkeit durch die Nutzung von globalen Identifikationsdiensten wie dem Handle System.

Weitere wichtige Bestandteile möglicher Integrationsmaßnahmen umfassen dabei:

- Jedes der neun CLARIN-D Zentren unterhält ein Repositorium für die langfristige Archivierung und Bereitstellung digitaler Ressourcen. Die Repositorien sind mit dem Data Seal of Approval zertifiziert.
- Bereitstellung Webservice-basierter Zugriffsmethoden auf Daten und Ressourcen. Hier wird die Grundlage für den granularen Zugriff auf Daten und Workflows und damit für die Möglichkeit der einfachen, gezielten und effizienten Wiederverwendung einer Ressource in neuen Kontexten gelegt.
- Nutzung der föderierten Authentifikations- und Autorisierungsinfrastruktur CLARIN-D AAI. Diese erlaubt es den Zugriff auf eine Ressource auf bestimmte Nutzer bzw. Nutzergruppen einzuschränken und verknüpft dies mit Methoden, welche Single-Sign-On Funktionalität ermöglichen.
- Die Anbindung an die CLARIN-D Federated Content Search stellt eine attraktive Möglichkeit der inhaltsbasierten Suche auf einer Vielzahl verschiedener textueller Ressourcen gleichzeitig dar.

- Das Problem der Speicherung von Zwischenergebnissen kann mit Hilfe der Anbindung an die CLARIN-D Workspaces gelöst werden.
- Die TeLeMaCo-Sammlung von Lehr- und Lernmaterialien erlaubt ein gezieltes Finden von Kurzanleitungen, Handbüchern sowie kleinen und großen Lerneinheiten.

Gesamtziel des Vortrags ist es den Teilnehmern zu verdeutlichen, welche neuen Möglichkeiten sich durch die oben genannten Formen der Integration bieten, mit welchen Methoden und Werkzeugen dabei gearbeitet werden kann und in welcher Form Dokumentation zu jedem dieser Schritte verfügbar ist. Zudem sollen bekannte Problemfelder, wie die Problematik der geeigneten Granularität der Metadaten und des Zugriffs auf eine Ressource diskutiert und exemplarische Lösungen vorgestellt werden. Die Teilnehmer sollen in die Lage versetzt werden, den für die Integration von Ressourcen notwendigen Aufwand korrekt abschätzen zu können. Es soll zudem verdeutlicht werden, welche Synergien und Potentiale durch eine solche Integration in eine Forschungsinfrastruktur wie CLARIN-D mit relativ geringem Aufwand möglich sind.

Webdienste und WebMAUS

Thomas Kisler

Die Common Language Ressources and Technology Infrastucture, CLARIN, ist eine Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften. Ziel der Infrastruktur ist es, sprachbasierte Ressourcen und Dienste dezentral anzubieten, sowohl für geschriebene als auch gesprochene Sprache.

Ein wichtiger Bestandteil von CLARIN sind Webservices, die es Wissenschaftlern ermöglichen, speziell auf ihre Anwendungsbedürfnisse hin entwickelte Softwarepakete im WWW zu nutzen, ohne diese auf dem eigenen Rechner installieren zu müssen. Ebenso können etablierte und weitverbreitete Softwaretools auf diese Weise um neue, webbasierte Dienste erweitert werden und somit neue Funktionalität anbieten. Ein besonderer Vorteil webbasierter Dienste besteht zudem darin, dass ohne Zutun der Nutzer die jeweils neueste Version der Software verwendet werden kann.

Der Zugriff auf Ressourcen gesprochener Sprache erfordert mindestens eine orthographische Transkription der Audiodateien; für weitere Analysen sind detailliertere Transkriptionen, etwa eine breite phonemische oder gar eine enge phonetische Segmentation und Etikettierung notwendig. Ein orthographisches Transkript kann in der Regel einfach und ohne besondere Kenntnisse erstellt werden, häufig ist es sogar, z.B. im Falle von Korpora gelesener Sprache, bereits verfügbar. Die Erstellung einer phonemischen Transkription oder einer phonetischen Segmentation dagegen erfordert Spezialkenntnisse und ist extrem zeitaufwendig – so kann eine enge phonetische Segmentation bis zu tausendmal so lang dauern wie die Äußerung lang ist.

Das Münchner automatische Segmentationssystem MAUS erstellt aus dem orthographischen Transkript einer Äußerung eine Segmentation auf Phonemebene, d.h. jeder Laut der Äußerung wird mit dem Sprachsignal aliniert. Die Besonderheit von MAUS ist, dass es aus dem orthographischen Transkript aufgrund statistischer Aussprachеregeln Aussprachehypothesen generiert und diese mit dem Sprachsignal abgeglichen werden; MAUS gibt als Ergebnis die wahrscheinlichste Aussprachehypothese zurück und kann damit die typischen Koartikulationsphänomene gesprochener Sprache berücksichtigen.

Im Kontext von CLARIN-D wurden eine grafische, webbasierte Benutzeroberfläche für MAUS sowie Programmierschnittstellen in Form von Webservices zum einfacheren Einbinden der Funktionalität in externe Anwendungsprogramme entwickelt. Diese Webservices sind, wie alle CLARIN-D Ressourcen, in CMDI-konformen Metadaten beschrieben und können somit von geeigneten Anwendungsprogrammen automatisch benutzt werden.

Die grafische Benutzeroberfläche im WWW erlaubt es einem Nutzer, interaktiv Audiodateien und dazugehörige orthographische Transkripte im Browser hoch- und nach Abschluss der Bearbeitung die Segmentationsdaten herunterzuladen.

WebMAUS unterstützt aktuell neun Sprachen und ist ihm Rahmen von CLARIN-D Showcases in bestehende linguistische Workflows integriert.

Motivation einer Sektion zu CLARIN-D und Zusammenfassung der Abstracts

Christoph Draxler

Bayerisches Archiv für Sprachsignale

Institut für Phonetik und Sprachverarbeitung

LMU München

draxler@phonetik.uni-muenchen.de

1. Übersicht

Für die 1. Jahrestagung der DHd planen wir eine Sektion zu CLARIN-D, der deutschen Forschungsinfrastruktur-Initiative für die Geistes- und Sozialwissenschaften mit dem Schwerpunkt auf Sprachressourcen. Diese Sektion besteht aus drei Vorträgen:

1. *CLARIN-D und Forschungsinfrastrukturen*: Thorsten Trippel (Universität Tübingen), Dieter van Uytvanck (Max-Planck Institut für Psycholinguistik, Nijmegen]
2. *Schritte zur Integration einer Ressource in CLARIN-D*: Thomas Eckart, Volker Boehlke (Uni Leipzig), Jörg Knappen (Universität des Saarlands)
3. *Web-Services und WebMAUS*: Thomas Kisler (LMU München)

2. Abstract

Die drei Vorträge bauen aufeinander auf: der erste ordnet CLARIN-D in die europäischen Infrastrukturinitiativen ein, der zweite beschreibt den Umfang der CLARIN-D Ressourcen und das prinzipielle Vorgehen der Integration von externen Ressourcen in die CLARIN-D Infrastruktur, der dritte zeigt exemplarisch einen für CLARIN-D entwickelten Webdienst.

Die europäische CLARIN Infrastruktur besteht aktuell aus acht Ländern sowie einer internationalen Organisation als Mitgliedern, und einem Land als Beobachter. Die deutsche CLARIN-D-Initiative ist dezentral organisiert: neun Zentren decken die Erstellung und Pflege von sowie den Zugriff auf sprachbasierte Ressourcen und Dienste ab. In der aktuellen Implementierungsphase sind in CLARIN-D bereits wichtige Infrastrukturkomponenten erfolgreich implementiert worden, so z.B. das einheitliche Login, die Einrichtung von menschen- und maschinenzugänglichen Repositories an allen CLARIN-D Zentren, Unterstützung sowohl des automatischen Harvesten der Metadaten für den globalen Datenkatalog des Virtual Language Observatory (VLO) als auch eine erste Version der interaktiven verteilten Inhaltssuche in den Datenbeständen der Zentren. Darüberhinaus sind viele Textkorpora und Sprachdatenbanken CLARIN-konform aufbereitet und in die Repositories aufgenommen worden, und Tools für die web-basierte Bearbeitung der Daten entwickelt und implementiert worden.

Im zweiten Vortrag beschreiben wir, wie bislang verteilte und uneinheitlich aufgebaute Text- und Sprachressourcen so aufbereitet werden, dass sie CLARIN-D konform beschrieben und über ein einziges Login genutzt werden können. Dabei ist es zentrales CLARIN-D Anliegen, dass die Daten – inklusive aller Zugangs- und Nutzungsrechte – bei den bisherigen Eigentümern bleiben.

Natürlich besteht auch die Möglichkeit, Datenbestände an ein CLARIN-D Zentrum zu übertragen, z.B. um die dauerhafte Verfügbarkeit und Langzeitarchivierung zu sichern.

In diesem Vortrag werden die wichtigsten Schlüsseltechnologien wie die komponentenbasierte Metadaten-Architektur CMDI, die Verwendung von dauerhaft gültigen Persistent Identifiern, der zentrale Metadatenkatalog der VLO und die verteilte Inhaltssuche im Detail, die Verkettung linguistischer Verarbeitungstools im Web vorgestellt, und der Aufbau einer Dokumentation und Sammlung von Lehr- und Lernmaterialien vorgestellt.

Der dritte Vortrag beschreibt die Entwicklung von CLARIN-D Webdiensten und geht auf den für CLARIN-D implementierten Dienst WebMAUS ein. Konkret geht es um das in der Verarbeitung gesprochener Sprache notwendige, aber sehr zeitaufwendige Etikettieren und Segmentieren von Sprachdateien. Bei dieser Segmentation wird das Sprachsignal mit einer Wort- oder Lautfolge aliniert, so dass Analysen des und Suchen im Sprachsignal über den Inhalt der Äußerung möglich sind. Grundlage der Entwicklung von WebMAUS ist, dass eine einfache orthographische Transkription einer Äußerung relativ einfach und ohne spezielles phonologisches oder phonetisches Wissen möglich ist. Liegt eine solche orthographische Transkription vor, dann können in einem automatischen Verfahren daraus Aussprachehypothesen vorgeschlagen und diese dann mit dem Sprachsignal abgegelichen werden.

WebMAUS ist in die linguistische Verarbeitungskette von CLARIN-D eingebunden und hat sich in kurzer Zeit zu einem vielgenutzten Webdienst entwickelt. Zu den Anwendern zählen nicht nur Phonetiker und Linguisten, sondern zunehmend auch unter anderem die Ethnologie, Dialektologie, Sprachtechnologie und Kommunikationswissenschaften.

3. Vortragende

Thorsten Trippel: Seminar für Sprachwissenschaft, Universität Tübingen; Liaison zur europäischen CLARIN Forschungsinfrastruktur

Dieter van Uytvanck: Language Archives, Max-Planck Institut für Psycholinguistik, Nijmegen; Leiter der technischen Infrastruktur in CLARIN-D

Volker Boehlke: Abteilung Automatische Sprachverarbeitung, Uni Leipzig; Betreuung der CLARIN-D Facharbeitsgruppen

Thomas Eckart: Abteilung Automatische Sprachverarbeitung, Uni Leipzig; Repositories und Inhaltssuche

Jörg Knappen: Englische Sprach- und Übersetzungswissenschaft, Universität des Saarlands; Repositories und Inhaltssuche

Thomas Kisler: Bayerisches Archiv für Sprachsignale, LMU München; WebDienste