MPI-Psycholinguistik - The Language Archive Nijmegen/NL

Begehung am 28.11.2012
Die Angaben entsprechen dem Stand von Nov. 2012 und stellen nicht die aktuelle Situation im TLA dar

Aufgaben

Hauptaufgaben:	Erhebung, Analyse und Archivierung der Dokumentation von bedrohten Sprachen; Entwicklung von Software-Tools und einer portablen Forschungsdaten-Infrastruktur (The Language Archive Technology (LAT)), die auch Archivierungsaufgaben übernimmt (z. B. zur teilautomatisierten Verschlagwortung von Metadaten)
weitere Aufgaben:	Unterstützung des Forschungsdatenmanagements
nicht Teil der Aufgaben:	nicht relevant
relevante Fachrichtung:	Psycholinguistik, Sprachforschung, Ethnologie

Allgemeine Angaben

Rechtsform:	zugehörig zur Max-Planck-Gesellschaft und. zum MPI für Psycholinguistik, TLA aber teilfinanziert, getragen von MPI, KNAW und BBAW
Gründung/Geschichte::	seit 1990 werden digitale Daten der Disziplin archiviert, seit 2000 – 2016 Projekt „Dokumentation bedrohter Sprachen“ (DoBeS), Finanzierung durch VW-Stiftung 09.11. Gründung „The Language Archive“ am MPI für Psycholinguistik
Anzahl Mitarbeiter:	30-33 MA (davon 7 feste Stellen) Archive und Digitalisierung 8 MA Entwicklung Archivsoftware 7 MA Entwicklung Annotationssoftware 4 MA Entwicklung Knowledgesoftware 4 MA technisches Personal 5 MA Anzahl FTE (Planstellen): 7 Anzahl Hilfskraftstellen (Planstellen): unbekannt Anzahl FTE (befristet, Drittmittel): 23 Anzahl Hilfskraftstellen (befristet, Drittmittel): unbekannt Anzahl WissMA: 8 (Datenkuratierung) Anzahl IT: unbekannt
Zertifizierung:	ISO TC37/SC4 ISO DCR 12620 DCR, 24618 PISA, 24622 CMDI Data Seal of Approval
Referenzmodell:	kein Referenzmodell

Datenbestand

Verpflichtung Daten zu sichern?	projektbezogene Verpflichtung Daten zu archivieren
Findet gezielte Datenakquise statt?	Nein

Anzahl an Projektarchiven:	ca. 200 Sprachen archiviert (davon 80 bedrohte Sprachen) ca. 21.000 h (Audio- und Videoaufnahmen), ca. 167.000 mit Metadaten versehene Aufzeichnungen, über 5 Mio. kommentierte Textabschnitte, über 90 Lexika
Anzahl an „grauer Literatur“:	nicht relevant
Anzahl an Doktorarbeiten:	unbekannt
Anzahl an retrodigitalisierten Zeitschriften:	nicht relevant
Anzahl an Einträgen Nachweiskatalog:	nicht relevant
Gesamtgröße Archiv:	65.000 GByte (65 TB)
Datensatzgröße:	ca. 1 TByte
Datenzuwachs jährlich:	17 TB/Jahr
Altdaten vorhanden:	Nein

Datengenerierung

technische Unterstützung bei Datenerhebung (z. B. durch Tools, Software, etc):	Ja
fachliche Unterstützung bei Datenerhebung:	Ja

Datenupload

Online-Upload möglich?	Ja, mit Hilfe von Language Archive Management and Upload System (LAMUS)
Auswahlkriterien für Aufnahme von Daten:	fachliche Relevanz, technische Beurteilung und enthaltene Metadatenbeschreibung in CMDI
Werden angebotene Daten abgelehnt? (Gründe)	Nein
Vertragsverhältnis Datengeber - Datenarchiv:	unbekannt.
Software/Collection Registry:	verschiedene Tools zur Registrierung von Projektdokumenten

Kuratierung von Daten

Arbeitsschritte zur Datenarchivierung:	nicht OAIS-konformes Archiv; kompletter Datenzyklus von Erhebung, Analyse Archivierung und Bereitstellung wird durch Software-Suite aus insgesamt 13 Tools abgebildet und unterstützt Vorbereitung Roh-Daten werden manuell mittels der Eigenentwicklungen ins System eingespielt Integration Bis zur Archivierung werden die Daten automatisch mit wenig manuellem Aufwand durch verschiedene Anwendungen weitergereicht und archivtauglich aufbereitet.

Werkzeuge für Datenarchivierung:	ANNEX, LEXUS, IMEX, TROVA, VICOS Eigenentwicklung: Ja, alle verfügbare Software (kommerziell, OpenSource): Open Source
Versionskontrolle:	Ja
Verhältnis manuelle/automatisierte Datenkuratierung:	vergleichsweise wenig bis keine manuelle Kuratierung notwendig, da Daten nur über ein vorgegebene Software-Suite erzeugt und verwaltet werden können
Aufwand Datenkuratierung insgesamt (Personen, Zeit):	8 MA
wissenschaftliche Bewertung von Daten:	Nein
Verwendung Repository-System:	Nein
Software zur Verwaltung/Ablage der Archivdateien:	LAMUS, IMDI-Browser; im Backend strukturierte, verzeichnisbasierte Dateiablage Eigenentwicklung: verschiedene Scripts für Synchronisation, Checksummen, Kopiervorgänge, Aktualisierungen etc. Verfügbare Software (kommerziell, OpenSource): Open Source; Versuche mit iRods als Logical Replication Layer
Software zur Verwaltung Metadaten/Rechte/Versionen:	ARBIL, Access Management System (AMS) Eigenentwicklung: Ja verfügbare Software (kommerziell, OpenSource): Open Source

Archivierung von Daten

technischer Partner für LZA:	RZ der MPG Garchingen, GWDG Göttingen
Anzahl redundanter Datenkopien:	4 volle dynamische Kopien mit 50-jähriger Garantie (2 RZ in München, 1 davon Garching, 2 Rechenzentren in Göttingen, 1 davon GWDG 2) 2 Kopien in Nijmegen, 1 Kopie in MPI Leipzig 11 zusätzliche regionale Repositories
Administration Archivsystem inhouse/extern:	intern
garantierter Zeitraum:	50 Jahre
Erfahrungen mit technischer Migration von Archivdateien:	Ja

Bereitstellung von Daten

Arbeitsschritte zur Datenbereitstellung:	kaum zusätzliche Arbeitsschritte notwendig, um aktuelle und archivierte Daten zur Nachnutzung freizugeben; Daten können mittels verschiedener Tools onlinegesucht, durchsucht und aufgerufen werden
Werkzeuge für Datenbereitstellung:	LEXUS, TROVA, IMDI-Browser Eigenentwicklung: Ja, alle verfügbare Software (kommerziell, OpenSource): Open Source
Registrierung bzw. Login für Nutzer notwendig?	Ja
Wird die IP und der Zugriff geloggt?	unbekannt
Gibt es verschiedene Authentifizierungsstufen?	ja, Authentifizierung über Shibboleth

Zugriff:	Online über Portal Open Access → direkter Zugriff: wenige Datensammlungen Restricted Open Access → Verhaltenskodex muss durch User aktiv zugestimmt werden: weniger als 3 Datensammlungen Protected → User fragt Eigentümer an, Art der Nutzung wird festgehalten, Zustimmung Verhaltenscodex, Zugang: die meisten Datensammlungen Closed → nur Eigentümer hat Zugang: wenige Datensammlungen
Wie wird ein Missbrauch von Daten verhindert/kontrolliert (z.B. wirtschaftliche Nutzung)?	k.A.
Rollenkonzept:	Datenbearbeiter/-geber Projektmitarbeiter Datennutzer Administrator
Qualitätssicherung:	Daten werden durch entsprechende Tools qualitätsgesichert erstellt, archiviert und bereitgestellt
Schutzmechanismen für bestimmte Informationen:	Access rights können für jeden Datensatz individuell gesetzt werden; Eigentümer kann Zugriffsrechte in bestimmten Fällen einschränken personenbezogene Metadaten/Dateien: unbekannt raumbezogene Metadaten/Dateien: unbekannt

Metadaten & Interoperabilität

Mindestanforderungen an Daten:	Mindestangaben für Datensätze müssen vorliegen, geringe technische Vorgaben, hohe Formatkonsistenz wird angestrebt
Verwendung von Standards:	Ja, s. u. ISO 12620
Metadatenmodell:	IMDI, CMDI (bis Juni 2013 Umstellung auf CMDI) eigene Anpassungen an Metadaten?: Ja, falls erforderlich aktive Beteiligung an der Weiterentwicklung von Standards
Vorhandene Schnittstellen:	OAI-PMH, Mapping auf Dublin Core
Sichtbarkeit der Metadaten in anderen Portalen/Aggregatoren:	unbekannt
Sichtbarkeit der archvierten Dateien in anderen Portalen/ Aggregatoren:	unbekannt
System für Persistente Identifikatoren:	handle: jedes Datenobjekt (= Strukturelle Metadaten, beschreibende Metadaten, Audio/Video-Dateien, Annotationen, Text) wird in separater Datei gespeichert und erhält individuell ein PID Anzahl PIDs insgesamt: unbekannt

Nicht-Technische Dienstleistungen

Bereitstellung Guidelines, Ratgeber etc.:	- zu den verschiedenen Anwendungen der LAT - zur Dokumentation von Sprache - Dokumentenarten (Audio/Video/Photo, Lexika) - Metadaten und Standards.
Antragsberatung:	Ja
Schulungen/Workshops:	OnlineTutorials, Klassische Schulungen hauptsächlich für Projektbeteiligte, aber einige Kurse sind auch offen für externe Wissenschaftler

Weiterentwicklung Standards:	Ja, aktive Beteiligung
Support Datenmanagement:	unbekannt

Finanzen: Einnahmen

Jahresbudget:	unbekannt davon speziell für das Datenzentrum k.A.
Basisfinanzierung:	7,5 Stellen durch bi-nationale Kooperation (NL 53 %, D 47 %) NL – MPI Nijmegen (3 MA) D – Max Planck Gesellschaft (MPG) (2,5 MA) D – Berlin Brandenburgische Akademie (BBAW) (1MA) NL – Dutch Academy of Science (KNAW) (1MA) andere MA durch externe Projekte Zeitraum der vertraglich Finanzierung 7 Jahre (momentan für 5 Jahre), nach 3 Jahren Evaluation (Stand Nov. 2012) nachhaltige Finanzierung? Nein
Drittmittel:	unbekannt
restliche Finanzierung:	100 % durch MPI für Psycholinguistik mit Ende des DoBeS-Projekts läuft Finanzierung für Entwicklung der LAT aus
Gebühren für das Archivieren von Daten:	keine, da in TLA nur Daten aus Projekten (wie z. B. DoBeS, u. a.) archiviert werden
Gebühren für die Nutzung von Daten:	keine

Finanzen: Ausgaben

Kostenstruktur nach Kostenträger:	Personalkosten unbekannt Planstellen 7 Stellen Drittmittelstellen 23-26 Stellen Hardware/Software unbekannt Reisekosten, Gebäudeinstandhaltung unbekannt

Kostenstruktur nach Kostenstellen:	Daten akquirieren und aufbereiten unbekannt Daten archivieren unbekannt Daten zugänglich machen unbekannt Softwareentwicklung MPI-TLA hat 10 Jahre lang 8 FTE´s für die Toolentwicklung größtenteils durch Projekte und externe Gelder finanziert Kosten für PIDs unbekannt
Höhe Durchschnitts-Kosten pro 1 GB Daten:	unbekannt

Negativa

nicht funktionierende/ optimierbare Workflows oder Prozesse:	unbekannt
nicht gelöste Probleme:	Aufrechterhaltung der gesamten Infrastruktur nach Auslaufen der aktuellen Projektförderung noch unklar