



Nummer 6, 2014, Mai

PID: 11858/00-1779-0000-0023-9537-F

Editorial

Sechster CLARIN-D-Newsletter

„Alles neu macht der Mai“ (kommt laut CLARIN-D Federated Content Search in zwei Ressourcen vor) – ein abgedroschener Spruch, dennoch nicht ganz ohne Bezug zu CLARIN-D.

Denn erstens haben wir in den vergangenen Monaten mit großem Einsatz darauf hingearbeitet, dass im Mai 2014 endlich die Bewilligung für die nächste CLARIN-Phase kommen kann, zweitens hat der Newsletter mit Steffi Pletzer eine neue Co-Herausgeberin, drittens findet im Mai die LREC in Reykjavik mit einem eigenen CLARIN-D Tutorial zu „Online Speech and Language Resources“ statt, und viertens kommen im Mai die ersten Anmeldungen für die Joint ESU-CLARIN Summer School in Leipzig herein.

In diesem Newsletter haben wir gleich zwei Repository-Beschreibungen: Alex Geyken gibt eine Übersicht des CLARIN-Servicezentrums Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften, und Peter Fankhauser gibt einen ausführlichen Einblick in das Repository am IDS. Diese beiden Repositories sind für CLARIN-D von herausragender Bedeutung, weil sie a) bereits bestehende Ressourcen in neuartiger Weise zugänglich machen, und b) ein standardisiertes Verfahren zur Aufnahme neuer Ressourcen zur Verfügung stellen.

Zwei Workshops mit CLARIN-D Relevanz zeigen, dass der Disseminations-Gedanke ernst genommen wird. Im Workshop „Annotation: Anwenderbedarf und Support“ des Hamburger Zentrums für Sprachkorpora ging es darum, den Kontakt zwischen CLARIN-D externen Anwendern und den Tool-Entwicklern in CLARIN-D herzustellen, um die Bedürfnisse der Anwender und die Imple-

mentationen der Entwickler aufeinander abzustimmen. „Gesprochene Sprache: von der Aufnahme zur Publikation“ war ein Workshop des BAS, der sich explizit an Studenten und junge Wissenschaftler/innen richtete und praxisnah Sprachaufnahmen, ihre Annotation und anschließende Aufnahme in das BAS Repository vermittelt hat. Die Resonanz sowohl von Vortragenden als auch den Teilnehmern war bei beiden Workshops außerordentlich positiv.

Das große CLARIN-D Ereignis des Jahres 2014 wird die gemeinsame Sommerschule der European Summer University und CLARIN-D vom 22.07.-02.08.2014 in Leipzig. In diesem Newsletter gibt es dazu eine Übersicht der geplanten Kurse und Termine – merken Sie sich den Termin, geben Sie ihn an Ihre Mitar-

beiter weiter und laden Sie ausgewählte Studenten zur Teilnahme ein, denn eine solch große Bandbreite an Themen in derart kompakter Form finden Sie woanders kaum!

Nun noch ein großes Dankeschön an alle Autoren und viel Vergnügen beim Lesen dieser Ausgabe des CLARIN-D-Newsletters!



Christoph Draxler & Steffi Pletzer

V. i. S. d. P./Impressum:

Christoph Draxler
Ludwig-Maximilians-Universität München
Institut für Phonetik und Sprachverarbeitung
Schellingstr. 3
80799 München

Telefon: +49 (0) 89 / 2180 - 2807
E-Mail: draxler@phonetik.uni-muenchen.de

Für die Inhalte der Artikel sind die jeweiligen Autoren verantwortlich

Webpräsenz des europäischen Langzeitprojekts:

www.clarin.eu

Ein CLARIN-Center stellt sich vor: Das CLARIN-Servicezen- trum des Zentrum Sprache der BBAW

Ein Kurzbericht über die Sprach- datensammlung

Das Repository des CLARIN-Servicezentrum des Zentrum Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) [1] dient der Langzeitarchivierung von sprachwissenschaftlichen Primärdaten und konzentriert sich dabei vorwiegend auf historische Textkorpora und lexikalische Ressourcen.

Es wurde im Juni 2013 als „CLARIN Centre B“ zertifiziert [2]. Die Zertifizierung durch das „Data Seal of Approval“ [3] umfasst insbesondere die Aspekte der Organisation, des Workflows, der Qualitätssicherung und Nachhaltigkeit der Daten.

[1] <http://clarin.bbaw.de/>

[2] Siehe das Zertifikat unter PID: <http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-93>
(Wie alle URLs in diesem Dokument zuletzt abgerufen am 12. Mai 2014).

[3] Siehe dazu <http://www.datasealofapproval.org/>

[4] <http://www.django-de.org/>

[5] <https://github.com/emory-libraries/eulfedora>

[6] URL zur Schnittstelle des CLARIN-Repositorys der BBAW:
<http://clarin.bbaw.de:8088/oaiprovider/?verb=Identify>

Sämtliche Metadaten im Repository sind entsprechend dem CLARIN-spezifischen CMDI-Profil kodiert. Die einzelnen Datensätze werden mit persistenten Identifizierern (PIDs) versehen, wodurch eine langfristige Referenzierbarkeit ermöglicht wird. Jede neue Version eines Datensatzes wird mit einer eigenen PID versehen, während die jeweils älteren Versionen über ihre ursprüngliche PID verfügbar bleiben.

Das Repository basiert, wie die meisten anderen deutschen CLARIN-Repositories, auf der Software Fedora Commons. Die Weboberfläche wurde mithilfe des Python-Web-Frameworks Django [4] bzw. der Eulfedora-Komponente [5] implementiert, die über eine gegenüber der reinen Fedora Commons Software intuitivere Suchmaske und Abfragesyntax verfügt. Für den automatisierten Zu-

griff besitzt das Repository eine OAI-PMH-Schnittstelle[6]

Die derzeit im Repository verfügbaren historischen Textkorpora stammen überwiegend aus dem DFG-Projekt „Deutsches Textarchiv“[7] der BBAW. Die Volltexte des DTA sind anhand des auf XML/TEI P5-basierenden DTA-Basisformats (DTABf) annotiert, welches als Best-Practice Modell für geschriebene Korpora in CLARIN-D fungiert. Derzeit sind 1300 im Rahmen des DTA digitalisierte Werke im Repository verzeichnet, etwa 700 weitere sind in Arbeit[8]. Seit kurzem ist auch der vollständige ‚Dingler‘, also sämtliche 370 Bände des von Johann Gottfried Dingler begründeten Polytechnischen Journals (Erscheinungszeitraum 1820–1931) im Repository zu finden[9]. Die Volltexte aus dem ‚Dingler‘-Projekt wurden ebenfalls entsprechend des DTABf annotiert und in die Korpusinfrastruktur des DTA integriert. Seit kurzem ist auch das ‚Berliner Wendekorpus‘ aus Transkripten von narrativen Interviews, die zwischen 1992 und 1996 mit Ost- und Westberlinern über deren persönliche ‚Wende‘-Erfahrungen geführt wurden, im Repository verfügbar.

Bei den im BBAW-Repository verwalteten lexikalischen Ressourcen handelt es sich um Wörterbücher aus dem Bestand des „Digitalen Wörterbuchs

der Deutschen Sprache“ (DWDS[10]). Dazu gehören unter anderem die gegenwartssprachliche Wörterbuchkomponente des DWDS, das Etymologische Wörterbuch des Deutschen nach Wolfgang Pfeifer und die Erstbearbeitung des von Jacob Grimm und Wilhelm Grimm begründeten Deutschen Wörterbuchs (¹DWB, 1854–1960). Über das Repository kann beispielsweise auf Stichwortlisten, angereichert mit grammatischen Informationen, aus den verschiedenen Wörterbüchern des DWDS zurückgegriffen werden. Hinzu kommt eine lexikalische Datenbank, die im dlexDB-Projekt [11] erarbeitet wurde. Diese enthält Informationen über die Auftretenshäufigkeit von Wörtern und deren Kategorien, von Wortsequenzen sowie von sublexikalischen Einheiten wie z.B. Silben für den Einsatz in der psychologischen und linguistischen Forschung.

Es ist auch für externe, d. h. nicht an der BBAW arbeitende Wissenschaftler möglich, Daten zur Langzeitarchivierung im Repository zu deponieren. Voraussetzung dafür ist die inhaltliche Übereinstimmung der Daten zu den primären Aufgaben des BBAW-Repositorys sowie eine Konvertierung der Daten in die im Repository verwendeten Formate, und drittens müssen die Urheberrechte für eine Aufnahme in das Repository geklärt sein.

[6] <http://clarin.bbaw.de:8088/oai/provider/?verb=Identify>

[7] DTA, <http://www.deutschestextarchiv.de>

[8] Das DTA bietet darüber hinaus unter <http://www.deutschestextarchiv.de/download/> die Texte seines Kernkorpus sowie des bislang veröffentlichten Teils des Ergänzungskorpus zum Download an. Mit einem Zeitstempel versehen, können dort die gesamten Textdaten oder nach Genre (Belletristik – Gebrauchsliteratur – Wissenschaft) bzw. Erscheinungszeitraum zusammengestellte Teilkorpora als ZIP-Dateien heruntergeladen werden.

[9] Weitere Informationen unter <http://www.polytechnischesjournal.de/>

[10] <http://www.dwds.de>

[11] <http://www.dlexdb.de/>



DAS CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBAW

Author:

Urbanitzky

Title:

Elektricität

Text class:

Bibliographical information:

Date:

between

example: 1800-01-01 or 1800-01-01 - 2013-12-31

Results per page:

10

Results source:

Deutsches Textarchiv

Result fields to display:

- Pid
- Title
- Source
- Publisher
- Author
- Type
- Date
- Text class

Search [Simple search](#)

1 total objects

Number	Author	Title	Type	Bibliographical information	Date	Publisher	Text class
dta:1216	Urbanitzky, Alfred von	Die Elektrizität im Dienste der Menschheit – Eine populäre Darstellung der magnetischen und elektrischen Naturkräfte und ihrer praktischen Anwendungen. Nach dem gegenwärtigen Standpunkte der Wissenschaft (vollständige digitalisierte Ausgabe)	Text	Urbanitzky, Alfred von: Die Elektrizität im Dienste der Menschheit. Wien; Leipzig, 1885. [Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, SBB-PK, Op 29650<a>]	1885-01-01	Deutsches Textarchiv	Gebrauchsliteratur. Populärwissenschaft

[↑ ZUM SEITENANFANG](#)



DAS CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBW

Bibliographic information:

Title:	Die Elektrizität im Dienste der Menschheit – Eine populäre Darstellung der magnetischen und elektrischen Naturkräfte und ihrer praktischen Anwendungen. Nach dem gegenwärtigen Standpunkte der Wissenschaft (vollständige digitalisierte Ausgabe)
Author:	Urbanitzky, Alfred von
Date:	1885-01-01
Language:	de
Text class:	Gebrauchsliteratur: Populärwissenschaft
Rights:	Creative Commons Attribution-NonCommercial 3.0 Unported License
Source:	Urbanitzky, Alfred von: Die Elektrizität im Dienste der Menschheit. Wien; Leipzig, 1885. [Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, SBB-PK, Op 29650<a>]
Link to book:	link
Version date:	Jan 21, 2014
DC:	view content
CMDI:	view content (Download: http://hdl.handle.net/11858/00-203C-0000-0023-3823-D)
OAI_DC:	view content
XML:	view content (Download: http://hdl.handle.net/11858/00-203C-0000-0023-2C9C-4)
RELS-EXT:	view content

▼ Versions DC:

2014-01-21 17:02:41	link
2014-01-20 21:57:13	link

▶ Versions CMDI:
 ▶ Versions OAI-DC:
 ▶ Versions XML:
 ▶ Versions RELS-EXT:

Uploaded at Jan 20, 2014; last modified Jan 21, 2014 (3 months, 1 week ago).

Das Repositorium des Instituts für deutsche Sprache in Mannheim

Die Kernaufgabe des Instituts für Deutsche Sprache (IDS) ist die Erforschung und Dokumentation der deutschen Sprache. Dazu sammelt und archiviert das IDS einen umfangreichen Bestand an Forschungsprimärdaten in Form von Korpora der geschriebenen und gesprochenen Sprache sowie Sekundärdaten, wie zum Beispiel lexikographische Ressourcen.

Das IDS-Repositorium[1] hat zum Ziel, sowohl die am IDS erhobenen Forschungsprimärdaten, als auch Daten von externen Datengebern im Bereich der Germanistik nachhaltig zur archivieren und innerhalb geltender gesetzlicher Rahmenbedingungen allgemein verfügbar zu machen.

Abbildung 1 zeigt schematisch die Rolle

des Repositoriums im Kontext der Ressourcen innerhalb und außerhalb des IDS. Die internen Ressourcen umfassen Korpora der geschriebenen Sprache (Deutsches Referenzkorpus, DeReKo), der gesprochenen Sprache (Archiv für Gesprochenes Deutsch, AGD) sowie lexikographische Ressourcen. Für die Aufbereitung und Nutzung dieser Ressourcen existieren jeweils eigene Aufbereitungsprozesse und Systeme, die ständig weiterentwickelt werden: Für geschriebene Sprache sind das derzeit das Korpus-Suche-, -Management- und -Analyse-System COSMAS II, sowie die Aufbereitungsprozesse im Bereich DeReKo, für gesprochene Sprache die Datenbank Gesprochenes Deutsch (DGD) und für lexikographische Ressourcen das Online-Wortschatz-Informationssystem Deutsch (OWID).

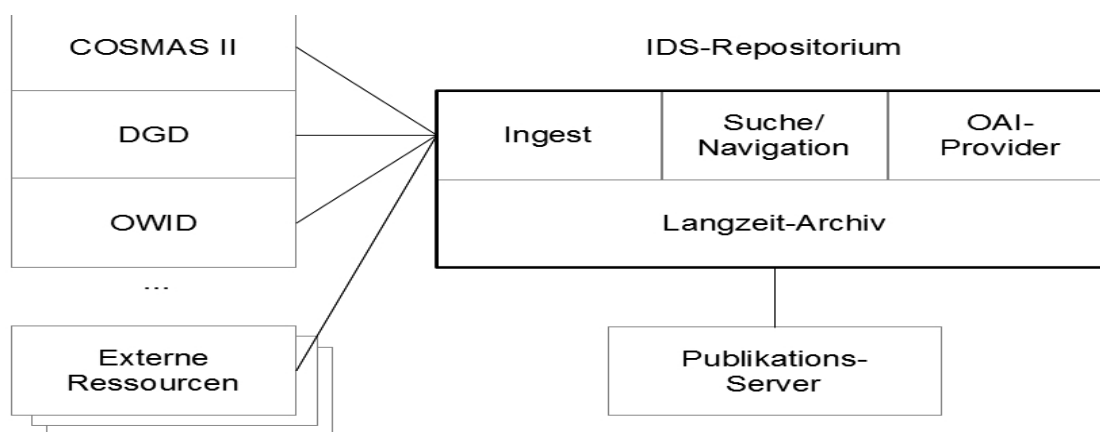


Abbildung 1: IDS-Repositorium im Kontext

[1] <http://repos.ids-mannheim.de/>

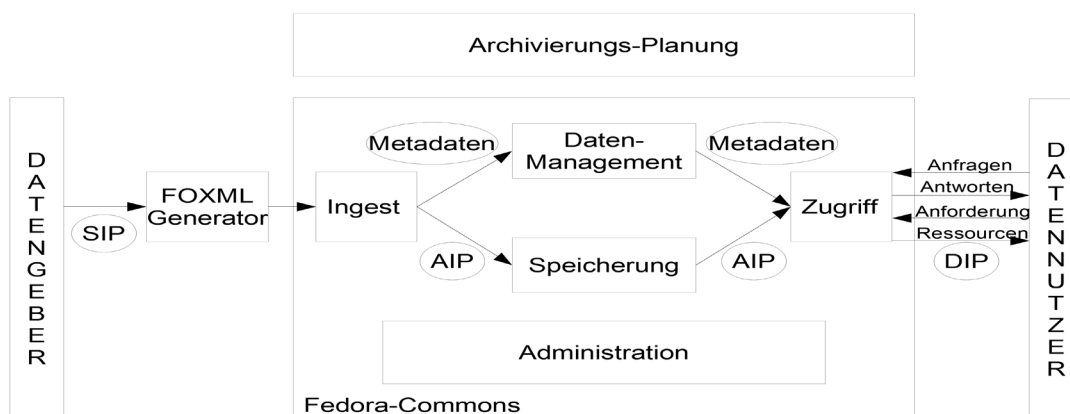


Abbildung 2: Funktionale Architektur des IDS-Repositoryms

Funktionale Architektur

Abbildung 2 zeigt die funktionale Architektur des IDS-Repositoryms anhand des OAIS-Referenzmodells. Die wesentlichen Komponenten wurden auf Basis von Fedora-Commons realisiert.

„Archive Information Packages“ (AIPs) verknüpfen Primärdaten mit deskriptiven und technischen Metadaten. Sie werden als digitale Objekte im Fedora-XML-Format (FOXML) serialisiert und gespeichert. Eine Ressource besteht typischerweise aus mehreren miteinander verknüpften digitalen Objekten. Jedes digitale Objekt besteht wiederum aus mehreren sogenannten Datenströmen für Daten und Metadaten in unterschiedlichen Formaten. Für alle Ressourcen wird eine Metadatenbeschreibung in CMDI und eine daraus abgeleitete Dublin-Core-Beschreibung abgelegt.

Die Einspeisung (Ingest) von „Submission Information Packages“ (SIPs) wird von einem einfachen Webinterface sowie einem REST-API und Skripten für den Ingest von mehreren digitalen Objekten (Batch-Ingest) unterstützt. Das IDS-Repositorym verwendet typischerweise Batch-Ingest auf Basis von SIPs, die von

einem für die jeweilige Ressource angepassten Verarbeitungsprozess (FOXML-Generator) erzeugt werden.

Für die Suche unterstützt Fedora-Commons ein REST-API, das von einer einfachen formularbasierten Webschnittstelle angesprochen wird. Die Metadaten werden einerseits maschinenlesbar über eine OAI-PMH-konforme Schnittstelle (OAI-PMH) und andererseits in einer für Webbrowser geeigneten Darstellung zur Verfügung gestellt. Nutzer haben direkten Zugriff auf die archivierten digitalen Objekte, wenn die geeignete Zugangsautorisierung besteht; einige Ressourcen und alle Metadaten werden ohne Zugangsbeschränkung zur Verfügung gestellt. Diese „Dissemination Information Packages“ (DIPs) bestehen entweder aus einzelnen Datenströmen oder einem gepackten Archivformat für die gesamte Ressource.

Modellierungsprinzipien

Sprachressourcen sind sehr heterogen in Bezug auf verwendete Datenmodelle für Metadaten und Primärdaten, Medienformate, Speicherungssysteme sowie Umfang und Nutzungsform. Diese

Heterogenität hat teilweise historische Gründe, ist aber auch auf die spezifischen Anforderungen und Bedingungen zurückzuführen, unter denen die Ressourcen zusammengestellt wurden.

Ziel der Langzeitarchivierung ist es einerseits, die Ressourcen in ihrem aktuellen Nutzungskontext langfristig verfügbar zu machen, und andererseits, die Ressource auch in neuen Kontexten für eine Nachnutzung aufzubereiten. Dabei gelten folgende Prinzipien:

1. **Erhaltung der Originalformate:** Die Originaldaten werden in jedem Fall in ihrem Originalformat mit möglichst geringen Anpassungen (z.B. Kodierung in UTF-8) archiviert. Damit soll zumindest ihre aktuelle Nutzungsform gesichert sein.
2. **Ergänzung der Originalformate um nachhaltige Formate:** Falls ein Originalformat nicht einem der für die nachhaltige Archivierung und Nutzung empfohlenen Formate entspricht, werden die Originaldaten zusätzlich in geeignete Formate konvertiert und archiviert.
3. **Erhaltung und Formalisierung der Metadaten:** Alle vorhandenen Metadaten werden verlustfrei in ein formales CMDI-Profil übersetzt. Dieses Vorgehen beinhaltet die formale Abbildung der einzelnen Metadatenfelder in geeignete ISOcat-Kategorien.
4. **Ergänzung von fehlenden Metadaten:** Insbesondere für die Gesamtressource werden fehlende Kernmetadaten der Dublin-Core-Initiative (Ersteller, Titel, Beschreibung, Typ, Sprache) aus den Daten extrahiert oder manuell ergänzt.
5. **Persistente Identifikation:** Alle Ressourcen und ihre Datenströme erhalten einen sogenannten Persistenten Identifikator (PID), sodass sie auch bei einem Umzug des Repositoriums zu einer anderen Adresse unveränderlich referenzierbar bleiben.

Weitere Infos auch im Wiki:

<http://de.clarin.eu/mwiki>

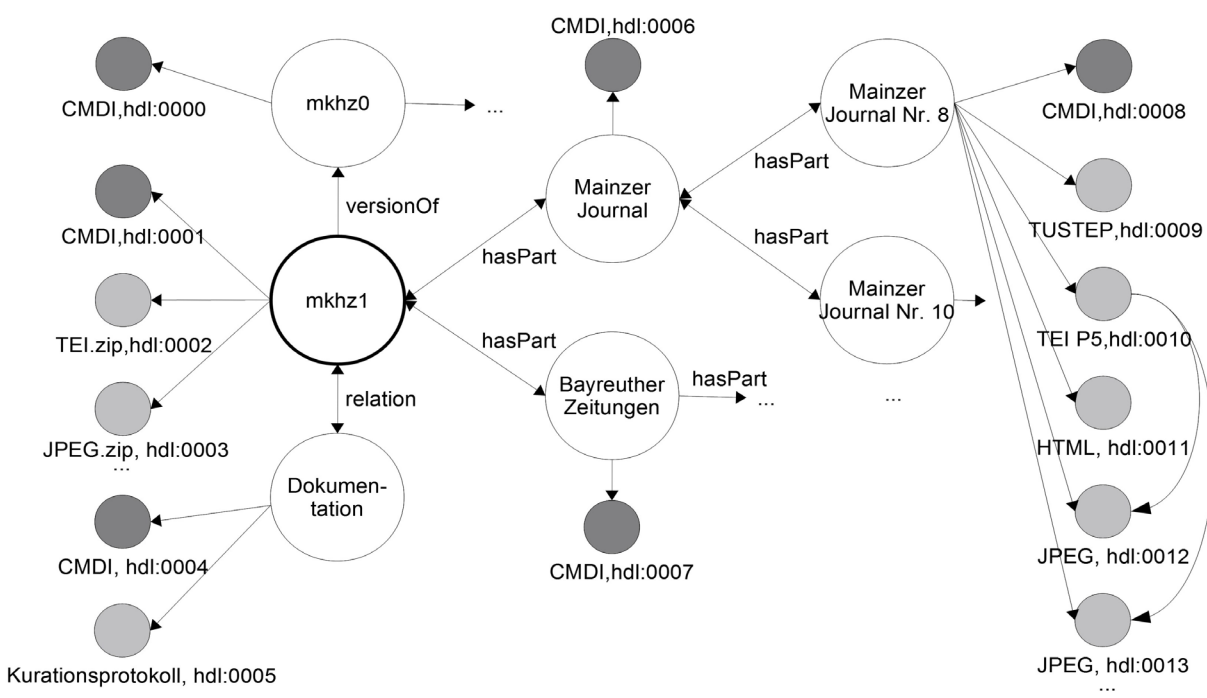


Abbildung 3 Objektmodell

Objektmodell

Digitale Sprach-Ressourcen bestehen typischerweise aus mehreren Teilen. Abbildung 3 zeigt am Beispiel des Mannheimer Korpus Historischer Zeitungen und Zeitschriften (mkhz)[2] das dem Repositoryum zugrunde liegende Objektmodell.

Das Korpus (mkhz1) besteht aus mehreren Zeitungen (Mainzer Journal, Bayreuther Zeitungen etc.), die ihrerseits wieder aus mehreren Ausgaben bestehen. Diese Beziehungen werden als hasPart repräsentiert. Für jedes dieser Objekte werden Metadaten im CMDI-Standard angelegt, die Dublin-Core-Metadaten repräsentieren und dem Objekt einen persistenten Identifikator zuordnen - in der Abbildung exemplarisch dargestellt mit hdl:nummer. Die eigentlichen Daten werden als sogenannte Datenströme repräsentiert und ebenfalls mit einem persistenten Identifikator ausgestattet. So enthält das Objekt „Mainzer Jour-

nal Nr. 8“ die einzelnen Druckseiten in hochaufgelöstem JPEG-Format, die ursprüngliche Transkription im TUSTEP-Format, die daraus generierte Transkription im TEI P5-Format und eine zum Schmökern geeignete Version in HTML. Die Transkriptionen verweisen ihrerseits wieder auf die ihnen zugrundeliegenden Druckseiten unter Verwendung derer persistenten Identifikatoren. Damit wird eine enge Verknüpfung zwischen den Primärdaten (Druckseiten) und den daraus abgeleiteten Sekundärdaten (Transkription) erreicht. Die einzelnen Formate (TUSTEP, TEI, JPEG) werden zusätzlich als komprimiertes Archiv (.zip) abgelegt und als Datenströme der Gesamtressource verfügbar gemacht. Das Objekt Dokumentation dient zur Archivierung der Aufbereitungsprozesse und die Relation versionOf setzt das aktuelle Korpus in Bezug mit einer früheren Aufbereitung (mkhz0).

[2] <http://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content>

Die anderen Korpora und Ressourcen des IDS sind ähnlich modelliert, unterscheiden sich aber in den für das jeweilige Korpus adäquaten Metadaten, der gewählten Granularität und den zugrundeliegenden Formaten. Die Prinzipien der Repräsentation von Sprach-Ressourcen für die Langzeitarchivierung – geeignete Granularität zur persistenten Identifizierbarkeit, Metadatenvollständigkeit zur Auffindbarkeit, Formatvollständigkeit zur nachhaltigen Nutzung sowie Dokumentation der Aufbereitung und aktuellen Nutzung zur Nachvollziehbarkeit – gelten jedoch für alle archivierten Ressourcen.

Metadatenmodellierung

Anders als zum Beispiel bibliographische Daten weisen Sprachressourcen eine hohe Heterogenität in ihren Metadaten auf. Korpora der geschriebenen Sprache erfordern andere Metadaten als Korpora der gesprochenen Sprache oder lexikographische Ressourcen.

Um dieser Heterogenität Rechnung zu tragen verwendet das IDS-Repositoryum

die in CLARIN entwickelte Metadaten-Infrastruktur CMDI. CMDI ermöglicht es, beliebige Metadatenschemata auf Basis von wiederverwendbaren Komponenten zu spezifizieren. Die Semantik der Metadaten-Felder wird dabei über eine obligatorische Zuordnung zu einer ISocat-Kategorie[3] eindeutig spezifiziert. Durch die Wiederverwendung von Komponenten und die explizite Spezifikation der Semantik können so die spezifischen Metadaten-Anforderungen einer Sprachressource berücksichtigt, aber auch die Interoperabilität zwischen verschiedenen Metadaten-Schemata aufrechterhalten werden.

Für jede Ressource wird ein Minimum von Dublin-Core-Metadaten erhoben. Konkrete Beispiele der Metadaten sind im IDS-Repositoryum einzusehen.

Ingestprozesse

Die Aufbereitung von Ressourcen erfordert häufig ressourcenspezifische Prozesse, die sich jedoch möglichst aus wiederverwendbaren und konfigurierbaren Komponenten zusammensetzen.

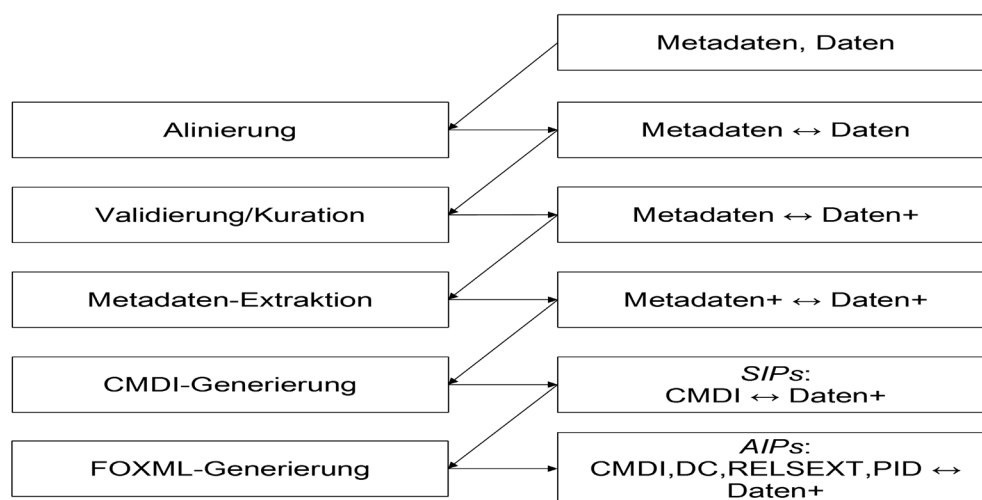


Abbildung 4: Allgemeiner Aufbereitungsprozess von Sprachressourcen

[3] <http://www.isocat.org>

Abbildung 4 zeigt eine verallgemeinerte Darstellung der wesentlichen Aufbereitungsschritte.

Im ersten Schritt (Alinierung) werden Metadaten und Daten in einen expliziten Bezug miteinander gesetzt. Metadaten werden häufig in Form von Tabellen, deren Zellen durch Kommata voneinander getrennt sind, oder durch ad hoc XML-Strukturen mit Referenzen auf die eigentlichen Daten angegeben. Dieser Schritt stellt sicher, dass für jede Ressource Metadaten eindeutig vorhanden sind. Er erfordert häufig eine Normalisierung von Referenzen und Dateinamen.

Im zweiten Schritt (Validierung/Kuration) werden Datenformate validiert, und – falls erforderlich – nicht nachhaltig nutzbare Datenformate in eines der zur Langzeitarchivierung empfohlenen Datenformate konvertiert. Typischerweise werden dabei sowohl XML-basierte Formate (zum Beispiel auf Basis von TEI P5) als auch PDF/A-Versionen erstellt. Nicht valide Dateien bzw. Du-

plikate werden in Abstimmung mit den Datengebern behandelt. Dieser Schritt zielt auf eine nachhaltige Nutzbarkeit der Daten. Abhängig vom Datenformat kann dieser Schritt recht aufwändig sein.

Im dritten Schritt (Metadaten-Extraktion) werden fehlende Metadaten aus der XML-Repräsentation der Daten extrahiert bzw. manuell hinzugefügt. Dieser Schritt nutzt die XML-Aufbereitung des vorherigen Schritts, um mithilfe von standardisierten XML-Werkzeugen weitere Metadaten zu erzeugen.

Im vierten Schritt (CMDI-Generierung) werden die Metadaten in ein geeignetes Profil der CMDI-Metadaten-Infrastruktur transformiert. Falls kein geeignetes Profil verfügbar ist, wird ein bestehendes Profil um entsprechende Metadatenfelder erweitert. Mit der Zuordnung von Metadatenfeldern zu formalen ISOcat-Kategorien zielt dieser Schritt auf eine formale Interpretierbarkeit der Metadaten. Die Aufbereitungsschritte zielen jeweils auf einen Aspekt der Daten- bzw. Metadatenqualität, bauen modular auf-

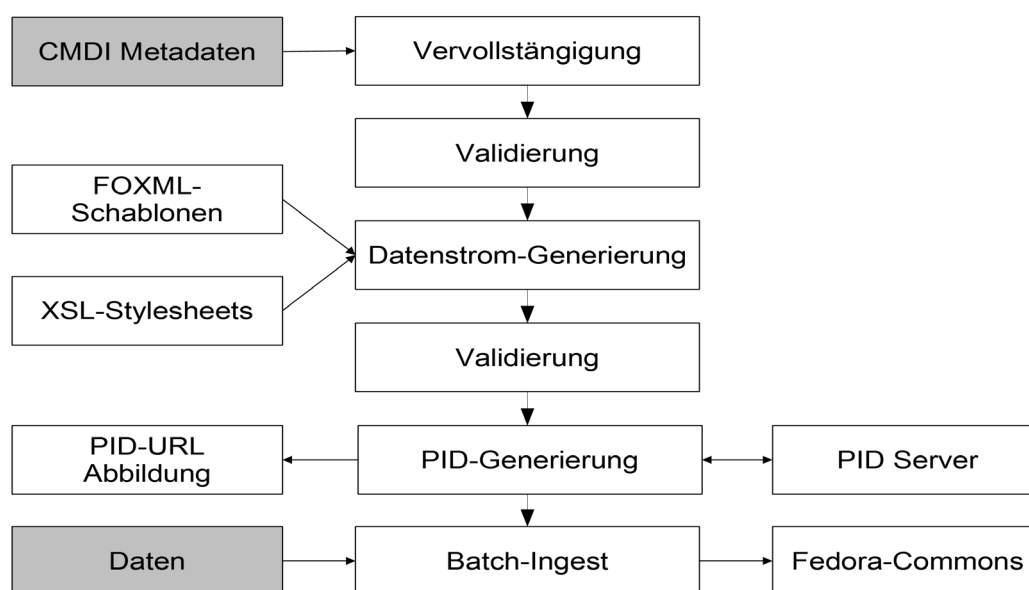


Abbildung 5: Generierung von FOXML

einander auf und werden umfassend dokumentiert. Das Ergebnis der Aufbereitung ist ein „Submission Information Package“ (SIP), das aus mehreren digitalen Objekten (Daten) zusammen mit ihren CMDI-Metadaten besteht.

Im letzten Schritt werden diese SIPs für den Ingest in Fedora-Commons aufbereitet. Dieser Schritt ist als eine einfach konfigurierbare Verarbeitungskette realisiert und in Abbildung 5 schematisch dargestellt.

Zunächst werden CMDI-Metadaten mit fehlender Information ergänzt und gegen ihr Profil, das als XML Schema repräsentiert ist, validiert. Daraufhin werden für jedes digitale Objekt, bestehend aus CMDI-Metadaten und Daten, FOXML-Datenströme generiert. Diese Generierung wird von FOXML-Schablonen gesteuert, die weitere technische Metadaten für Fedora-Commons bereitstellen, wie zum Beispiel Mime-Typ, lokale Identifikatoren und Kontrollgruppe. Zudem werden die von Fedora-Commons vorgegebenen Metadatenströme für Dublin Core und die OAI-PMH-Schnittstelle aus den CMDI-Metadaten mit Hilfe von XSLT-

Stylesheets generiert (OAI-PMH). Diese Datenströme werden ebenfalls validiert.

Für alle Datenströme wird ein persistenter Identifikator (PID) registriert, sowie alle lokalen Identifikatoren im SIP mit dem zugehörigen PID ersetzt. Die Abbildung zwischen PIDs und lokalen Identifikatoren wird ebenfalls im Repository gespeichert. Die so generierten digitalen FOXML-Objekte werden in einem Batch-Prozess in Fedora-Commons eingespeist.



Peter Fankhauser

Webpräsenz des europäischen Langzeitprojekts:

www.clarin.eu

CLARIN-D Disseminationsworkshop in München

„Sprachdatenbanken – von der Aufnahme zur Publikation“

Das CLARIN-D Zentrum, Bayerisches Archiv für Sprachsignale (Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München), hat als Teil seiner Aktivitäten in Arbeitspaket 9 „Dissemination“ einen Workshop für fortgeschrittene Studenten, Doktorande und PostDocs veranstaltet.

Stattgefunden hat dieser Workshop am Mittwoch, den 31.03.2014 im Internationalen Begegnungszentrum München IBZ (Amalienstr. 38, 80799 München). Zum Thema „Sprachdatenbanken – von der Aufnahme bis zur Publikation“ kamen 29 Teilnehmer überwiegend aus Deutschland, aber auch von der österreichischen Akademie der Wissenschaften in Wien sowie der Universität Zürich. Der Workshop war sehr praxisorientiert aufgebaut: die Teilnehmer hatten vor dem Workshop bereits die notwendige Software und Aufnahmeskripte installiert, so dass nach einer kurzen Einführung von Christoph Draxler in das Thema Sprachaufnahmen die praktische Arbeit beginnen konnte. Jeweils in Gruppen wurden die Rechner konfiguriert, Sprecherdaten angelegt und Aufnahmesitzungen aufgezeichnet. Die Aufnahmen

wurden anschließend auf einen Server übertragen, dort um eine Text-Annotation ergänzt und waren damit bereit für den nächsten Verarbeitungsschritt.



Thomas Kisler führte in WebMAUS ein. WebMAUS ist ein frei verfügbarer Web Service zur automatischen Segmentierung und Etikettierung von gesprochener Sprache. WebMAUS kann, geeignete Signale und eine orthographische Rohtranskription der Äußerung vorausgesetzt, eine qualitativ hochwertige phonetische Segmentierung erstellen. Die Teilnehmer im Kurs konnten WebMAUS sofort mit den von Ihnen aufgenommenen Daten ausprobieren.



Am Nachmittag standen die Erstellung von Metadaten sowie der Aufbau eines Repository im Vordergrund. Bernhard Jackl stellte das von ihm entwickelte Skript COALA vor, mit dem man aus einfachen Texttabellen automatisch CMDI-kompatible Metadaten generieren kann; wenn Ersteller von Sprachdatenbanken diese Daten gleich in einem passenden Format speichern, dann können sie ohne Mehraufwand automatisch korrekt formatierte und CLARIN-kompatible Metadaten erzeugen.

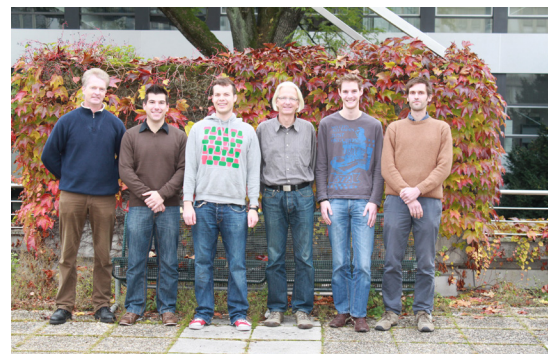


Aus diesen Metadaten erzeugt Uwe Reichel mit einem perl Skript ein Repository für CLARIN-D. In diesem Repository sind alle Ressourcen per Persistent Identifier adressiert, das Repository wird jede Nacht vom Harvester des Virtual Language Observatory und anderen Suchmaschinen indiziert, und sein Inhalt kann über die Federated Content Search in CLARIN durchsucht werden.



Den Abschluss des Workshops bildete eine Fragestunde, in der die Teilnehmer ihre aktuellen Projekte vorstellen und Fragen dazu stellen konnten. Die Diskussion war lebhaft.

Sowohl die Teilnehmer als auch die Dozenten haben vom Workshop profitiert: das BAS hat Rückmeldung zur Bedienerfreundlichkeit und Nützlichkeit seiner Tools erhalten, die Teilnehmer haben einen kompakten und sicherlich auch recht anstrengenden Einblick in den Prozess der Erstellung einer Sprachdatenbank erhalten.



v.l: Florian Schiel, Thomas Kisler, Fabian Bross, Christoph Draxler, Bernhard Jackl, Uwe Reichel

CLARIN-D Workshop des Hamburger Zentrums für Sprachkorpora

Bericht zum Workshop „Annotation: Anwender- bedarf und Support“ der HZSK

Die Idee des Workshops, der am 28. und 29. November 2013 stattfand, war es, den Kontakt zwischen CLARIN-D-externen Anwendern und den Entwicklern von Werkzeugen aus dem CLARIN-D-Umfeld über den reinen Helpdesk hinaus zu intensivieren und einen Rahmen zu schaffen, in dem beide Communities einander ihre Arbeit in Form von Vorträgen und Systemdemonstrationen vorstellen.

Basierend auf den Erfahrungen des HZSK bei der Konzeption und Durchführung des CLARIN-D-Helpdesk war dem Workshop die Erkenntnis vorausgegangen, dass in den Anwendercommunities oft sehr spezifische und komplexe Anforderungen an Dienste und Werkzeuge zur Annotation linguistischer Daten gestellt werden. Um diese noch besser bedienen zu können, sollte durch den Workshop

sowohl auf Seiten der Entwickler das Verständnis für die individuellen Anwenderbedürfnisse, und auf der Anwenderseite das Wissen um das Potential der CLARIN-D-Infrastruktur und natürlich den Helpdesk verbessert werden.

Zu diesem Zweck gliederte sich das Programm in die drei Themenblöcke

- Annotation von Korpora gesprochener Sprache und Lernerkorpora,
- Werkzeuge und Standards und
- Historische Korpora und Nicht-Standard-Varietäten.

Themenblock 1: Korpora Gesprochene Sprache und Lernerkorpora

Thomas Schmidt (Institut für deutsche Sprache) eröffnete den Workshop mit seinem Beitrag „Token Annotation in FOLK“, in dessen Rahmen er die (semi-)automatischen Verfahren der orthographischen Normalisierung, Lemmatisierung und des Part-Of-Speech-Taggings vorstellte, die bei der Aufbereitung des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)[1] zum Einsatz kommen. Es wurden zudem Perspektiven einer Einbindung dieser Verfahren

[1] <http://agd.ids-mannheim.de/folk.shtml>

als Webservice in die CLARIN-D Infrastruktur sowie die notwendige Erweiterung der WebLicht-Umgebung zum Zweck einer Verarbeitung von Transkriptionen gesprochener Sprache diskutiert. Anknüpfend hieran stellten Cordula Meißner und Daisy Lange, Mitarbeiterinnen des Projektes GeWiss – Gesprochene Wissenschaftssprache (Universität Leipzig)[2] die im Rahmen des Projektes eingesetzten Verfahren der Annotation von Sprachwechseln und Diskurskommentierungen vor. Karoline Kühl (Universität Kopenhagen) diskutierte unter dem Titel „Neue Sprachen findet man nicht nur im Urwald“ am Beispiel des Written Faroe Danish Corpus (WriFD) Methoden zur Feststellung, inwiefern eine empirische Ressource neue Varietäten adäquat widerspiegelt. Im Vordergrund der Betrachtungen standen dabei unter anderem Fragen der syntaktischen sowie morphosyntaktischen Annotation mehrsprachiger Korpora. Melanie Andresen, Mitarbeiterin an der Schreibwerkstatt Mehrsprachigkeit (Universität Hamburg)[3] gab Einblicke in die heterogenen schriftlichen und mündlichen Daten, die im Rahmen der Beratungsarbeit des Projektes erhoben wurden und eröffnete Perspektiven für die Generierung eines hochattraktiven Korpus hieraus.

Auch bei dem anschließend von Hagen Hirschmann (HU Berlin) vorgestellten Falko Korpus[4], einem fehlerannotierten Lernerkorpus des Deutschen als

Fremdsprache, standen Fragen der Annotation und Aufbereitung von Lernerdaten im Vordergrund. Anders als bei den vorangegangenen Ansätzen lag der Focus hier jedoch auf der gleichzeitigen Analyse und Annotation grammatischer (normgerechter) oder ungrammatischer (normabweichender) Strukturen, zu deren Zweck für das Falko-Korpus eigens eine auf der Formulierung von Zielhypothesen basierende Mehrebenenarchitektur entwickelt wurde.

Themenblock 2: Werkzeuge und Standards

Die im ersten Block vielfach zu Tage getretenen methodologischen Fragen der Mehrebenenannotation und Interpretativität von Annotation wurden von Wolfgang Menzel (Universität Hamburg) in seinem Beitrag mit dem Titel „Korpusannotation und Meinungsvielfalt“ aus der Perspektive der automatischen Sprachverarbeitung aufgegriffen. Im Zentrum stand eine kritische Auseinandersetzung mit dem Begriff des Gold Standard Annotation vor dem Hintergrund individueller Interpretationen von sprachlichen Phänomenen. Ausgehend von dem Beitrag wurden allgemeine Anforderungen an Standards und Benutzeroberflächen zur linguistischen Annotation diskutiert. Den Versuch einer Umsetzung dieser Anforderungen in Form eines Werkzeugs zur manuellen Annotation präsentierte Said Yimam (TU Darmstadt) aus dem Entwicklerkreis von WebAnno[5], einem flexiblen webbasierten Werkzeug zur Mehrebenenan-

[2] <https://gewiss.uni-leipzig.de>

[3] <http://www.universitaetskolleg.uni-hamburg.de/de/projekte/tp05>

[4] <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko>

[5] <https://code.google.com/p/webanno/>

notation aus dem CLARIN-D Umfeld. Direkt an diesen Beitrag knüpfte Daniel Jettka (HZSK - Uni Hamburg) mit seiner Demonstration von WebLicht (Web-Based Linguistic Chaining Tool) [6] als integraler Teil der CLARIN-D Infrastruktur an. Diskutiert wurden im Anschluss verschiedene Anwendungsszenarien für beide Werkzeuge in Forschung und Lehre sowie die Nutzung beider Werkzeuge an individuelle Forschungsfragen.

Themenblock 3: Historische Korpora und Nicht-Standard-Varietäten

Zu Beginn des dritten Blocks stellte Fabian Barteld (Universität Hamburg) die in dem DFG-Projekt „Entwicklung der satzinternen Großschreibung im Deutschen“ eingesetzten Verfahren zur syntaktischen und semantischen Annotation frühneuhochdeutscher Hexenverhörprotokolle vor. Im Anschluss gaben Sarah Ihden und Timm Lehmberg (Universität Hamburg) in ihrem Beitrag Grenzfälle der Annotation historischer Daten einen Überblick über den Arbeitsstand des DFG Projektes ReN – Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650). Der Schwerpunkt lag dabei auf Verfahren der automatischen Annotation von Nicht-Standard-Daten mithilfe einer für diesen Zweck geschaffenen Variante des STTS Tagsets. Direkt daran knüpfte Sarah Kwekkeboom (Universität Bochum) mit einer Diskussion von Verfahren der manuellen Vorannotation (bzw. Präeditierung)

historischer Texte für sprachstufenübergreifende Referenzkorpora an. Dies erfolgte am Beispiel der mit ReN (s.o.) assoziierten Korpora ReM (Referenzkorpus Mittelhochdeutsch) und ReF (Referenzkorpus Frühneuhochdeutsch). Christian Brockmann und Vito Lorusso (Universität Hamburg – Sonderforschungsbereich 950 Manuskriptkulturen in Asien, Afrika und Europa)[7] führten anschließend in ihrem Beitrag Philosophisches und naturwissenschaftliches Wissen in griechischen Manuskripten des Kardinals Bessarion (1403-1472) am Beispiel von Randnotizen, Kommentierungen, Edierungen und Subskriptionen in historischen Handschriften einen von den digitalen Methoden abweichenden Annotationsbegriff ein.

Abgeschlossen wurde der dritte und letzte Block mit einem Beitrag von Christina Vertan (Universität Hamburg) die unter dem Titel „Annotation multilingualer historischer Dokumente“, einen Entwurf für Annotationsverfahren zum Zweck einer historischen und ethnologischen Auswertung von Daten aus dem osteuropäischen und osmanischen Raum für ein sich in der Beantragung befindliches Projekt vorstellte.

Zusammenfassung

Trotz stark variierender Daten, Zielsetzungen und Bearbeitungsstände stellte sich heraus, dass bei allen vorgestellten Projekten vergleichbare methodologische Probleme auftraten, die nach Auffassung der Anwesenden eine wichtige

[6] <http://weblicht.sfs.uni-tuebingen.de/>

[7] <http://www.manuscript-cultures.uni-hamburg.de/>

Rolle für einen Support und die Integration in CLARIN-D sowie die Entwicklung von Werkzeugen und Standards spielen. Diese waren vornehmlich die allgemeine Problematik der Theoriebindung existierender Datenformate, Annotations schemata (bzw. -Vokabularien) und Standards sowie die besonderen Anforderungen, die aus der Aufbereitung von Nicht-Standarddaten resultieren. Die überwiegende Mehrheit der Vortragenden von der „Anwenderseite“ formulierte den Wunsch nach mehr Flexibilität und Skalierbarkeit von Werkzeugen und Standards, um diese für ihre individuellen Forschungsansätze nutzen zu können. So wurde in Bezug auf zu annotierenden (Text-)Segmente beispielsweise die Möglichkeit eine Loslösung vom auf Normorthographie basierenden Token- und Satzbegriff gewünscht, ebenso wurden die Restriktionen in Annotationswerkzeugen in Hinblick auf Definition und Korrelierbarkeit von Annotations-

ebenen von den Anwendern kritisiert. Hervorgehoben wurde ebenfalls der Bedarf an intuitiv bedienbaren Oberflächen für Annotations- und Analysewerkzeuge, die zudem nachhaltig gepflegt und in Hinblick auf die Bedürfnisse der Anwendercommunity weiterentwickelt werden müssen.

Die erfreulich lebhaften und konstruktiven Diskussionen in allen drei Blöcken können jedoch vor allen Dingen als Indiz für die Wichtigkeit von gemeinsamen Workshops verschiedener Teildisziplinen, vor allen Dingen mit unterschiedlich starker technischer (bzw. technologischer) Ausrichtung, gewertet werden. Das HZSK möchte sich bei allen Referenten und Besuchern für den gelungenen Workshop bedanken.

hzsk hamburger zentrum
für sprachkorpora

Mitmachen!

Liebe Leser des CLARIN-D-Newsletters, wenn ihr Ideen für einen kurzen Beitrag zu diesem Newsletter habt oder dringend einen Gedanken loswerden wollt, schickt euren kurzen Artikel samt Bild an newsletter@phonetik.uni-muenchen.de

Hinweise zur Beitragsgestaltung findet ihr im Wiki.

Ankündigung des Hamburger Zentrums für Sprachkorpora

hzsk hamburger zentrum für sprachkorpora

Die an der Universität Hamburg, am Fachbereich Informatik, Arbeitsbereich Natürlichsprachliche Systeme - NATS, unter Leitung von Prof. Dr.-Ing. Wolfgang Menzel erstellte Hamburg Dependency Treebank (HDT) wird zur Zeit in die Bestände des Hamburger Zentrums für Sprachkorpora integriert und über CLARIN-D zugänglich gemacht.

Bei der HDT handelt es sich um eine große Baumbank des Deutschen, die mehr als 250.000 Sätze umfasst, welche mit Abhängigkeitsstrukturen, zu etwa 80%

manuell, annotiert wurden. Die verwendeten Texte stammen vom Nachrichtenticker Heise [1] (heise.de) aus den Jahren 1996 bis 2001. Im Zuge der Integration in die CLARIN-Infrastruktur wird eine Konvertierung der Daten in das TCF-Format angeboten. CLARIN-Nutzer können die Daten auf diese Weise in WebLicht weiter verarbeiten und visualisieren lassen. Des Weiteren werden die Daten Interessenten aus der akademischen Gemeinschaft über die webbasierte Abfrage und Visualisierung im HZSK Repository zugänglich gemacht.

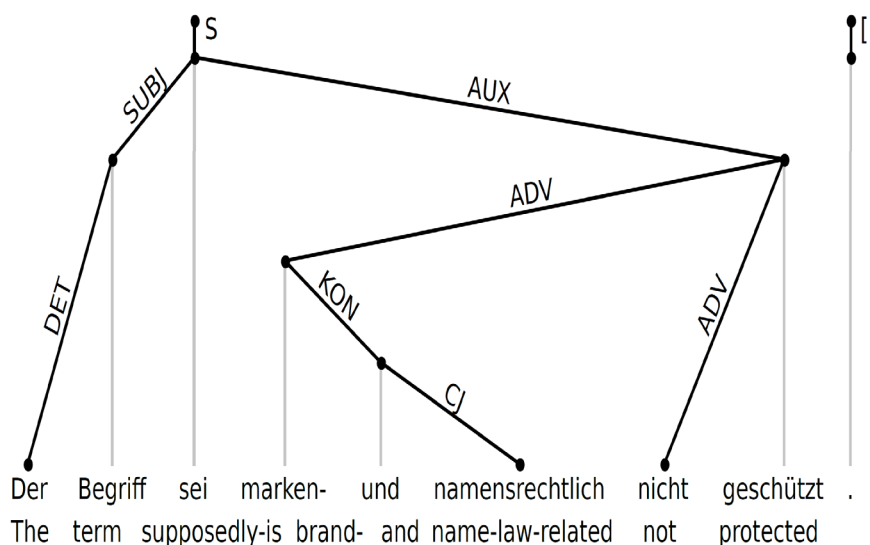


Abbildung 1 : hdt-parse

[1] <http://www.heise.de>

Europäische Sommerschule “Digital Humanities und Sprachressourcen”

Das CLARIN-D Ereignis des Jahres

ESU DH C & T - CLARIN-D

Die gemeinsame Sommerschule der European Summer University und CLARIN-D vom 22.07.-01.08.2014 in Leipzig.

http://www.culintec.uni-leipzig.de/ESU_C_T/

Die Sommerschule will einen Platz schaffen, nicht nur zum Lernen und Weiterbilden, sondern sie will Doktoranden, junge Forscher und Wissenschaftler aus den Bereichen der Geistes-, Bibliotheks- & Ingenieurwissenschaften und der Informatik zusammenbringen. Das Ziel ist es, einen interdisziplinären Erfahrungs- und Wissensaustausch zwischen gleichrangigen Partnern in einem multilingualen und multikulturellen Kontext zu schaffen. Die Sommerschule schafft so die Basis für zukünftige gemeinsame Projekte und Kooperationen und fördert auf besondere Weise die Vernetzung über Disziplin-, Länder-, oder Kulturgrenzen hinaus.

Ziele

Die Sommerschule bietet eine anregende Umgebung zum Diskutieren, zum Lernen und zur Verbreitung von Wissen und Kenntnissen der Methoden und Technologien im Bereich der computergestützten Geisteswissenschaften. Sie bietet ein Forum für Fragen, Antworten und Diskussionen über die Auswirkungen und Implikationen der Anwendung von computergestützten Methoden und Tools auf Kulturgüter. Sie gibt darüber hinaus Einblick in die Komplexität geisteswissenschaftlicher Daten und in die Herausforderung, die die Verbindung der Geisteswissenschaften mit der Informatik und den Ingenieurwissenschaften mit sich bringt.

Programm

Die Sommerschule dauert 11 Tage. Das intensive Programm besteht aus Workshops, öffentlichen Vorlesungen, Projektpräsentationen, einer Postersession und einer Diskussionsrunde. Folgende Workshop-Themen sind geplant:

- XML-TEI Kodierung, Strukturierung und Präsentation

- Suchabfragen in Textkorpora
- Vergleichen von Korpora
- Historische Textkorpora für die Geistes- und Sozialwissenschaften. Digitalisierung, Annotation, Qualitätsmanagement und Analysen
- Open Greek und Open Latin
- Fortgeschrittene Anwendungen für die Geisteswissenschaften in Python
- Stilometrie: computergestützte Analyse literarischer Texte
- Herausgeben im Digitalen Zeitalter: historische Texte und Dokumente
- Raum – Zeit – Objekt: Digitale Methoden in der Archäologie
- Gesprochene Sprache – Aufnahme, Annotation, Analyse
- Multimodale Korpora: wie erstellt, wie nutzt man sie?
- Planung und Management großer Projekte
- Digital Humanities für Lehrstuhlinhaber und Dekane

Jeder Workshop besteht aus 16 Sitzungen in 32 Wochenstunden. Die Teilnehmerzahl ist auf 12 pro Workshop begrenzt.

Bewerbung

Informationen zur Bewerbung auf die Workshopplätze finden Sie unter:

http://www.culingtec.uni-leipzig.de/ESU_C_T/.

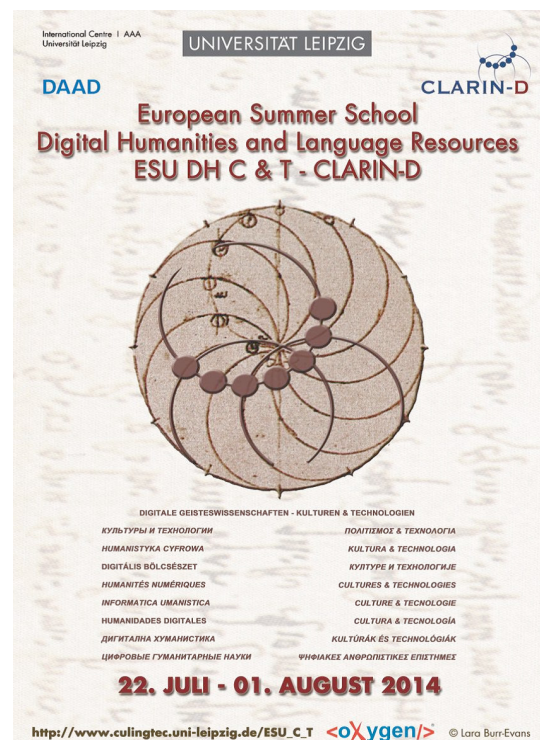
Bei der Platzvergabe werden junge Forscher in den Geisteswissenschaften, die ein technologie-basiertes Forschungsprojekt planen oder bereits in einem solchen Projekt mitarbeiten vorrangig behandelt.

Angehende Ingenieure und Informatiker beschreiben ihr Spezialgebiet und ihre Forschungsinteressen in allgemeinverständlicher Weise, und formulieren ihre Erwartungen an die Sommerschule.

Die Bewerbungen werden laufend entgegengenommen. Die Auswahl der Teilnehmer erfolgt durch den wissenschaftlichen Beirat der Sommerschule und die Leiter der Workshops.

Weitere Informationen

http://www.culingtec.uni-leipzig.de/ESU_C_T/



Abkürzungsverzeichnis (NELCA)

AAI	Authentication and Authorization Infrastructure
ABaC:us	Austrian Baroque Corpus
AEDit	Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit
ALLEA	ALL European Academies
AP	Arbeitspaket
AsiCa	Atlante Sintattico della Calabria
BAS	Bayerisches Archiv für Sprachsignale (München)
BBAW	Berlin-Brandenburgische Akademie der Wissenschaften
BMBF	Bundesministerium für Bildung und Forschung
CiNaViz	City Name Visualization
CMDI	Component MetaData Infrastructure
CLARIN	Common Language Resources and Technology Infrastructure
DAI	Deutsches Archäologisches Institut
DAITF	Data Access and Interoperability Task Force
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DDDTs	DeutschDiachronDigital-Tagset
DH	Digital Humanities
DSA	Data Seal of Approval
DTA	Deutsches Textarchiv
DTAQ	DTA-Qualitätssicherung
ELAN	EUDICO Linguistic Annotator
eAQUA	Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum for Research Infrastructure
EUDICO	European Distributed Corpora Project
EXMARaLDA	Extensible Markup Language for Discourse Annotation
F-AG	Fachspezifische Arbeitsgruppen
FCS	Federated Content Search
FI-Initiative	Forschungs-Infrastruktur-Initiative
FOLK	Forschungs- und Lehrkorpus gesprochenes Deutsch
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
GeWiss	Gesprochene Wissenschaftssprache kontrastiv
GIS	Geographisches Informationssystem
HAB	Herzog August Bibliothek Wolfenbüttel
HPC	High Performance Cluster
HZSK	Hamburger Zentrum für Sprachkorpora
IAIS	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme

ICRI	International Conference on Research Infrastructures
IDS	Institut für Deutsche Sprache (Mannheim)
IETF	Internet Expert Task Force
IMDI	ISLE Meta Data Initiative
IMS	Institut für Maschinelle Sprachverarbeitung (Stuttgart)
InfAI e.V.	Gemeinnütziger Verein des Instituts für Angewandte Informatik in Leipzig
IWiST	Informationswissenschaft und Sprachtechnologie (Hildesheim)
KiDko	KiezDeutsch-Korpus
LiS	Literatur- und Informationsversorgungssysteme
MPI	Max-Planck-Institut
NELCA	Never-Ending-List der CLARIN-Abkürzungen
NMMoCap-Korpus	Natural Media Motion Capture-Korpus
NSF	National Science Foundation (USA)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PID	Persistent Identifier
SaGA	Speech and Gesture Alignment Corpus
SHARE	Survey of Health, Ageing and Retirement in Europe
SuUB	Bremen Staats- und Universitätsbibliothek Bremen
STTS	Stuttgart-Tübingen Tagsets
TCF	Text Corpus Format
TeLeMaCo	Teaching and Learning Materials Collection
TLA	The Language Archive
TüNDRA	Tübingen aNnotated Data Retrieval Application
TüPP-D/Z	Tübinger Partiiell Geparstes Korpus des Deutschen/Zeitungskorpus
TUSTEP	Tübinger System von Textverarbeitungs-Programmen
VLC	Virtual Linguistic Campus
VLO	Virtual Language Observatory
WADL	Web Application Description Language