



# Data Preparation Tool neo

Datenaufbereitungssoftware für Archivdaten

KIM Workshop 2019  
02.04.2019, Mannheim

# Data Preparation Tool

## Motivation

- große Anzahl archivischer Datenpartner mit heterogener Datenhaltung und Erschließungspraxis
- Wunsch nach regelmäßigen Updates in DDB und Archivportal-D
- Neue Anforderungen an Datenlieferungen, die nicht immer erfüllt werden können
  - vor allem: fehlende Schnittstellen in älterer Erschließungssoftware

# Data Preparation Tool

## Zielsetzung



- Dezentralisierung der Datenaufbereitung
  - insbesondere Update-Lieferungen durch Datengeber selbst aufbereiten lassen
- Workflows in der Fachstellenarbeit zentral abbilden
- Nachnutzbarkeit erhöhen
  - Nachnutzung providerspezifischer Anpassungen durch stärkere Modularisierung
  - Einfache Einbindung neuer Ein-/Ausgangsformate
- Ziel: Archive können Updates selbständig aufbereiten
  - → mehr Kontrolle über Ihre Daten
  - → **Ergänzung** zur individuellen Unterstützung, *kein Ersatz!*

# Data Preparation Tool

## Besonderheiten von Archivdaten



Nach emotionsgeladenen Diskussionen und intensiven Beratungen stimmt die Mehrheit der Abgeordneten der Volkskammer in ihrer 30. Tagung am 23. August 1990 für den Beitritt zur Bundesrepublik nach Artikel 23 des Grundgesetzes. Bedingung ist, dass die notwendigen Voraussetzungen dafür geschaffen werden, wie zum Beispiel die Gründung der Länder oder der erfolgreiche Abschluss der Zwei-plus-vier-Verhandlungen.

Beispiel für ein archivistisches Einzelobjekt: Akte

# Data Preparation Tool

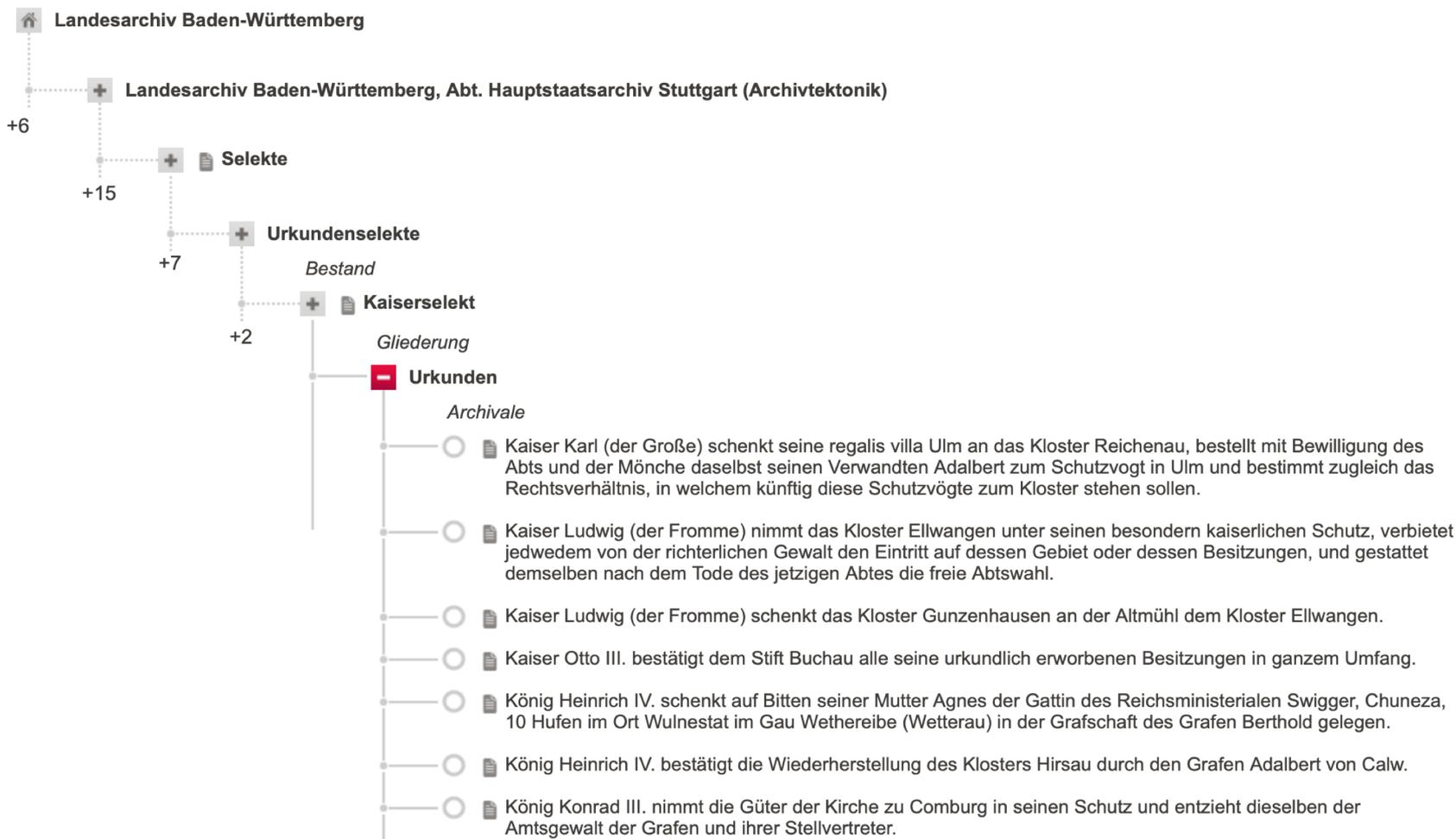
## Besonderheiten von Archivdaten



Beispiel für ein archivistisches Einzelobjekt: Urkunde

# Data Preparation Tool

## Besonderheiten von Archivdaten



Kontextualisierung eines Einzelobjekts durch übergeordnete Hierarchien

# Data Preparation Tool

## Besonderheiten von Archivdaten

Administrative Informationen

Bestandsbezogene Informationen

Gliederungsebene 1

- 
- 
- 

Gliederungsebene 1.1

- 
- 
- 

Verzeichnungseinheit 1

- 
- 
-

# Data Preparation Tool

## Konzept und Funktionalität



- **Transformation** beliebiger Quelldaten  
(vorzugsweise XML) nach EAD(DDB)
- **Vorprozessierung** für den DDB-Ingest
  - Anziehen der Binaries, OAI-Harvesting,  
Normalisierung, METS/MODS-Erzeugung zum  
Bedienen des DFG-Viewers
- **Plausibilitätsprüfung** der Datenlieferung
  - Identifier, Hierarchieverknüpfung, obsoleete Objekte
- Auszeichnung der **Rechte-** und  
**Aggregatorinformation**

# Data Preparation Tool

Data Preparation Tool neo

Datengeber (ISIL):

Liste aktualisieren

Transformationseinstellungen Analyseeinstellungen





Binaries anziehen

METS/MODS für DFGviewer generieren

Eigene XSL-Transformation einbinden

Pfad zum Stylesheet:

**Willkommen beim Data Preparation Tool!**

Um zu beginnen, legen Sie Ihre Exportdatei(en) im Ordner data\_input/\*Datengeber-ISIL\* ab und starten Sie den Prozess mit "Transformation starten".  
Weitere Informationen finden Sie im Menü "Info".

Analyse abgeschlossen.

Analyse



**Analyse abgeschlossen.**

Benötigte Zeit: 0:00:13.408851

Analyseergebnisse verfügbar

# Data Preparation Tool

Analyseergebnisse

Analyseergebnisse

 Data Preparation Tool **neo** 

## Technische Validierung

Prüfung der technischen Voraussetzungen

Damit Ihre Daten korrekt in die Portale geladen werden können, muss jedem Objekt ein eindeutiger Identifier zugeordnet sein.  
Zudem sollten die Bestandsfindbücher mit der Archivtektonik verknüpft sein, damit die hierarchische Beziehung korrekt abgebildet werden kann.

Die Ergebnisse der Prüfung finden Sie im Folgenden:

Es wurden keine Probleme gefunden.

Verknüpfung zwischen Findbuch und Tektonik  

Doppelte Identifier  

[Im Browser öffnen ...](#) Close

# Data Preparation Tool

## Konzept und Funktionalität

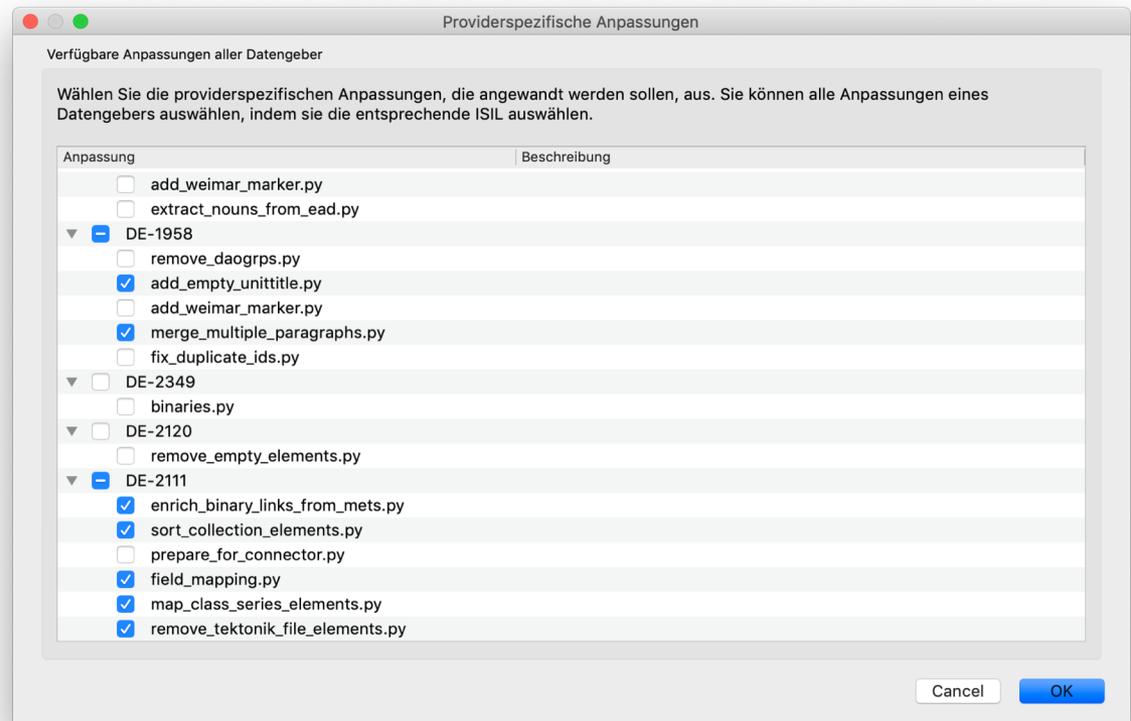


### – Providerspezifische Anpassungen

feingranulare, nachnutzbare Mappinganpassungen

kombinierbar über mehrere Datengeber hinweg

Python-Skripte bieten große Flexibilität



# Data Preparation Tool

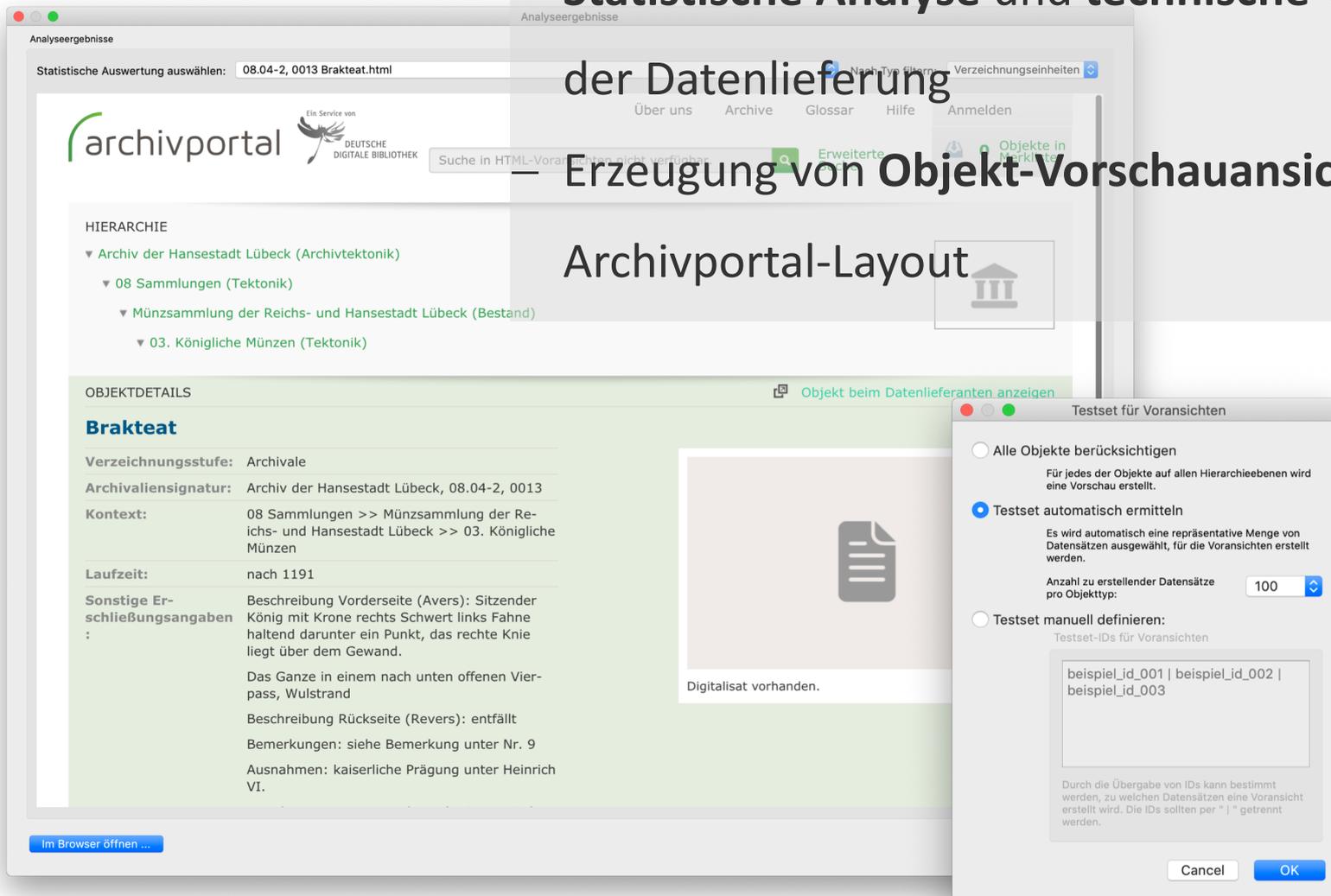
## Konzept und Funktionalität

### – Statistische Analyse und technische Validierung

der Datenlieferung

### – Erzeugung von Objekt-Vorschauansichten im

Archivportal-Layout



The screenshot displays the Archivportal interface. The main window shows search results for '08.04-2, 0013 Brakteat.html'. The left sidebar contains a hierarchy: Archiv der Hansestadt Lübeck (Archivtektonik) > 08 Sammlungen (Tektonik) > Münzsammlung der Reichs- und Hansestadt Lübeck (Bestand) > 03. Königliche Münzen (Tektonik). The main content area shows details for 'Brakteat', including its archival level, signature, context, and description. A preview icon is visible, and a dialog box titled 'Testset für Vorsichten' is open in the foreground.

**OBJEKTDDETAILS**

**Brakteat**

Verzeichnungsstufe: Archivale

Archivsignatur: Archiv der Hansestadt Lübeck, 08.04-2, 0013

Kontext: 08 Sammlungen >> Münzsammlung der Reichs- und Hansestadt Lübeck >> 03. Königliche Münzen

Laufzeit: nach 1191

Sonstige Erschließungsangaben:

Beschreibung Vorderseite (Avers): Sitzender König mit Krone rechts Schwert links Fahne haltend darunter ein Punkt, das rechte Knie liegt über dem Gewand.

Das Ganze in einem nach unten offenen Vierpass, Wulstrand

Beschreibung Rückseite (Revers): entfällt

Bemerkungen: siehe Bemerkung unter Nr. 9

Ausnahmen: kaiserliche Prägung unter Heinrich VI.

Im Browser öffnen ...

**Testset für Vorsichten**

Alle Objekte berücksichtigen  
Für jedes der Objekte auf allen Hierarchieebenen wird eine Vorschau erstellt.

Testset automatisch ermitteln  
Es wird automatisch eine repräsentative Menge von Datensätzen ausgewählt, für die Vorsichten erstellt werden.

Anzahl zu erstellender Datensätze pro Objekttyp: 100

Testset manuell definieren:  
Testset-IDs für Vorsichten

beispiel\_id\_001 | beispiel\_id\_002 |  
beispiel\_id\_003

Durch die Übergabe von IDs kann bestimmt werden, zu welchen Datensätzen eine Vorsicht erstellt wird. Die IDs sollten per " | " getrennt werden.

Cancel OK

# Data Preparation Tool

Analyseergebnisse

Analyseergebnisse

Data Preparation Tool **neo**

## Metadatenauswertung

Datenqualität & statistische Auswertung der Datenlieferung

Datenqualität

Ergebnisse, die für die Übernahme in Portale problematisch sein können, werden rot hinterlegt; solche, bei denen eine Verbesserung der Datenqualität möglich, jedoch nicht zwingend notwendig ist, werden hingegen gelb hinterlegt.

Weitere Punkte werden in späteren Versionen des Tools ergänzt.

Es wurde eines oder mehrere Probleme der Datenqualität gefunden, die sich auf die technische Validität der Daten auswirken.

Objekte ohne Titel 1014 ▾

Objekte mit Platzhaltertitel 0 ▾

Durchschnittliche Anzahl Erschließungsfelder pro VZE 0 ▾

Im Browser öffnen ...

Close

# Data Preparation Tool

Analyseergebnisse

Statistische Analyse (Findbuch)

Hierarchieebenen

Anzahl Objekte insgesamt	112235
Anzahl Objekte mit eigenem DDB-View 	106539
Objekte auf Verzeichnungsebene	105450
Objekte auf Teilebene	0
Gliederungsstufen (Klassifikationsebene)	5501
Gliederungsstufen (Serienebene)	1284

Binärinhalt

Anzahl von Digitalisaten	1164
Anzahl Objekte mit Digitalisat(en)	225

Medientyp (normalisiert)

[Im Browser öffnen ...](#) Close

# Data Preparation Tool

Analyseergebnisse

Statistische Analyse (Findbuch)

Hierarchieebenen

**Anzahl Objekte** 112235

**Anzahl Objekte mit eigenem DDB-View** ⓘ 106539

**Objekte auf Verzeichnungsebene** 105450

**Objekte auf Teilebene** 0

**Gliederungsstufen (Klassifikationsebene)** 5501

**Gliederungsstufen (Serienebene)** 1284

Binärinhalt

**Anzahl von Digitalisaten** 1164

**Anzahl Objekte mit Digitalisat(en)** 225

Medientyp (normalisiert)

Im Browser öffnen ...

Close

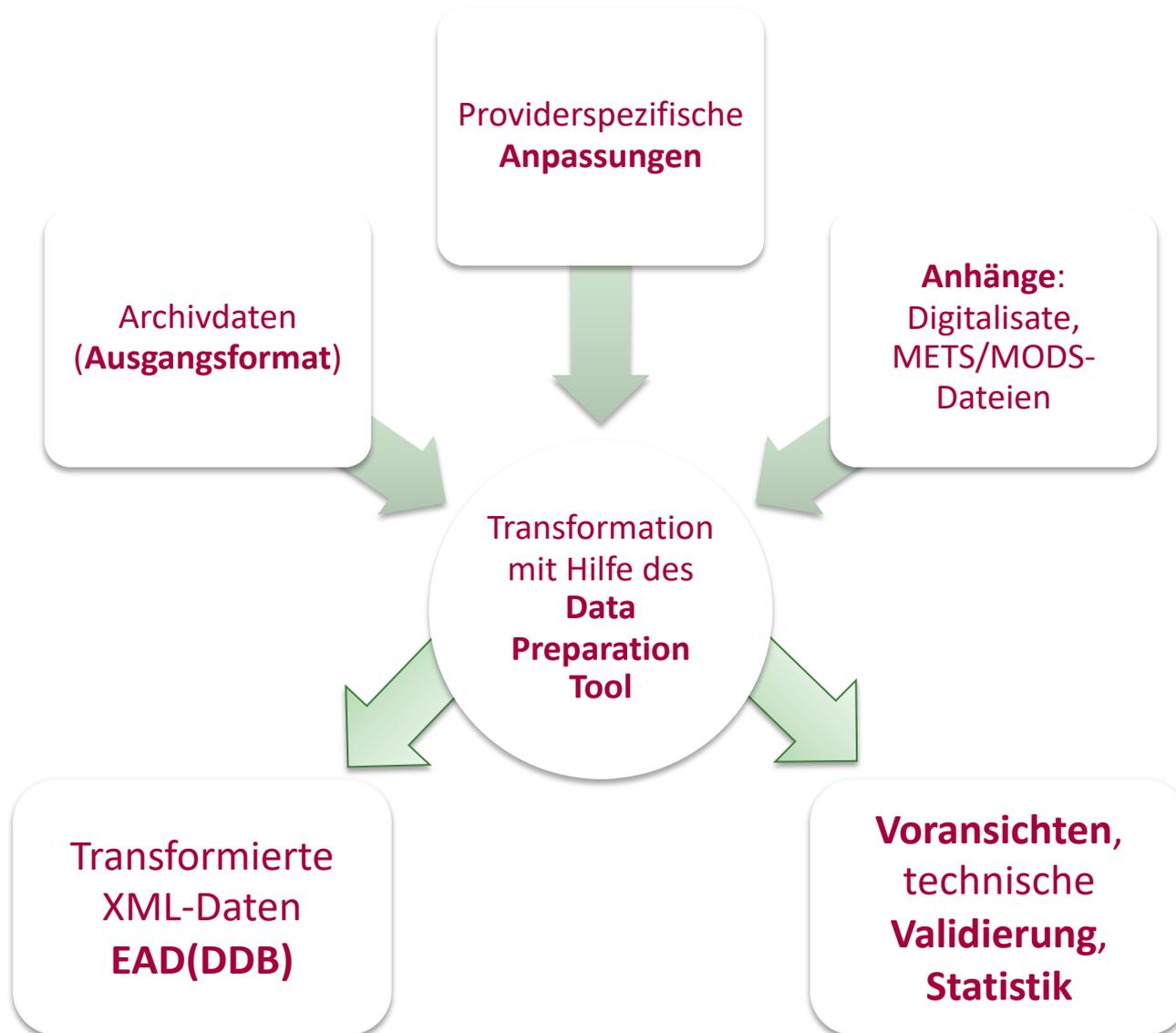
Dies umfasst diejenigen Objekte, die im Portal eine eigene Objektseite erhalten. Dazu zählen alle Objekte auf Bestands-, Verzeichnungs- und Vorgangsebene. Zusätzlich werden Objekte auf Gliederungs- und Serienebene berücksichtigt, wenn diese eine Kurzbeschreibung (Element 'abstract') besitzen.

Kategorie	Anzahl
Anzahl Objekte	112235
Anzahl Objekte mit eigenem DDB-View ⓘ	106539
Objekte auf Verzeichnungsebene	105450
Objekte auf Teilebene	0
Gliederungsstufen (Klassifikationsebene)	5501
Gliederungsstufen (Serienebene)	1284

Kategorie	Anzahl
Anzahl von Digitalisaten	1164
Anzahl Objekte mit Digitalisat(en)	225

# Data Preparation Tool

## Konzept und Funktionalität



# Data Preparation Tool

Open-Source-Technologien



lxml



Requests  
*http for humans*



# Data Preparation Tool

Cross-Platform



**Analyseergebnisse**

Data Preparation Tool **neo**

## Metadatenauswertung

Datenqualität & statistische Auswertung der Datenlieferung

Datenqualität

Ergebnisse, die für die Übernahme in Portale problematisch sein können, werden rot hinterlegt, wenn eine Verbesserung der Datenqualität möglich, jedoch nicht zwingend notwendig ist, weiter hinterlegt.

Weitere Punkte werden in späteren Versionen des Tools ergänzt.

Es wurde eines oder mehrere Probleme der Datenqualität gefunden, die jedoch keine Fehler sind, sondern als Anregung zu verstehen sind.

Objekte ohne Titel: 0 | Objekte mit Platzhaltertitel: 8

Durchschnittliche Anzahl Erschließungsfelder pro VZE: 0

Statistische Analyse (Findbuch)

Hierarchieebenen

Anzahl Objekte insgesamt

Anzahl Objekte mit eigenem DDB-View

Im Browser öffnen ...

**Analyse**

Analyse abgeschlossen.

100%

Benötigte Zeit: 0:00:00.288022

Analyseergebnisse verfügbar

Statistische Auswertung

Technische Validierung

Vorsichten

Close

**Willkommen beim Data Preparation Tool!**

Um zu beginnen, legen Sie Ihre Exportdatei(en) im Ordner data in den Ordner "data\_input/\*Datengeber-ISIL\*" ab und starten Sie den Prozess mit "Transformation starten". Weitere Informationen finden Sie im Menü "Info".

## Visuelle Validierung

Verbesserung der Persistenz von Objekt-Identifiern

Unterstützung der „großen“ EAD-Version 3.0

# Data Preparation Tool

Weiterentwicklung: Visuelle Validierung



- Erweiterung durch eine **visuelle Validierung**
  - analog zu XML-Schema-Validierung, aber ...
    - menschenlesbare Fehlermeldungen
    - Gruppierung der Ergebnisse nach Priorität
    - Zusammenfassung der Meldungen zu einer übergreifenden Aussage zur Verwendbarkeit der Datenlieferung
  - Zielgruppe: Fachstelle, Datengeber, Aggregatoren und Softwarehersteller

# Data Preparation Tool

Weiterentwicklung: Stabile Identifier



- Problematik der nicht immer gegebenen **stabilen Identifier** im Archivbereich, jedoch Voraussetzung für geplante Projekte
- Umsetzung einer **Objekt-ID-Konkordanz** bereits seit längerem ein Desiderat
- Ziel: Rückführung geänderter IDs auf Ihre ursprüngliche ID, wenn jeweiliges Objekt schon einmal geliefert
- Methode: Gewichteter Abgleich über Feldinhalte und Kontext

# Data Preparation Tool

Weiterentwicklung: Stabile Identifier

```
<c level="file" id="A91x60765755176544120161108133732140">
  <did>
    <unitid>1</unitid>
    <unittitle>Abel, Fritz</unittitle>
    <unitdate normal="1948-01-01/1954-12-31">01.01.1948 -
      31.12.1954</unitdate>
    <abstract type="Enthält">Enthält u.a.: Personenstandsunterlagen,
      Arztabrechnungen der Ehefrau</abstract>
    <physdesc><extent>1 cm</extent></physdesc>
  </did>
  <daogrp>
    <daodesc>
      <list>
        <item><name>B2_1</name><genreform>SONSTIGES</genreform></item>
      </list>
    </daodesc>
  <daoloc xlink:role="image_full" xlink:href="bilder/B2_1.pdf"/></daogrp>
</c>
```

Für Hash-Abgleich relevante EAD-Felder  
(u.a.)

Beispiel: Verzeichnungsebene

# Data Preparation Tool

Weiterentwicklung: Unterstützung für EAD 3.0



- Direkte Unterstützung des internationalen Standards **EAD 3.0**
  - EAD2002 bereits neben EAD(DDB) unterstützt
  - Herausforderung: sehr flexible Nutzung und hierarchische Verschachtelung von Elementen, offene Granularität im Vergleich zu EAD(DDB)
  - Lösung: Connector mit einem grundlegenden Mapping, welches sukzessive erweitert wird, wenn neue Anforderungen kommen
    - so bereits für Lieferungen in EAD2002 umgesetzt

# Data Preparation Tool

## Ausblick



- Zeitnahe Veröffentlichung als Open Source
- Bereitstellung insbesondere für ...
  - Datengeber zur Aufbereitung und Validierung Ihrer Daten
  - Aggregatoren und Softwarehersteller zur Unterstützung der Schnittstellenentwicklung
- Technische sowie Benutzerdokumentation
- Ermöglicht sukzessive Anpassung an neue Anforderungen

# Data Preparation Tool

## Fazit



- Die Lieferung an Portale stellt für viele Einrichtungen nach wie vor eine Herausforderung dar, da ...
  - ... sich die **heterogene Erschließung** nicht immer verlustfrei in den Portalen abbilden lässt
  - ... die Datengeber stark von den Möglichkeiten der genutzten **Softwareinfrastruktur** abhängig sind
  - ... die Nutzung weitergehender Features (z.B. die Verwendung des DFG-Viewers) z.T. eine Überarbeitung der eigenen Infrastruktur voraussetzt

- Die **komplexe Struktur** von EAD, insbesondere der hierarchische Aufbau, bietet viel Potenzial für Fehler und setzt eine **Persistenz** der Datenhaltung voraus, die bei den Einrichtungen nicht immer gegeben ist
- Teilnahme an Portalen soll möglichst vielen Archiven ermöglicht werden
- Data Preparation Tool als nachhaltige Lösung zur Konsolidierung sowie Evaluation/Validierung von Exportdaten



Vielen Dank  
für Ihre Aufmerksamkeit.  
... Fragen?

**Oliver Götze**

Landesarchiv Baden-Württemberg

DDB-Fachstelle Archiv / Archivportal-D

[oliver.goetze@la-bw.de](mailto:oliver.goetze@la-bw.de) / [archiv@deutsche-digitale-bibliothek.de](mailto:archiv@deutsche-digitale-bibliothek.de)

# Bildnachweis

Python-Logo. Python Software Foundation. (Folie 18)

<https://www.python.org/community/logos/>

lxml-Logo. Stefan Behnel: Implementing XML languages with lxml. (Folie 18)

<https://lxml.de/s5/lxml-ep2008.html>

requests-Logo. (Folie 18)

<http://docs.python-requests.org/en/master/>

Qt-Logo. Qt Project. (Folie 18)

[https://de.wikipedia.org/wiki/Datei:Qt\\_logo\\_2016.svg](https://de.wikipedia.org/wiki/Datei:Qt_logo_2016.svg)

Bulma Logo. (Folie 18)

<https://github.com/jgthms/bulma>