DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

MAPS

Workflow-Tool zur Validierung und Transformation von XML Daten

Karl-Ulrich Becker, Timo Schleier (SUB Göttingen)

KIM Workshop 2019, Mannheim, 2.+3. April

Aufgaben der Fachstelle Bibliothek (METS/Mods)

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Ansprechpartner für die Datenpartner zu Fragen zum Datenformat
- Harvesten von METS/Mods-Daten über OAI
- Validierung und Reporterstellung
- Bereinigung und Anreicherung von Datensätzen
- Lieferung der Datensätze an die DDB
- Erstellung und Pflege des konzeptionellen Mappings für den Ingest in die DDB

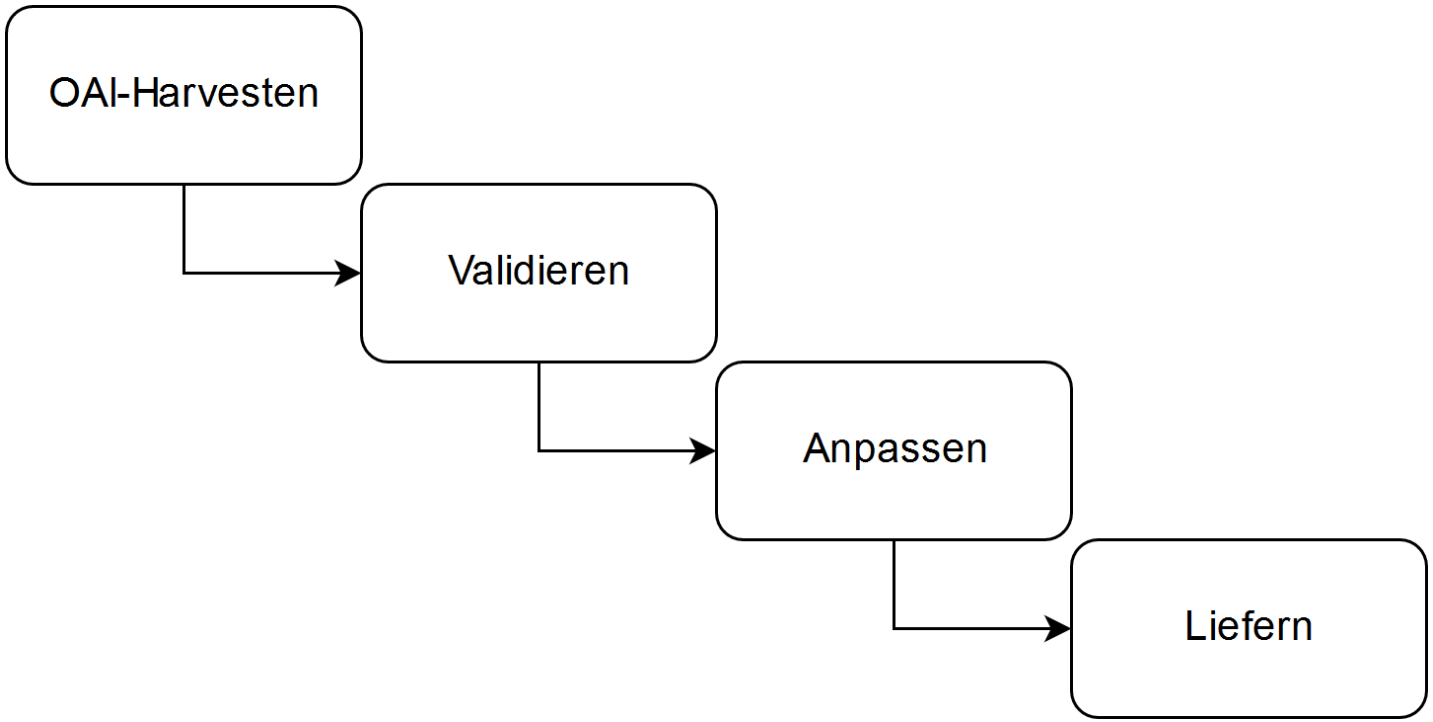
→ Viel "Handarbeit", geht das nicht besser?

Die MAPS-Idee

Anforderungen an MAPS

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

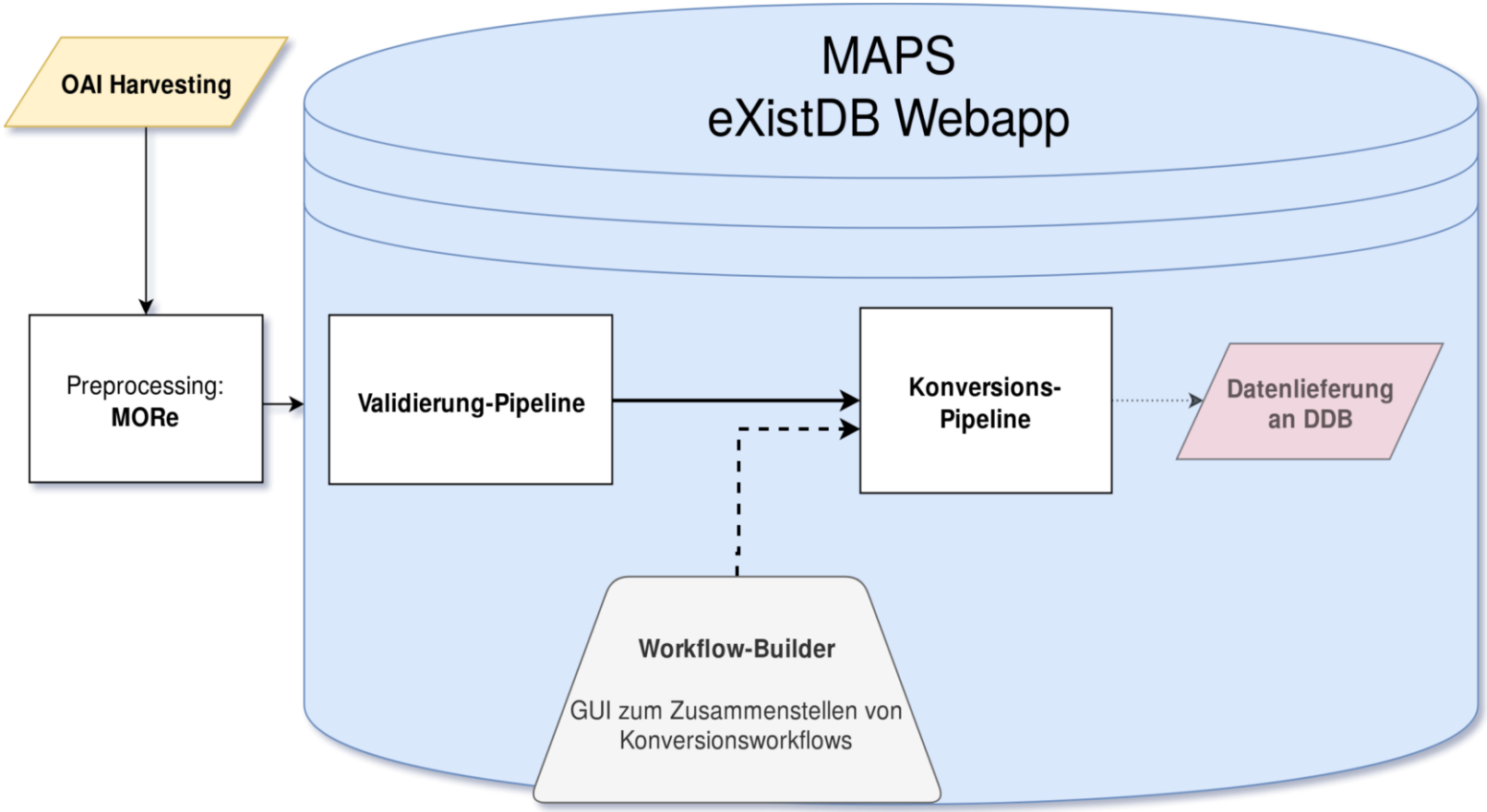
- Datenaggregation und Lieferung so weit automatisiert wie möglich und so flexibel wie nötig
- GUI statt Kommandozeile
- Datenbank statt Dateisystem



Das MAPS-System

MAPS

- Das MAPS-System besteht aus einer eXist-XML Datenbank, einer Webapp und einem externen Vorbereitungs-Tool
- Eingesetzte Technik: xQuery/XSLT/JavaScript/LaTeX



OAI Harvesting

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

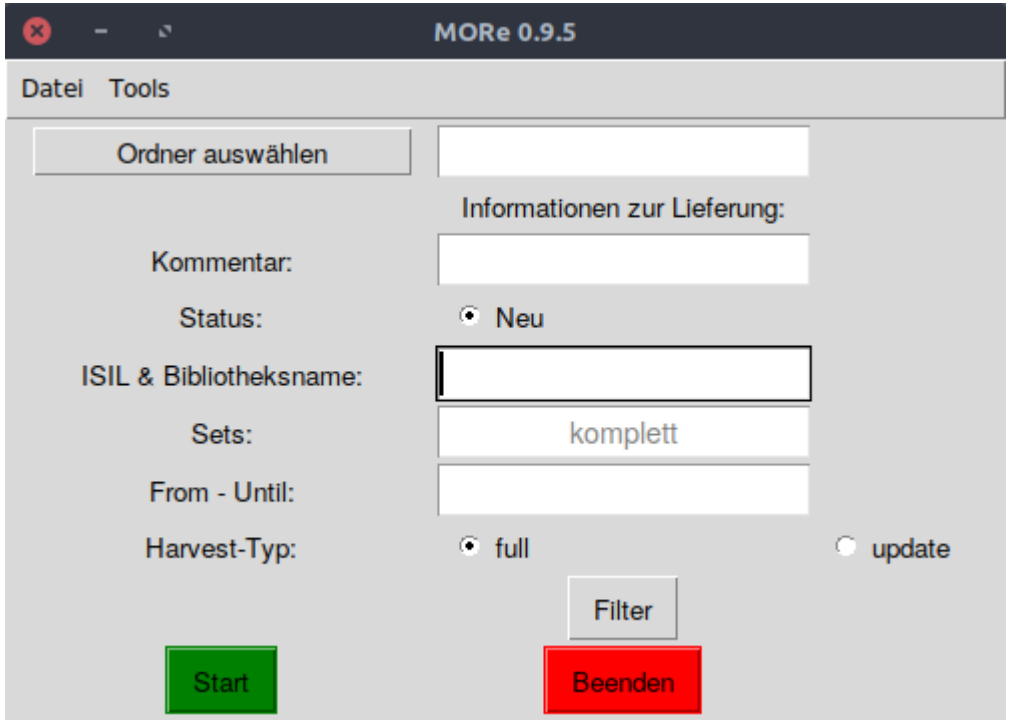
- Harvesting schwierig: Keine zufriedenstellende Lösung (für unsere Daten)
- Vorgehen Aktuell: Harvesting mit an der SUB enwickelten Tools (erst per ListRecords, dann harvesting einzelner Dateien per GetRecord)
- z. T. spezielle Anpassungen an einzelne Schnittstellen

→ **Ziel für die Zukunft von MAPS**: OAI Harvesting (auch für große Datensätze) zuverlässig umsetzen (und in Infrastruktur einbinden)

Preprocessing: MORe

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Renamer & Ingesttool für geharvestete XML Dateien
- geschrieben in Python, mit GUI (plattformunabhängig)



MORe: Funktionen

1. Benennt geharvestete METS-Dateien konsistent um, zBsp.:

   ```
   DE-3--oai_digitale_bibliothek_uni-halle_de_9994545.xml
   ```

   → Nutzt dafür Informationen aus den Dateien und eine Konkordanzliste für die Zuordnung dv:owner zu ISIL
2. Erstellt eine XML Datei, die Informationen über die aktuell bearbeiteten Dateien enthält → "Lieferungs-XML"
3. Ingestiert in die MAPS-Webapp

In Entwicklung:

- Filterfunktion (Daten nach bestimmten Kriterien schon vor dem Ingest aussortieren)

Das Lieferungs-Konzept

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Identifiziert eine Datenlieferung und alle dazugehörigen Dateien
- Enthält Informationen zum Bearbeitungsstand und zu der Schnittstelle, den geharvesteten Sets usw.

```
<lieferung xmlns="dcg:maps" id="20190307_DE-16_Heidelberg_UB-komplett">
 <status>validiert</status>
 <sets>komplett</sets>
 <oaiUrl>http://digi.ub.uni-heidelberg.de/cgi-bin/digioai.cgi</oaiUrl>
 <fromUntil/>
 <responseDate>2019-01-28T11:23:18Z</responseDate>
 <harvestTyp>full</harvestTyp>
 <creationDate>2019-03-07T06:18:11Z</creationDate>
 <comment>test</comment>
 <supplier>DE-16_Heidelberg_UB</supplier>
 <files>
  <file fatal="no">DE-16--oai_digi_ub_uni-heidelberg_de_10.xml</file>
  <file fatal="yes">DE-16--oai_digi_ub_uni-heidelberg_de_100.xml</file>
  <file fatal="no">DE-16--oai_digi_ub_uni-heidelberg_de_1000.xml</file>
  <file fatal="no">DE-16--oai_digi_ub_uni-heidelberg_de_1001.xml</file>
  <file fatal="no">DE-16--oai_digi_ub_uni-heidelberg_de_1002.xml</file>
  <file fatal="no">DE-16--oai_digi_ub_uni-heidelberg_de_10020.xml</file>
  <file fatal="yes">DE-16--oai_digi_ub_uni-heidelberg_de_10022.xml</file>
  <file fatal="yes">DE-16--oai_digi_ub_uni-heidelberg_de_10023.xml</file>
 </files>
</lieferung>
```

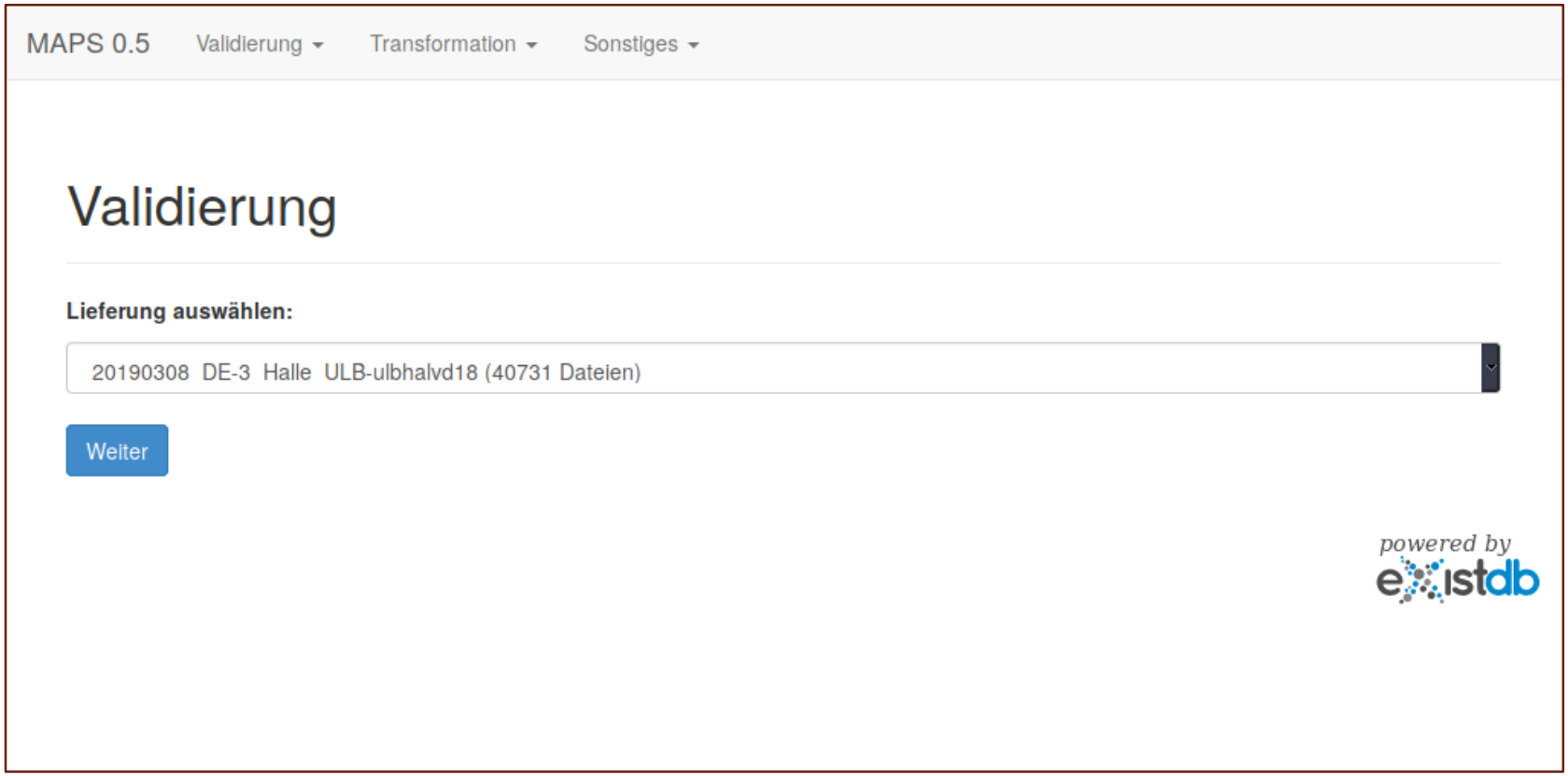
Validierung & Reporterzeugung
Anforderungen

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Automatische Generierung des versandfertigen Reports
- Filtern von Datensätzen mit kritischen Fehlern
- Prüfung aller Datensätze

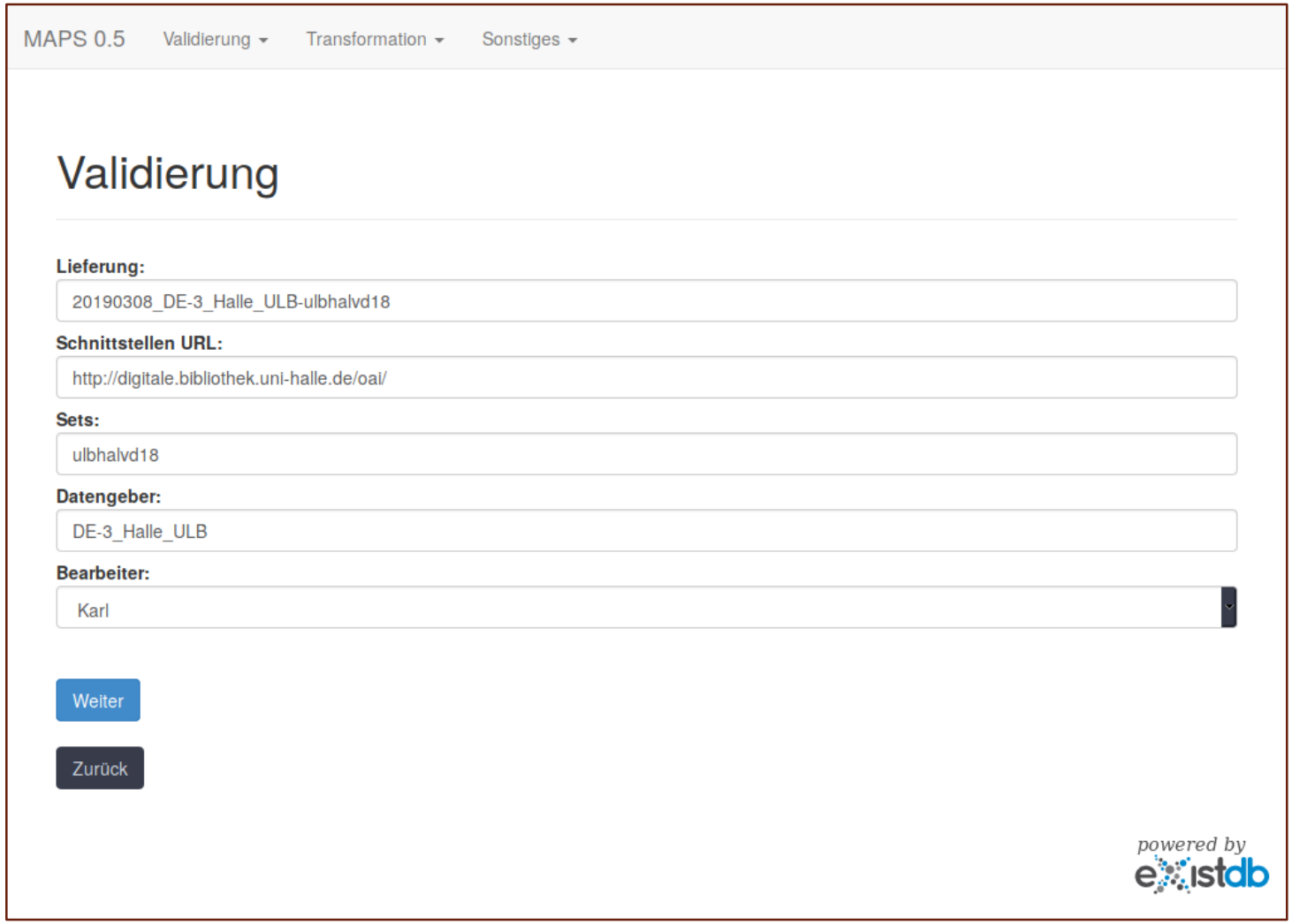
Validierung & Reporterzeugung
Screenshot: Validierung starten

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online



Validierung & Reporterzeugung
Screenshot: Validierung starten

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online



Validierung & Reporterzeugung
Screenshot: Report

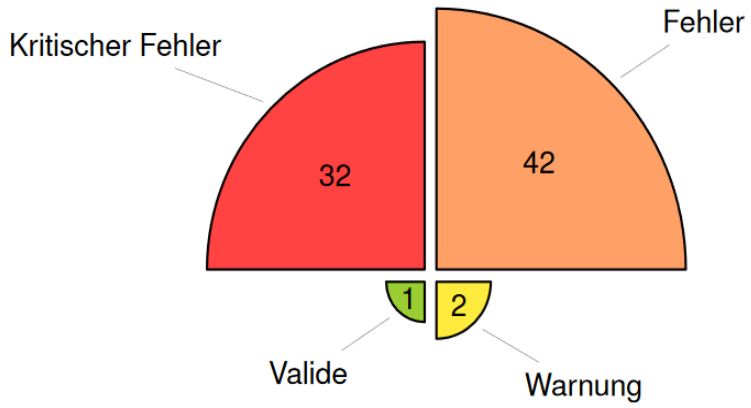
DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

1 Übersicht

Sehr geehrte Damen und Herren,
hiermit erhalten Sie den aktuellen Analysereport Ihrer Daten. Bitte beachten Sie die aufgelisteten Fehler unter Punkt 2 und korrigieren Sie ggf. Ihre Daten, um sie korrekt und vollständig in der DDB anzeigen zu können.

1.1 Analysierte Datensätze

- Von der OAI-Schnittstelle wurden 77 Datensätze geharvestet.
- Es wurde folgendes Set geharvestet:
  ("komplett" bedeutet, dass keine Eingrenzung auf ein Set stattgefunden hat)



Validierung & Reporterzeugung
Screenshot: Report

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

2.8 Kein gültiger Wert im Attribut authority in `mods:classification`

`mods:classification` wird in der DDB nur dann berücksichtigt, wenn es sich um Klassen aus der DDC oder der Systematik der ZDB handelt. Um dies zu kennzeichnen, muss in dem Attribut authority entweder "ddc" oder "zdbs" stehen. Ist dies nicht der Fall, wird `mods :classificaiton` bei der Transformation der Daten entfernt. Die folgenden ungültigen Werte

wurden verwendet:

- ZVDD (27)
- GDZ (1)

Dieser Fehler tritt 29 mal auf:

- `oai:gdz.sub.uni-goettingen.de:PPN8876915123 mets:dmdSec = DMDLOG_0001`
- `oai:gdz.sub.uni-goettingen.de:PPN88767769151 mets:dmdSec = DMDLOG_0001`
- `oai:gdz.sub.uni-goettingen.de:PPN8845769151 mets:dmdSec = DMDLOG_0001`

→ *Zur vollständigen Liste*

Die Datensätze werden von der Fachstelle Bibliothek vor dem Einspielen korrigiert

2.9 `mets:dmdSec` fehlt

Die `mets:dmdSec` enthält die bibliographische Beschreibung des Werks. Innerhalb eines METS-Datensatzes muss es mindestens eine `mets:dmdSec` geben, die MODS-Daten enthält. Ist dies nicht der Fall, wird der Datensatz nicht in die DDB eingespielt.

Dieser Fehler tritt 4 mal auf:

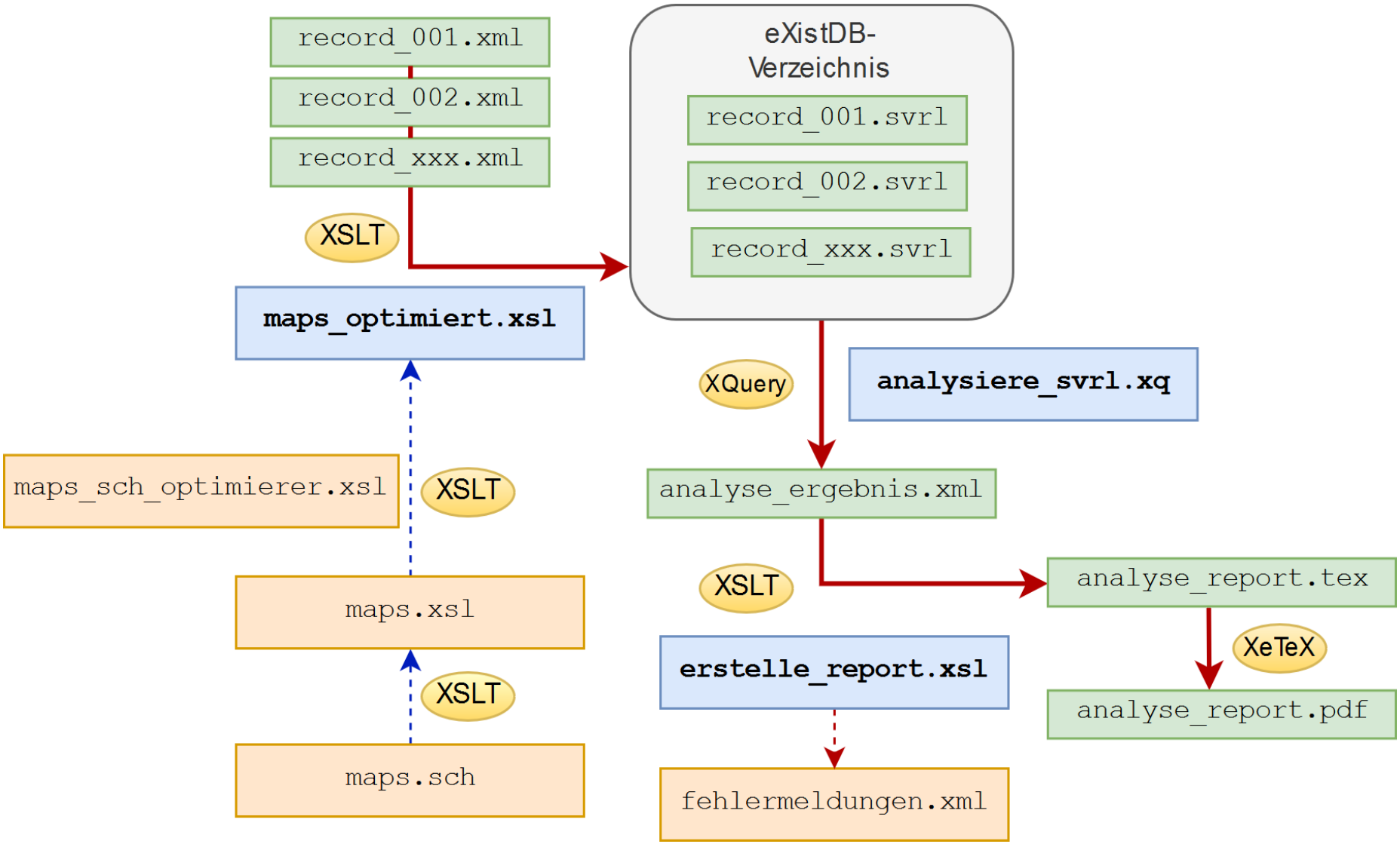
- `oai:gdz.sub.uni-goettingen.de:PPN8875569151 mets:dmdSec`
- `oai:gdz.sub.uni-goettingen.de:PPN887691513 mets:dmdSec`

→ *Zur vollständigen Liste*

Die Datensätze werden **nicht** eingespielt

Validierung & Reporterzeugung
eXistDB-Pipeline

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online



MAPS-Module & MAPS-Workflows

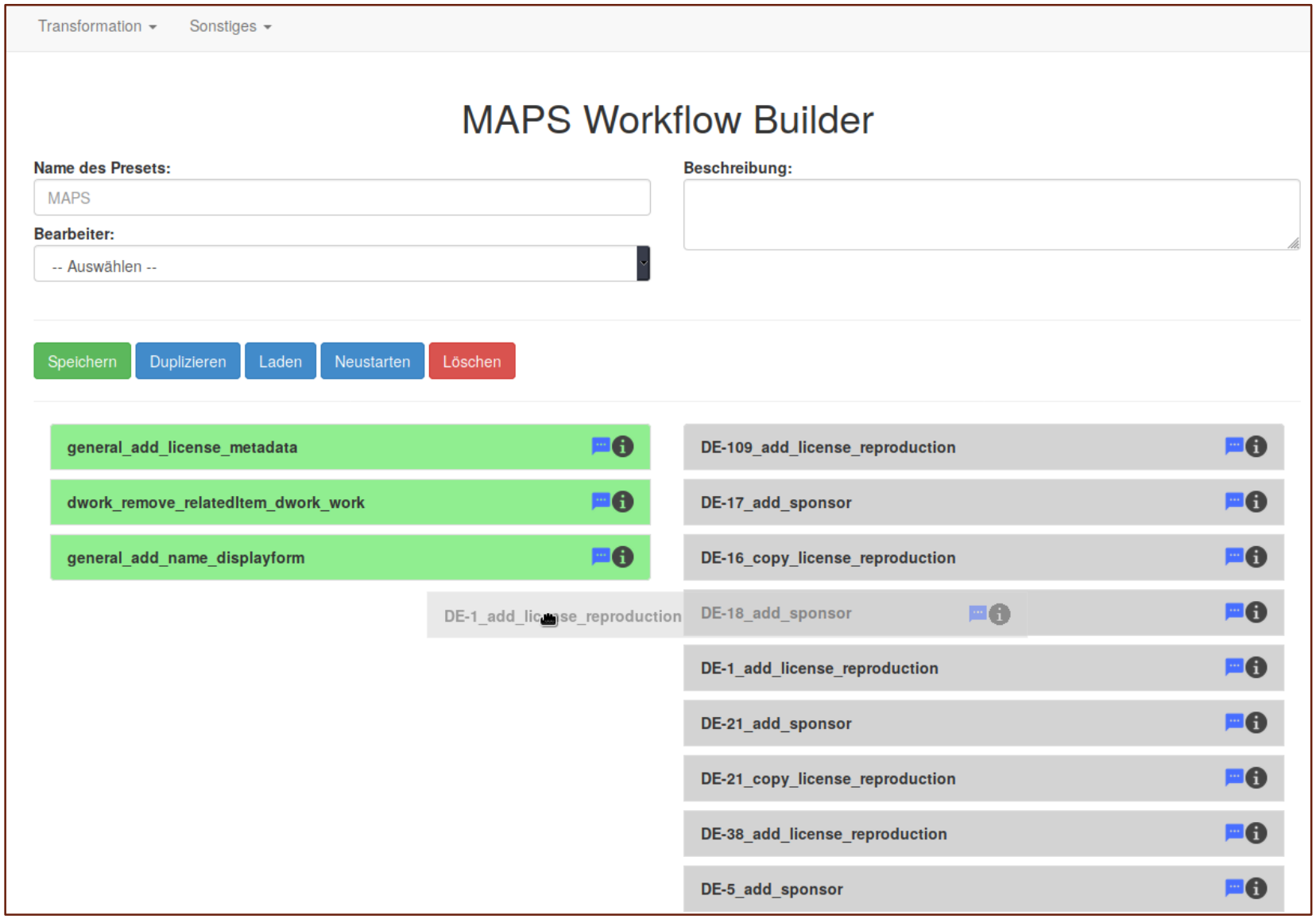
DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- "Baukasten für die Anpassung von Datensätzen"
- ein MAPS-Modul...
  - ist ein XSL-Script
  - führt eine möglichst granulare Anpassung durch
  - ist ausführlich dokumentiert
    - zur Nutzung im Workflow Builder
    - "einfach" und "technisch"
    - Abhängigkeiten
    - Typisierungen
- ein MAPS-Workflow...
  - fasst MAPS-Module geordnet zusammen
  - ist wiederverwendbar und leicht anpassbar

→ Vereinfachung des konzeptionellen Mappings durch teilweise Verlagerung des technischen Mappings zur Fachstelle

MAPS-Module & MAPS-Workflows
Screenshot: Workflow Builder

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online



Lieferung an die DDB

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Validierung vor der Lieferung
- Ziel: Validierte und bereinigte Initial- und Updatelieferungen aus MAPS heraus an die DDB senden
- Redundante Datenhaltung immer des aktuellsten Datenbestands
- Ausblick: DDB holt sich die Daten über unsere zukünftige OAI-Schnittstelle

Zu guter Letzt...

Lessons Learned...

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Warum eXistDB?
  - Beschränkung auf 2 Milliarden nodes in BaseX
  - Einfache Umsetzung von Webapps
- eXistDB Eigenheiten
  - Umständliche Systemupdates (dafür gute App-Updates)
  - mittelmäßiges Taskmanagement
  - Anpassen von Indizes schwierig
  - in der Anfangsphase durchaus viele Abstürze
  - schwieriges Debugging
  - XSLT Unterschiede zu "Saxon"
  - Performance: Mittelmäßig
- (zu) große Tex-Dateien
- Konzeption der Dokumentation der MAPS-Module

Die wichtigste Frage zum Schluss: Was bedeuten eigentlich MAPS & MORe?

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- **M**etadata **A**ggregation and **P**rocessing **S**ystem

- **M**ETS **O**AI **Re**namer

Vielen Dank für die Aufmerksamkeit