Archiving LEXUS 3 multimedia lexica

SEBASTIAN DRUDE, ANDRÉ MOREIRA, MENZO WINDHOUWER, SHAKILA SHAYAN
The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Archiving LEXUS 3 multimedia lexica

1. Hintergrund
2. Lexus 3
3. Archivierung

1. Hintergrund

Oft nicht ausreichend berücksichtigt:
LANGZEITPERSPEKTIVE von digitalen Resourcen

1. Hardware (fragile Speichermedien)
2. Lokalisierung (sich ändernde URLs)
3. Datenformate (Veralterung, proprietär)

Lösungsaspekte:

1. Viele automatische Kopien
2. Persistente Identifikatoren
3. Offene, gut dokumentierte Formate, Migrierung

→ Stabile Institutionen, Infrastrukturen

1. Hintergrund

- "The Language Archive" – technische Abteilung am Max-Planck-Institut für Psycholinguistik
- Datenarchiv: Schwerpunkt auf Multimedia, Korpora gesprochener Sprache
- Insb. Feldforschungsdaten, ca. 200 Sprachen
- Software-Entwicklung: Spracharchiv-Server (IMDI/CMDI), ELAN (Audio/Video-Annotation), ARBIL, ISOcat, KinOath, **LEXUS**
- Expertise in Datamanagement, -infrastructures, standards, many collaborations (CLARIN etc.)

2. LEXUS 3

- Web-based Tool für Bearbeitung und Anzeige von Multimedia-Lexica
- Entwicklung begann 2004 als Teil des LIRICS-Projekts zur Unterstützung von LMF

  Lexus 1 (2005): Back-end java, Front-end html
- Entwicklung ging weiter im Rahmen des DoBeS-Programms (2006 – 2010)

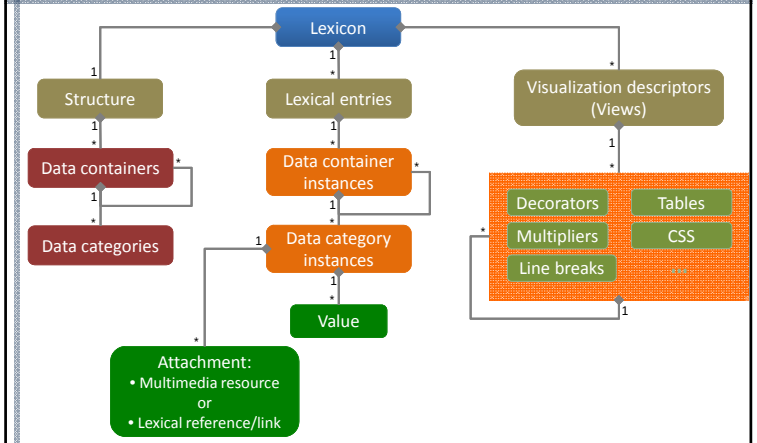
  Lexus 2 (2008): F-E in Flash
- 2012/11: **Version 3**, ohne spezifisches Projekt

  B-E rewrite, BaseX XML Datenbank, LMF support

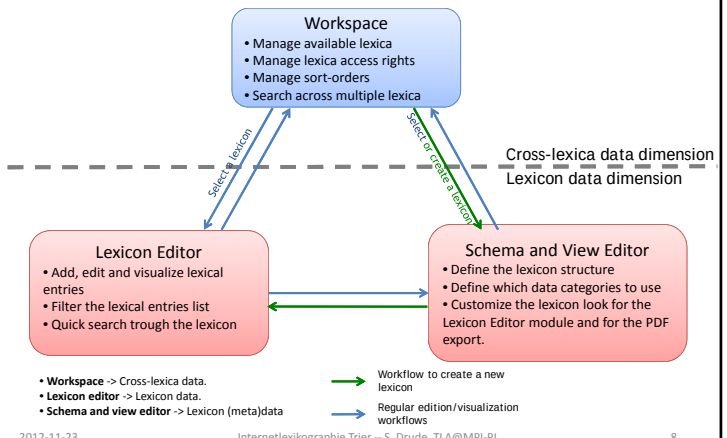
2. LEXUS 3

- Kollaborativ, web-based
  - Mehrere Benutzer können am selben Lexikon arbeiten
- Multimedia-Unterstützung
  - Verschiedene Typen Multimedia (z.B. Audio, Video, Bilder)
- LMF-kompatibel
  - Erlaubt, LMF-konforme Lexika zu erstellen/bearbeiten
- ISOcat-Verbindung
  - Datenkategorien werden direkt von ISOcat eingebunden
- Anpassbare Visualisierungs-Lay-outs
  - Lexikalische Liste, einzelne Einträge, Export (u.A. PDF)
- Konfigurierbare sort-orders
  - Beliebige Buchstaben & Kombinationen (unicode support)
- Komplexe Suchfunktion, auch über mehrere Lexika

2. LEXUS 3



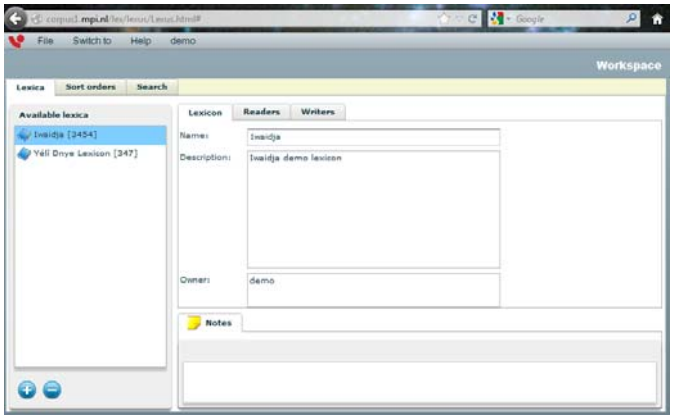
2. LEXUS 3



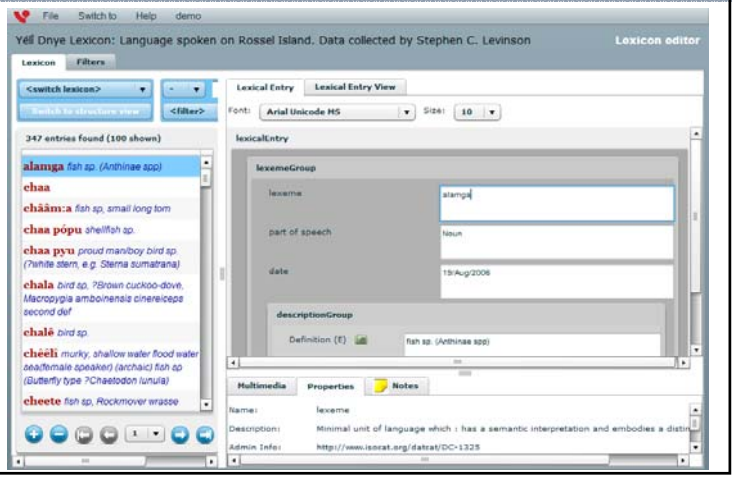
3. Lexus: DEMO

- http://corpus1.mpi.nl/lex/lexus/index.html

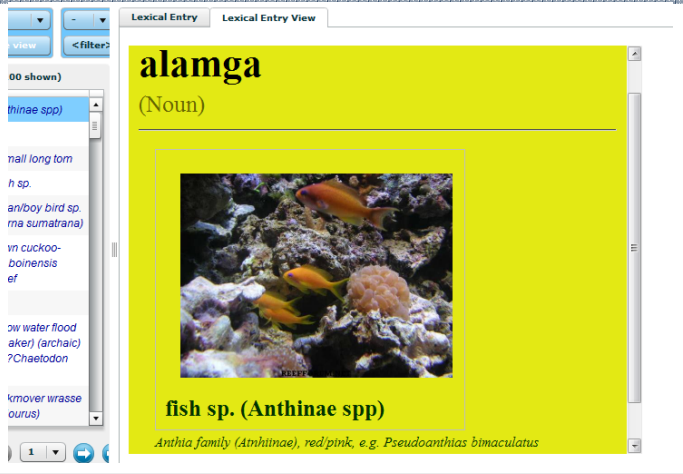
3. Lexus DEMO



3. Lexus DEMO



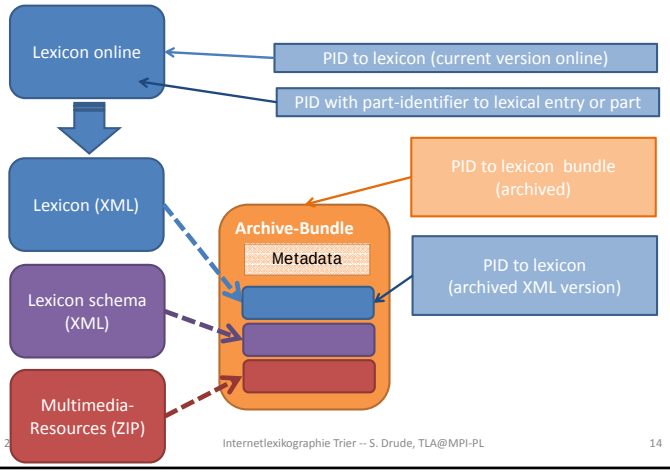
3. Lexus DEMO



3. Archivierung

- Datenarchiv derzeit zugepaßt auf Multimedia- und Annotations-Dateien
- Gegenwärtig können Lexikon und Schema als zwei XML-Dateien exportiert und dann (mit Metadata) zusammen archiviert werden
- Das Archiv ist kein Workspace, Archivierung ist manuell, nur nach größeren Überarbeitungen
- V.3.1: Multimedia-Dateien als ZIP-Datei
- Keine automatische Versionierung, noch keine PIDs ins Innere von älteren Versionen
- Ältere Vers. würden manuell in LEXUS geladen

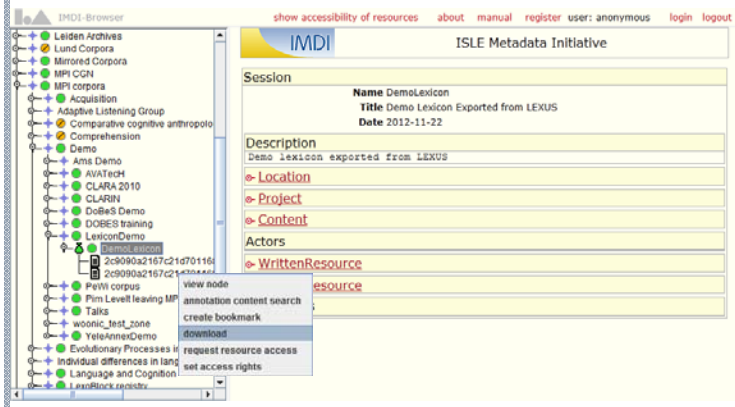
3. Archivierung



3. Archivierung: Demo

- http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI1569128%23

3. Archivierung: Demo



3. Archivierung: Demo

Lexicon file (XML, archived)

3. Archivierung: Demo

Lexicon schema file (XML, archived)

3. Archivierung

Ideal wäre:

- Volle Versionierung (wie bei MediaWiki u.Ä.)
- Automatische regelmäßige Archivierung
- Automatische separate Archivierung aller Resourcen (Multimedia), mit je eigenen PIDs
- Automatische Wiederherstellung älterer Versionen mit Multimedia möglich
- PIDs zu Artikeln und Artikelteilen werden immer problematisch sein – neue PID für jede Korrektur? Besser per Datum, Versionsnr. o.Ä.

Archiving LEXUS 3 multimedia lexica

Sebastian Drude

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands