INSTITUT FÜR DEUTSCHE SPRACHE

Anwendungsfall:
Germanistische Sprachwissenschaft

Begutachtung des WissGrid AP 3

28. Januar 2010, AIP Potsdam

Andreas Witt

Institut für Deutsche Sprache (IDS)

Bundesministerium
für Bildung
und Forschung

Allgemeiner Hintergrund zur Fachdisziplin

- Die germanistische Community untersucht die deutsche Sprache und die deutschsprachige Literatur in ihren historischen und gegenwärtigen Erscheinungsformen.
- Stellvertretend *Institut für Deutsche Sprache*: Stiftung des bürgerlichen Rechts, Mitglied der Leibnizgemeinschaft.
- Auftrag: außeruniversitäre Einrichtung zur Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte.
- Erstellen von Sprachdatensammlungen in Gestalt von Textkorpora als Forschungsgrundlage an empirischen Sprachdaten.

Motivation zur Langzeitarchivierung

- Unwiederbringlichkeit der Daten.
- Basis für die empirische Sprachwissenschaft.
- Lösungsstrategien in Lizenzrechtsfragen unter Berücksichtigung der Grundrechtskonflikte:

  Grundrecht auf Eigentum vs. Freiheit der Forschung.

  Grundrecht auf Eigentum vs. Grundrecht auf Informationelle Selbstbestimmung.
- Erste Schritte: Mitgliedschaft in DFN AAI.

Informationen zu den Daten

- 1,5 Mio. Einzeltexte in **Textkorpora**:

  Texte in geschriebener Sprache, Multimodale Daten, Rohdaten, versionierte Korpora.

  Metadaten und Annotationen.

  Virtuelle Kollektionen (PID-Verweise, mehrere Millionen Einzelressourcen).
- Umfang der Textdaten:

  ca. 1000 Dateien zu 32 GB; insgesamt etwa 5 TB (perspektivisch 10 TB in zwei Jahren).

  Umfang multimodaler Daten noch nicht abschätzbar.

Informationen zu den Daten

- **Datenformate**:

  Metadaten: Basisformat XML, XCES/ TEI-Header.

  Relationen in Metadaten über PIDs (z.B. Versionsrelationen, hierarchische Relationen, Ähnlichkeitsrelationen).

  Provenienz: XCES in TEI-Header angegeben (dokumentiert Aufarbeitung durch das IDS und Textquelle).

  Prozessdokumentation: für Textdaten im TEI-Header enthalten; für multimodale Daten in nichtstandardisierten Metadatenformaten.

  Geplant ist eine Überführung der Korpora von XCES in TEI P5.

Die (geplante) Grid-Forschungsumgebung

- **Archivierung**:

  Zeitplan: Schrittweise Einbringen der Daten.

  Einmaltransfer des bestehenden Materials.

  Täglich fallen neue Rohdaten an.

  Halbjährlicher Transfer in ein neue aktuelle Version.
- **Rechtliche Grundlagen:**

  Absprachen und Verträge mit Urhebern.

  Gelegentliche Interaktion mit den Textgebern.
- **Konsequenz**:

  Nachträgliches Abändern, Erweitern oder Löschen.

  Versionierung.

Die (geplante) Grid-Forschungsumgebung

- **IPR und Datenschutz** :

  130 Lizenzabsprachen: Korpora Eigentum der Urheber.

  Einfaches Nutzungsrecht durch den Bearbeiter.

  Geringer Teil der Ressourcen unterliegt nur GPL und CC-Bedingungen.

  Multimodale Daten unterliegen Datenschutzbe-stimmungen und gesonderten Einwilligungserklärungen.

  Datenspeicherung muss im IDS erfolgen.

Zusammenfassung

- Geschilderte Missstände erfordern *dringend* eine Strategie zur Langzeitarchivierung:

  Berücksichtigung der bestehenden **Lizenzvereinbarungen**.

  Repository B: Kollaborative Datenbearbeitung.

  Trust Zones: Daten bleiben formal im Haus, werden aber extern archiviert.

  Metadatenextraktion: Synergien mit IDS-Bestand.

  Validierung: Sicherung der Nutzbarkeit der archivierten Daten.

  Konvertierung: XCES in TEI P5.

Spezialfall TextGrid

- TextGrid als Ausgangsbasis für übergreifende Langzeitarchivierung.
- Berücksichtigung geisteswissenschaftlicher Anforderungen im Forschungsdatenarchiv .
- Übernahme des prototypischen Forschungsdatenarchivs von WissGrid.
- Einbindung der TextGrid-Dienste in das Repository.
- Implementierung von Schnittstellen zu anderen geisteswissenschaftlichen Datenarchiven.