



## Archiving LEXUS 3 multimedia lexica



SEBASTIAN DRUDE, ANDRE 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

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 2

## 1. Hintergrund

Oft nicht ausreichend berücksichtigt:  
LANGZEITPERSPEKTIVE von digitalen Ressourcen



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

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 3

## 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.)

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 4

## 2. LEXUS 3

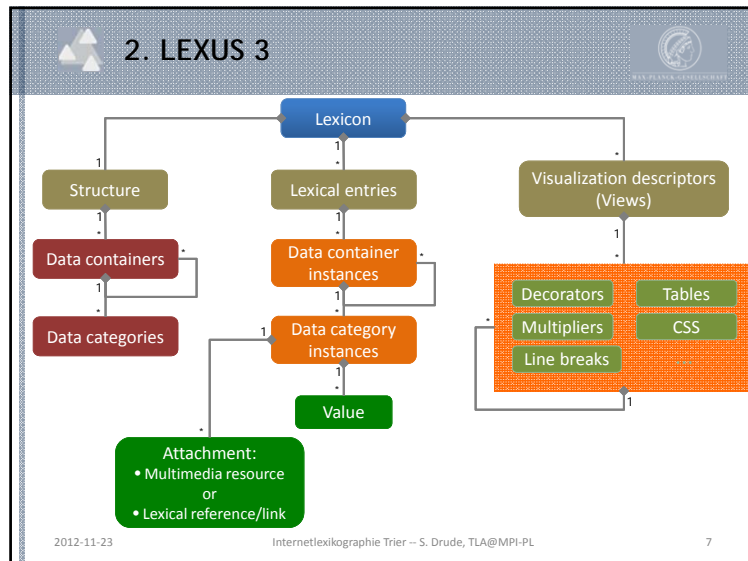
- Web-based Tool für Bearbeitung und Anzeige von Multimedia-Lexica
- Entwicklung begann 2004 als Teil des LIRICS-Projekts zur Unterstützung von LMF
- **Lexus 1 (2005):** Back-end java, Front-end html
- Entwicklung ging weiter im Rahmen des DoBeS-Programms (2006 - 2010)
- **Lexus 2 (2008):** F-E in Flash
- **2012/11: Version 3, ohne spezifisches Projekt**  
B-E rewrite, BaseX XML Datenbank, LMF support

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 5

## 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

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 6



3. Lexus: DEMO

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

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 9

3. Lexus DEMO

The screenshot shows a web browser window with the URL [corpus1.mpi.nl/lex/lexus/LexusIndex.html](http://corpus1.mpi.nl/lex/lexus/LexusIndex.html). The interface includes a menu bar (File, Switch to, Help, demo) and a 'Workspace' header. On the left, under 'Available lexica', two items are listed: 'Inaidja [3454]' and 'Yell Dnye Lexicon [347]'. The main area is divided into 'Lexicon', 'Readers', and 'Writers' tabs. The 'Lexicon' tab is active, showing fields for 'Name' (Inaidja), 'Description' (Inaidja demo lexicon), and 'Owner' (demo). There is also a 'Notes' section at the bottom.

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 10

3. Lexus DEMO

The screenshot shows the 'Lexicon editor' for the 'Yell Dnye Lexicon: Language spoken on Rossel Island. Data collected by Stephen C. Levinson'. The interface includes a menu bar (File, Switch to, Help, demo) and a 'Lexicon' header. On the left, there is a list of 347 entries found (100 shown), with 'alamga' selected. The main area is divided into 'Lexical Entry' and 'Lexical Entry View' tabs. The 'Lexical Entry View' tab is active, showing a form for editing the entry 'alamga'. The form includes fields for 'lexeme' (alamga), 'part of speech' (Noun), 'date' (19/Aug/2006), and 'descriptionGroup' (fish sp. (Anthinae spp)). There is also a 'Definition' field with the text 'fish sp. (Anthinae spp)'. At the bottom, there are sections for 'Multimedia', 'Properties', and 'Notes'.

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 11

3. Lexus DEMO

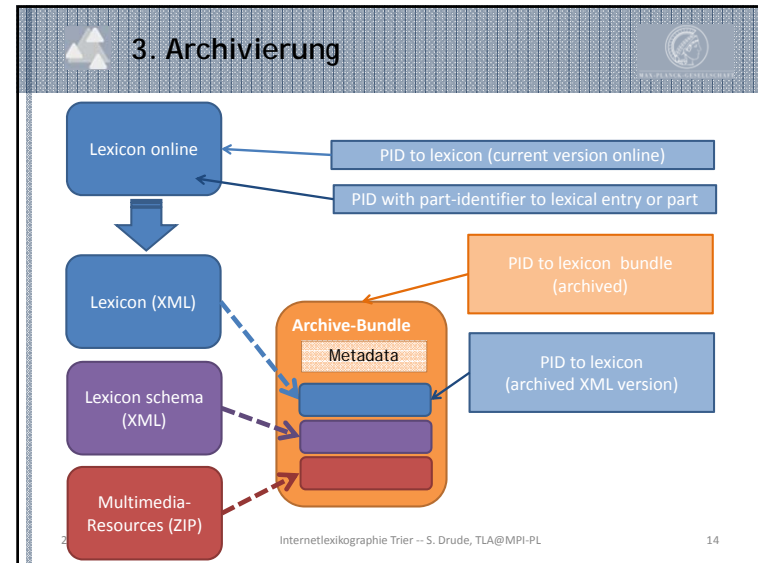
The screenshot shows the 'Lexical Entry View' for the word 'alamga'. The interface includes a menu bar (File, Switch to, Help, demo) and a 'Lexical Entry' header. On the left, there is a list of 00 shown entries, with 'alamga' selected. The main area is divided into 'Lexical Entry' and 'Lexical Entry View' tabs. The 'Lexical Entry View' tab is active, showing a detailed view of the entry 'alamga'. The entry is a noun, and it includes a photograph of a fish. The entry is described as 'fish sp. (Anthinae spp)' and 'Anthina family (Anthinae), red/pink, e.g. Pseudoanthias bimaculatus'. There is also a 'Definition' field with the text 'fish sp. (Anthinae spp)'. At the bottom, there are sections for 'Multimedia', 'Properties', and 'Notes'.

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 12

### 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

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 13



### 3. Archivierung: Demo

- [http://corpus1.mpi.nl/ds/imdi\\_browser/?open\\_path=MPI1569128%23](http://corpus1.mpi.nl/ds/imdi_browser/?open_path=MPI1569128%23)

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 15

### 3. Archivierung: Demo

2012-11-23 Internetlexikographie Trier -- S. Drude, TLA@MPI-PL 16



### 3. Archivierung: Demo

Lexicon file (XML, archived)

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon id="uid:2c9090a2167c21d701168047fdce4a45" version="1.0" xmlns="http://www.mpi.nl/lexus">
  <lexicon-information>
    <name>Yéll Dnye Lexicon</name>
    <description>Language spoken on Rossel Island. Data collected by Stephen C. Levinson Demo Yéll Dnye lexicon</description>
  </lexicon-information>
  <lexical-entry id="uid:c094e4a3-7ffa-4c38-946c-03e9402a1972" schema-ref="uid:2c9090a2167c21d70116804698914a02">
    <container id="uid:a4b5ab90-e428-40ad-9a3a-9c5c5d4a70d" schema-ref="uid:2c9090a2167c21d70116804698924a04" name="lexemeGroup">
      <data id="uid:442c6729-477c-4898-83a8-d3c3b27b9418" schema-ref="uid:2c9090a2167c21d70116804698984a3a" name="lexeme" start-letter="03" sort-key="022048015015000000000000000000000000000000000000000000000000000000">
        <value>chaa</value>
      </data>
      <data id="uid:4a6d3a98-cf93-47a8-8670-bf707191cbb4" schema-ref="uid:2c9090a2167c21d70116804698984a32" name="part of speech">
        <value>V</value>
      </data>
      <data id="uid:15555cc5-e3bd-4fd8-b7e7-aab60d478a84" schema-ref="uid:2c9090a2167c21d70116804698984a36" name="date">
        <value>26/Nov/2007</value>
      </data>
      <container id="uid:b0918f30-f005-4a35-a66d-5415b4a9748b" schema-ref="uid:2c9090a2167c21d70116804698924a06" name="descriptionGroup">
        <data id="uid:9f93f160-b0b0-44b4-9ba7-5931ee4b7ffc" schema-ref="uid:2c9090a2167c21d70116804698924a06" name="description">
          ...
        </data>
      </container>
    </lexical-entry>
  </lexicon>
</pre>


2012-11-23 Internetlexikographie Trier – S. Drude, TLA@MPI-PL 17


```

### 3. Archivierung: Demo

Lexicon schema file (XML, archived)

```
<?xml version="1.0" encoding="UTF-8"?>
<meta id="uid:2c9090a2167c21d701168047fdce4a45" version="1.0" xmlns="http://www.mpi.nl/lexus">
  <created>2012-11-05T13:34:33+01:00</created>
  <modified>2012-11-05T13:34:33+01:00</modified>
  <schema>
    <container id="uid:2c9090a2167c21d70116804698914a01" admin-info="" note="" multiple="true" mandatory="false" type="lexicon" name="lexicon" description="">
      <container id="uid:2c9090a2167c21d70116804698914a02" admin-info="" note="" multiple="true" mandatory="false" type="lexical-entry" name="lexicalEntry" description="Represents a word, a multi-word expression, or an affix in a given language">
        <container id="uid:2c9090a2167c21d70116804698924a04" admin-info="" note="" multiple="true" mandatory="false" type="container" name="lexemeGroup" description="">
          <datacategory id="uid:2c9090a2167c21d70116804698984a3a" admin-info="http://www.isocat.org/datcat/DC-1325" note="" multiple="true" mandatory="false" type="data" name="lexeme" description="Minimal unit of language which : has a semantic interpretation and embodies a distinct cultural concept." registry="12620" reference="http://www.isocat.org/datcat/DC-1325" sort-order="uid:2c9090a21782b9b01218bfe2f4503fb/">
            <datacategory id="uid:2c9090a2167c21d70116804698984a32" admin-info="" note="" multiple="true" mandatory="false" type="data" name="part of speech" description="Term used to describe how a particular word is used in a sentence." registry="12620" reference="http://www.isocat.org/datcat/DC-1345/">
              <datacategory id="uid:2c9090a2167c21d70116804698984a34" admin-info="" note="" multiple="true" mandatory="false" type="data" name="alternative forms" description="" registry="user defined" reference="a/">
                <datacategory id="uid:2c9090a2167c21d70116804698984a30" admin-info="" note=""

```


2012-11-23 Internetlexikographie Trier – S. Drude, TLA@MPI-PL 18

### 3. Archivierung

Ideal wäre:

- Volle Versionierung (wie bei MediaWiki u.Ä.)
- Automatische regelmäßige Archivierung
- Automatische separate Archivierung aller Ressourcen (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.

2012-11-23 Internetlexikographie Trier – S. Drude, TLA@MPI-PL 19



### Archiving LEXUS 3 multimedia lexica

Sebastian Drude  
 The Language Archive - Max Planck Institute for Psycholinguistics  
 Nijmegen, The Netherlands