EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
Seminar für Sprachwissenschaft
Allgemeine Sprachwissenschaft und Computerlinguistik

Prozesse zur Beschreibung und Archivierung linguistischer Forschungsdaten

Christina Hoppermann, Thorsten Trippel, Claus Zinn
Initiative zur Nachhaltigkeit Linguistischer Daten (NaLiDa)

Die Archivierung und Beschreibung von Forschungsdaten sind Aspekte, die nicht nur zur guten wissenschaftlichen Praxis zählen, sondern insbesondere von Forschungsförderungsorganisationen zunehmend über einen festgelegten Zeitraum, wie die von der Deutschen Forschungsgemeinschaft (DFG) geforderten 10 Jahre, vorgeschrieben werden. Diese Vorgaben stellen jedoch Herausforderungen an die Datenersteller, die über ihre fachwissenschaftliche Arbeit hinausgehen. Um diesen Mehraufwand minimieren zu können, stellt die Initiative zur Nachhaltigkeit Linguistischer Daten (NaLiDa) sowohl Verfahrensweisen als auch Anleitungen und Referenzen zur Beschreibung und Archivierung linguistischer Forschungsdaten zur Verfügung.

Archivierung

Wozu dienen *Repositorien*?

Ein Repositorium (Archiv) bezeichnet einen (digitalen) Ort zur nachhaltigen, langfristigen Sicherung von Daten und Dokumenten unter Einhaltung von Zugangsberechtigungen.
Repositorien bieten folgende Vorteile:

- Fähigkeit zur persistenten Datenspeicherung → Zugänglichkeit von Ressourcen,
- Zitierbarkeit von Ressourcen durch die Vergabe persistenter Identifikatoren (PIDs),
- Verbreitung von Ressourcen durch ihre Metadaten mittels OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting),
- Auffindbarkeit von Ressourcen durch Suchmöglichkeiten (z.B. durch einen Faceted Browser, der auf Metadaten basiert),
- Erfüllung der Vorgaben von Forschungsförderungsorganisationen bezüglich der Archivierung.

Organisation in digitale Objekte

Repositoriensysteme wie Fedora-Commons (www.fedora-commons.org) verwenden als zentrale Einheit *digitale Objekte,* die wiederum Datenströme enthalten.

- Digitales Objekt
  - Ressource mit ihren Metadaten und zugehörigen Dateien
  - Vergleichbar mit einem Ordner bzw. Verzeichnis bei Dateisystemen
  - Adressierbar mittels eines persistenten Identifikators (PID)
- Datenstrom:
  - Einzelner Bestandteil des digitalen Objektes
  - Vergleichbar mit Datei im Verzeichnis
- Vorteile dieses Ansatzes:
  - Nicht auf eine Ressourcenklasse beschränkt
  - Digitale Objekte beinhalten verschiedene Datentypen
  - Metadaten sind Datenstrom in dem Objekt
- Nachteile dieses Ansatzes:
  - Flache Struktur
  - Hierarchien nur durch Angabe von Relationen von Objekten
- Konsequenz: verschiedene Hierarchien sind gleichzeitig möglich

3 Phasen der Datenarchivierung

Datenersteller

Phase 1 Datenerstellung
- Abhängig vom Forschungsgebiet
- Auch mit spezieller Software
- z.B. Experimentaldaten, Korpora, Lexika, Faksimiles, Software
- Vorbereitung für die Archivierung

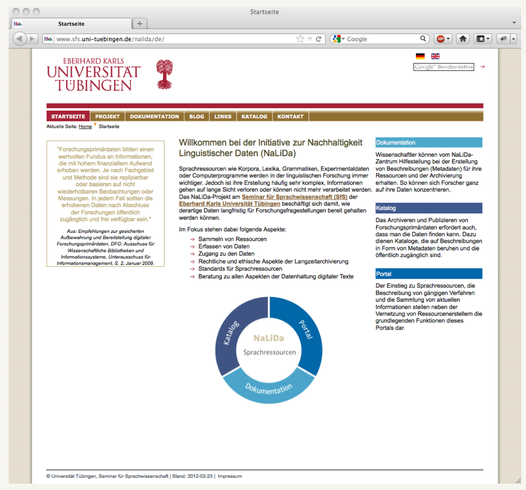
Wissenschaftler Support: Archivverwalter

Phase 2 Übertragung der Daten ins Repositorium
- Bestimmung des Ressourcentyps
- Hochladen der Daten (in ihren jeweiligen Formaten)
- Festlegung der Zugriffsrechte
- Beschreibung der Ressourcen

Archivverwalter

Phase 3 Überprüfung der Daten
- Konsistenz der Beschreibungen
- Verifikation des Eingangs der Daten
- Beantragung von persistenten Identifikatoren für die Ressourcen
- Eintrag der persistenten Identifikatoren
- Bereitstellung der Beschreibungen für Suchfunktionen

NaLiDa bei der Beschreibung und Archivierung von Forschungsdaten



Sprachressourcenportal der Initiative zur Nachhaltigkeit Linguistischer Daten (NaLiDa):
www.sfs.uni-tuebingen.de/nalida

NaLiDa hat folgende Schwerpunkte:

- Unterstützung der Archivierung durch die Erstellung von Metadaten:
  - Bereitstellung von Profilen (Beschreibungsvorlagen) für verschiedene Arten von Ressourcen
  - Unterstützung bei der Verwendung der Profile
  - Bearbeitung existierender Bestandsmetadaten in anderen Formaten
  - Hilfe bei der Erstellung von Metadaten, z.B. durch den Metadaten-Editor ProFormA
- Bereitstellung von Informationen zum Archivieren von Forschungs-primärdaten:
  - Erstellung von digitalen Objekten (Identifikation der Bestandteile)
  - Beschreibung der Arbeitsabläufe zur Archivierung
  - Beratung zu nachhaltigen Datenformaten für Primärdaten
- Erfassen von Daten:
  - Unterstützung bei der Erfassung von Primärdaten und Metadaten
  - Suche über die Metadaten
- Standards für Sprachressourcen (insbesondere für Metadaten)

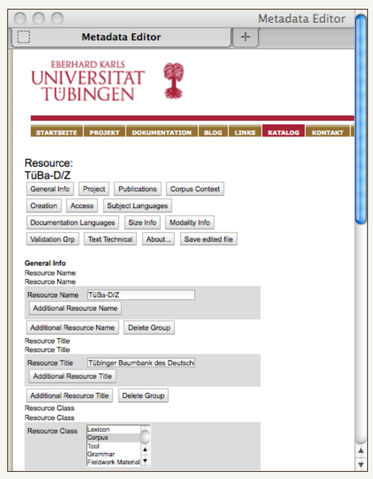
ProFormA: Ein Editor zur Unterstützung bei der Erstellung von Metadaten

ProFormA:

- Formular- und webbasierter Editor
- Verwendung: Erstellen und Editieren von Metadaten
- Basiert auf XForms, einem W3C-Standard zur Erstellung webbasierter Formulare
- Repräsentiert zugrunde liegende XML-Strukturen von Metadaten in Einer nutzerfreundlichen Formularansicht in HTML

Vorteile:

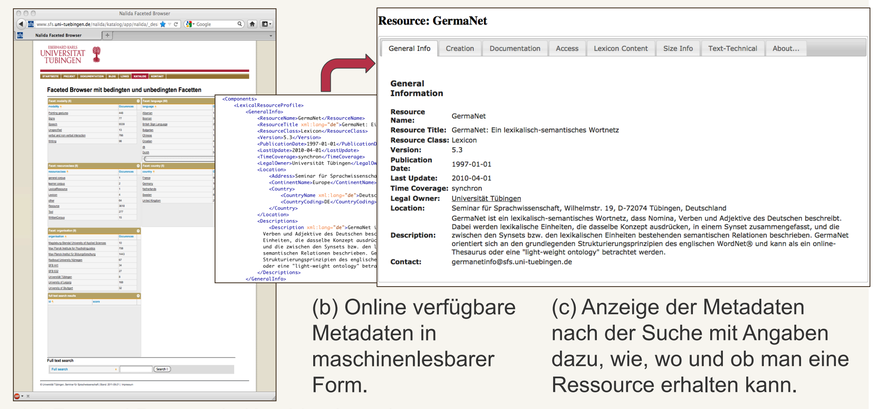
- Minimiert die Komplexität des Metadatenerstellungsprozesses
- Kann ohne (XML-)Vorkenntnisse angewendet werden
- Stellt keinen Mehraufwand für Nutzer dar



Beschreibung der Daten: Metadatenerstellung

Wozu dienen Metadaten?

- Metadaten sind strukturierte Daten, die sowohl zur Beschreibung als auch zum Auffinden von Forschungsdaten verwendet werden können
- Auffindbarkeit von Ressourcen
- Überblick über Ressourcen und Zugang zu Ressourcen
- Zitierbarkeit sowohl der Ressourcen selbst als auch der Metadaten



(a) Faceted Browser in Kombination mit Volltextsuche als Suchfunktion über Metadaten.

(b) Online verfügbare Metadaten in maschinenlesbarer Form.

(c) Anzeige der Metadaten nach der Suche mit Angaben dazu, wie, wo und ob man eine Ressource erhalten kann.

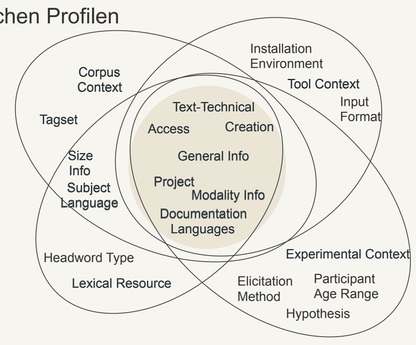
Component MetaData Infrastructure (CMDI)

CMDI basiert auf drei grundlegenden Konzepten:

- Profile
  - Beschreibungsvorlagen für Ressourcen-typen
  - beinhalten Komponenten als Bausteine
- Komponenten
  - gruppieren Datenkategorien zu semantischen Einheiten
  - können selbst weitere Komponenten enthalten
  - wiederverwendbar in unterschiedlichen Profilen
- Datenkategorien

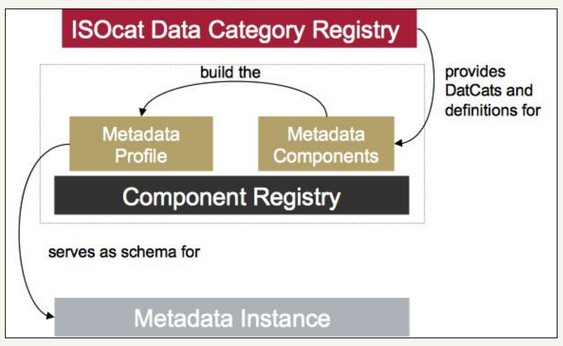
CMDI ist ein Framework für Metadaten-schemas, das im Kontext des EU-Projektes CLARIN entwickelt wurde (s. http://www.clarin.eu/cmdi):

- Wiederverwendbar und flexibel
- Eignung für die Beschreibung unterschiedlicher Ressourcentypen (z.B. Lexika, schriftliche/gesprochene Korpora, Grammatiken, Experimentaldaten, Software/Tools, etc.), die unterschiedliche Beschreibungsebenen erfordern



Zentrale Verzeichnisse zur Definition von CMDI Beschreibungsmitteln:

- ISOcat (http://www.isocat.org)
- Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/#)



Zusammenhang von Metadaten und Ressourcen

Zweitverwertung:

1. Daten und Ergebnisse können mit anderen Quellen verbunden werden (s. auch AG: *Linked Data in Linguistics*)
2. Bei einer Suche nach Autoren in Bibliothekskatalogen ist die Angabe von Daten und Ergebnissen aus Repositorien und Archiven möglich
3. Andere Projekte können mit Erlaubnis die Ausgangsdaten verwenden

Verwendung der Ressource ← **Ressource** → Zugriff auf die Ressource

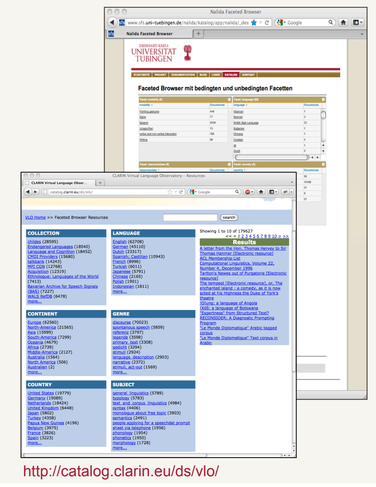
Evaluation:

1. Begutachtung und Überprüfung der Daten
2. Vergleich der Analysen: mit unter-schiedlichen Methoden
3. Reproduktion von Analysen

Ressource ← **Metadaten** ← Suche über Metadaten

Metadaten:

1. Finden vergleichbarer Methoden und Ansätze
2. Kontakt zu potentiellen Kooperations-partnern



Via http://www.sfs.uni-tuebingen.de/nalida/de/katalog/

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

Universität Tübingen · Philosophische Fakultät
Seminar für Sprachwissenschaft
Wilhelmstraße 19 · 72074 Tübingen · Germany
www.sfs.uni-tuebingen.de/nalida
nalida@sfs.uni-tuebingen.de