EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

SFB 833: Bedeutungskonstitution - Dynamik und Adaptivität sprachlicher Strukturen

Forschungsinfrastrukturen: Verfügbarkeit von Daten und deren Langzeitarchivierung

– Erfahrungen und Verfahren in Tübingen –

Prof. Dr. Erhard Hinrichs
Computerlinguistik
Seminar für Sprachwissenschaft
SFB 833: Bedeutungskonstitution

Projekte zur Forschungsinfrastruktur

D

CLARIN

SFB 833: INF

e SciDoc PROJECT
BW-eSci(T)

Verbundprojekt: Hamburg, Potsdam/Berlin, Tübingen
SFB 632: Informationsstruktur

SFB 441

Sonderforschungsbereich Mehrsprachigkeit

Nachhaltigkeit Linguistischer Daten

Wissenschaftsrat zu Forschungsprimärdaten

„Der Wissenschaftsrat empfiehlt den Trägereinrichtungen die umfassende und langfristige Archivierung qualitätsgesicherter und für die jeweilige wissenschaftliche Gemeinschaft langfristig relevanter Daten“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.57f

Freie Verfügbarkeit von Forschungsprimärdaten

„Forschungsprimärdaten bilden einen wertvollen Fundus an Informationen, die mit hohem finanziellem Aufwand erhoben werden. Je nach Fachgebiet und Methode sind sie replizierbar oder basieren auf nicht wiederholbaren Beobachtungen oder Messungen. **In jedem Fall sollten die erhobenen Daten nach Abschluss der Forschungen öffentlich zugänglich und frei verfügbar sein.**“

Aus: *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009

Wissenschaftsrat zur Plagiatsverhinderung

„Der Wissenschaftsrat begreift die umfassende öffentliche Sammlung und Bereitstellung von Forschungsprimärdaten auch als ein probates Mittel der Qualitätssicherung in der wissenschaftlichen Praxis, welches hilft, wissenschaftlichen Betrug und Plagiate leichter zu identifizieren, da die Herkunft von Forschungsdaten aus Repositorien in jedem Falle offen gelegt und die „Autoren“ der Daten zitiert werden müssen.“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.59

Konsequenzen aus der Verfügbarkeit

„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.“

Aus:
Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, DFG: Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, S. 2, Januar 2009

Zentrale Archivierung und Datenmanagement

- Datenmanagement von:
  - Forschungsprimärdaten
  - Metadaten
  - Verknüpfung von Forschungsprimärdaten und Publikationen
- Kooperation mit Rechenzentren, Bibliotheken, Archiven
  - Sicherung durch fortlaufende Migration auf aktuelle Hardware
  - Ausfallrisiken minimieren
  - Gewährleistung von Datenmanagement über das Projektende hinaus

Kooperation mit den Sondersammelgebieten der DFG

- Sondersammelgebiete an gegenwärtig 21 Universalbibliotheken und 12 Spezialbibliotheken

- dienen vor allem der überregionalen Versorgung der Geistes-, Rechts- und Sozialwissenschaften sowie einiger Fächer der Wirtschafts- und Naturwissenschaften

Forschungsprimärdaten

- Experimentaldaten
  - Erhebungen
  - Fragebögen
  - Versuche
- „Quellen“
- Annotationen
- Rohanalysen
- Unpublizierte Projektberichte
- ...

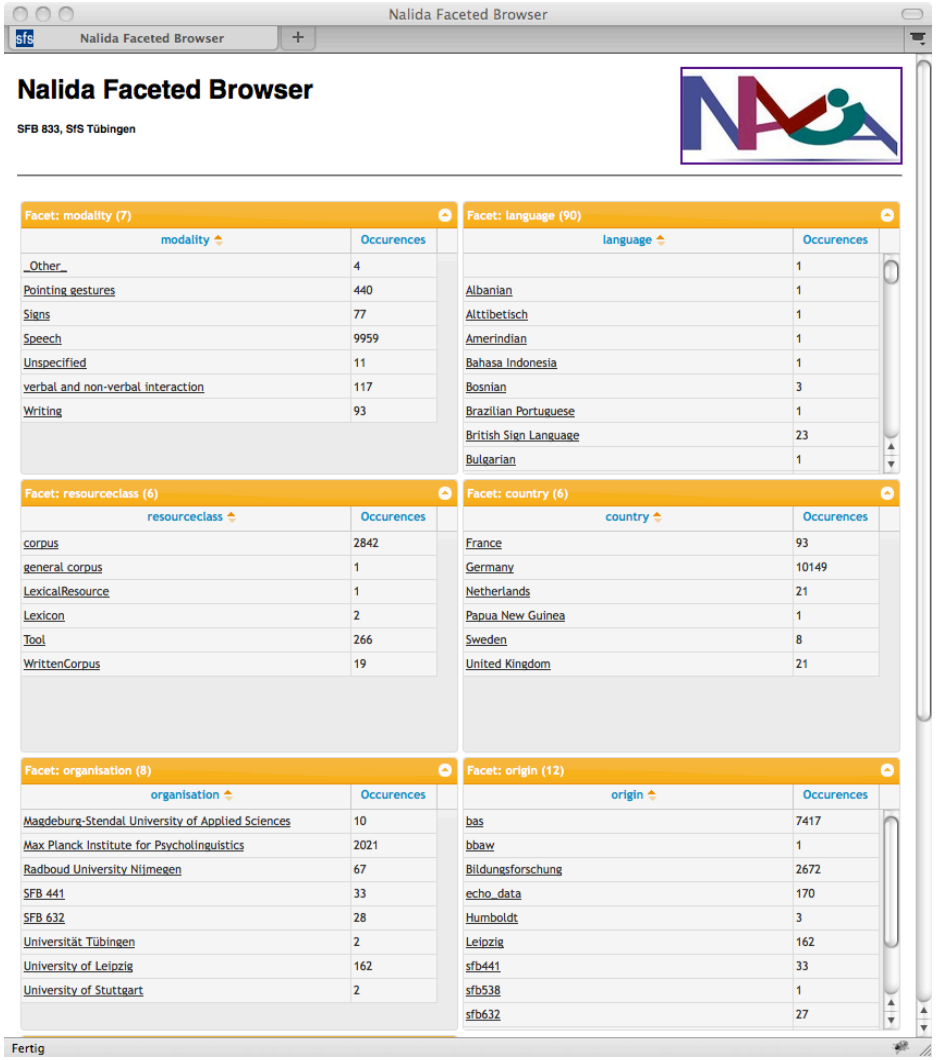
Metadaten: Dokumentation von Primärdaten

- Alle Informationen für:
  - (wieder) finden von Daten
  - (gefundene) Daten verstehen
    - Autorenschaft
    - Relevanz für die eigene Forschung
    - Format und Erstellungsdatum
- In einem spezifischen Format
  - Übersichtlich
  - Standardisiert
  - Durchsuchbar
- Maschinell verarbeitbar

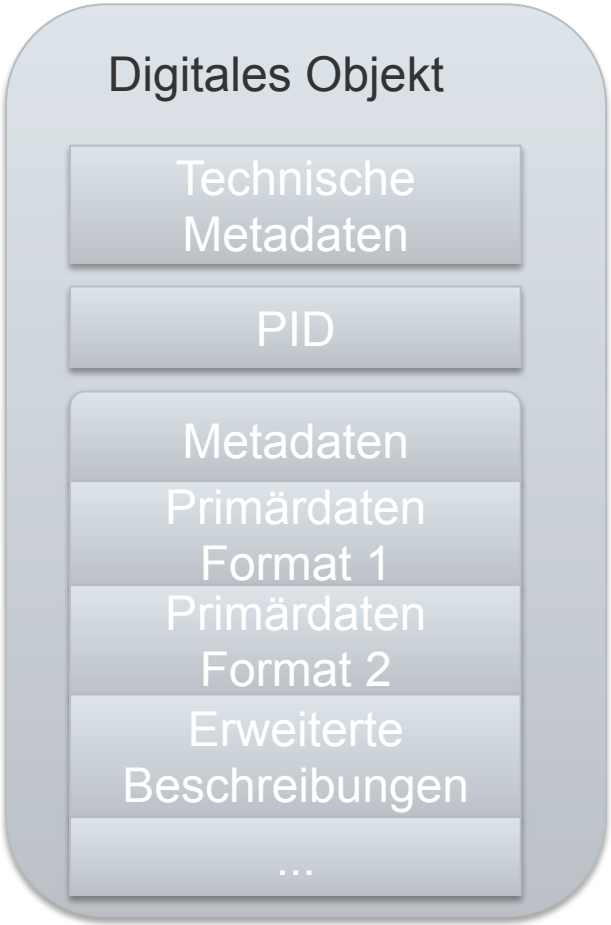
Katalog linguistischer Ressourcen

Faceted Browser

- Basierend auf OAI-PMH-Feeds
- Verwendung fachbezogener Facetten
  - z.T. vom Ressourcentyp gesteuert
  - Allgemeine Facetten für Suchraumbeschränkung
- Verwendung des Komponentenmodells für Metadaten
- Derzeit im „Community-Test“



Ablage in Verzeichnissen und Archivsystemen



- Von Archivsystemen automatisch erstellt
- Persistenter Identifikator, URN der DNB
- Beschreibende Metadaten vom Forscher erstellt
- Zusammengehörige Primärdaten vom Forscher erstellt
- Textuelle Beschreibungen, Dokumente, ... vom Forscher „beigelegt“

Eine Insel, eine Insel, noch eine Insel

Bild: Suat Eman / FreeDigitalPhotos.net

Wissenschaftsrat zur Archivierung und Zitierbarkeit

„Der Wissenschaftsrat fordert insbesondere die Forschungsförderer auf, Anreize zu schaffen, um qualitativ hochwertige Daten zu archivieren und langfristig zu sichern. Zu diesem Zweck sollten Referenz- und entsprechende Zitationsmöglichkeiten für Datensätze aufgebaut werden. *Persistent identifier* (PI) bzw*. digital object identifier* (DOI) erlauben eine eindeutige Identifizierung und Zitierbarkeit digital hinterlegter Daten selbst dann, wenn sie ihre Speicherorte (in der Regel referenziert über den *uniform resource locator*, URL) wechseln.“

Empfehlung zu Forschungsinfrastrukturen in den Geistes und Sozialwissenschaften, Wissenschaftsrat, Berlin, 28.01.2011, S.58

Standardisierung im Infrastrukturbereich

- Kommunikation der Lösungen miteinander
  - Austausch von Metadatenkatalogen
  - Zugriff über Webservices aufeinander
- Gemeinsame Metadatenstrukturen
  - Nach Typ der Ressource
  - Zentrale Definitionen von Datenkategorien
- Datenformate
  - Wo möglich: Datentypspezifische Standardformate
  - Offen für Reimplementierung
- Lösungen
  - Authentifizierung und Autorisierung
  - Datenschutz
  - Zentrale Infrastruktureinrichtungen

OAI-PMH

Komponentenmodell für Metadaten: CMDI

ISO 12620

Valides XML, TEI, ... (Datentyp abhängig)

Shibboleth

Verfahrensverzeichnis

Aufbau eines Archivsystems im SFB 833

- Kooperation mit der Bibliothek
  - Austausch über Technologien
  - Vermittlung von URNs der DNB
- Kooperation mit Zentrum für Datenverarbeitung (Rechenzentrum)
  - Nachhaltige Bereitstellung von Hardwareinfrastruktur
  - Zugangs-Authentifizierung
- Repository System
  - Fedora-Commons (derzeit im Test)
  - DSpace
  - Apache Jackrabbit (kein OAI-PMH integriert)

Persistente Datenhaltung im SFB 833

- Archivsystem für Forschungsprimärdaten
  - Hardware in Kooperation mit ZDV
  - Offenheit für unterschiedlichste Ressourcentypen
  - Persistente Identifikatoren vermittelt durch UB
- Offenheit gegenüber Austausch von Metadaten
  - OAI-PMH Schnittstelle
  - Zugriffsbeschränkungen nach Forschervorgaben für
    - Detaillierte Metadaten
    - Primärdaten
  - Komponentenbasierte Metadateninfrastruktur (CMDI)
- Suchfunktionalität für Ressourcen
  - Faceted Browser
  - Integriert Metadaten anderer Institutionen
  - Verweist auf jeweilige Archivsysteme/Institutionen

Kommunikation zwischen INF und den Teilprojekten

- Infrastruktur
  - Ansprechpartner auch innerhalb der Teilprojekte
  - Koordination mit zentraler Infrastruktur
  - Nationale und europaweite Abstimmung
- Assistenz für die Projekte
  - Erstellung angepasster Metadatenschemas
  - Beispielmetadaten
  - „Hands-on“ Training mit echten Projektdaten
    - Schulung zur Eingabe von Metadaten
    - Erläuterung von Datenkategorien
    - „der Blick des Außenstehenden“ auf die Beschreibung
  - Definition von Prozessen für die Archivierung
    - Werkzeuge
    - Eingabemasken
    - Checklisten
- Anpassungen der Archivprozesse nach Bedürfnissen der Forscher

Kommunikation mit dem Sondersammelgebiet Sprachwissenschaften (UB Frankfurt)

- Arbeitsteilung zwischen Fachwissenschaften und Bibliothek
  - Fachwissenschaften
    - Aufbau der Metadaten
    - Sammlung der Primärdaten
  - Bibliothek
    - Zugang für Fachwissenschaftler über Bibliotheksinfrastruktur
    - Zentraler Katalog
- Geplante Kooperation
  - Fachreferentin: Frau Heike Renner-Westermann
  - Projekt der UB Frankfurt: Einbindung von Primärdatensammlungen in eine Digitale Bibliothek für das Sondersammelgebiet

Zusammenfassung

- Primärdatenarchiv
  - Standardbasiert
  - Interoperabel
  - Für Fachdaten geeignet
- Metadaten
  - Standardbasiert
  - Verfügbar über offene Schnittstellen
  - Angepasst an Ressourcentypen
- Integration mit anderen Archiven
  - Bibliotheken
  - Nationalen und internationalen Archiven