LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Kontrolle der Vollständigkeit bei Wiley DEAL Journals

Peter Eisner
16. März 2022
AUdS-Tagung 2022

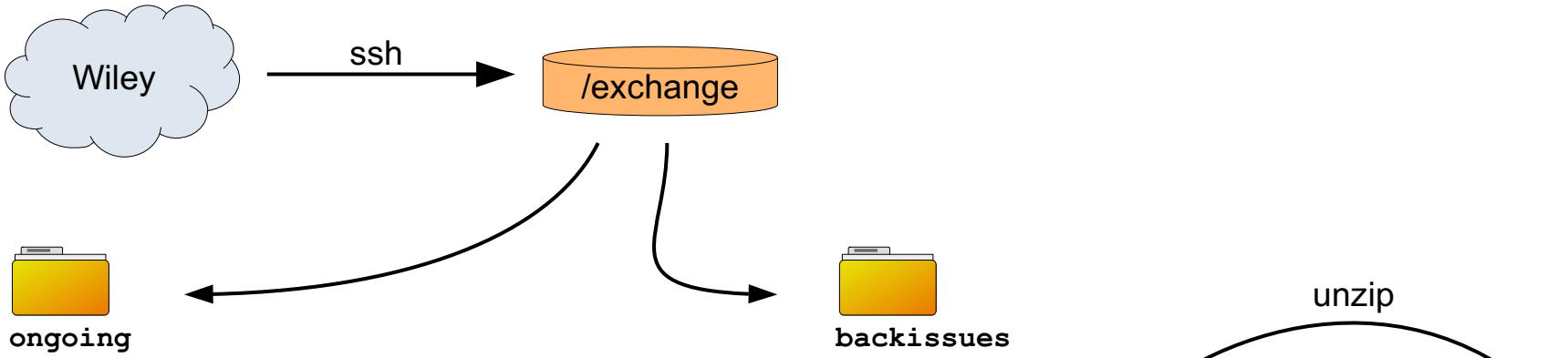
Leibniz Gemeinschaft

DEAL: Wiley

Überblick

- DEAL-Vereinbarung beinhaltet digitale Langzeitarchierung
- „Dark Archive Responsible Entity“ ist die TIB
- Team Langzeitarchivierung erhält riesige Datenmenge
- mehr als 2000 e-Journals
- mehr als 5 000 000 Artikel (1 Artikel = 1 IE im Archiv)
- mehr als 50 000 000 Dateien
- bisher im Archiv, ohne Wiley: ca. 150 000 IEs, ca. 2 500 000 Dateien
- neue Tools (Python, Bash, XSLT)
- neue Teil-Workflows (Pre-Ingest)
- Anpassung vorhandener Workflows

Datenlieferung Wiley DEAL



- 22054 ZIPs
- Aktuelles und Altbestände
- Lieferung nicht abgeschlossen
- derzeit nicht in Bearbeitung

- 290114 ZIPs
- Altbestände (1997–2021)
- Lieferung abgeschlossen
- Journal Inventory
- Issue Inventory

- 1 ZIP = 1 Issue
- 2391 Journals
- ca. 5 300 000 Artikel (IEs)
- ca. 52 000 000 Dateien
- ca. 11 Terabyte

Vollständigkeit als Vergleich

Beispiel:
Drei grüne Sofas

Anzahl: 3		✓	Anzahl: 3	**Quantität**
Farbe: grün	✓	✓	Farbe: grün	**Qualität**
grün				
nicht grün		✗		

Vollständigkeit als Vergleich

Datenlieferung
(Ist-Zustand)

- 1 ZIP = 1 Issue
- 1 Artikel = 1 Unterordner
- Dateiname nach Schema:
  JGC_ISSN_DATE_VOL_ISSUE
- Jeder Artikel enthält XML mit Metadaten
- in den meisten Fällen auch PDF
- in vielen Fällen Unterordner (Anhänge, Ressourcen)

Elemente im Wiley Journal Inventory
(Informationen zum Soll-Zustand)

- Journal group code (JGC)
- publication_title
- print_identifier (ISSN)
- date_first_issue_ online
- num_first_vol_online
- num_first_issue_ online
- date_last_issue_online
- num_last_vol_online
- num_last_issue_online
- number of issues
- number of articles
- number of PDF
- number without PDF

Wie zählt man PDF-Repräsentationen?

PDF-Repräsentation eines Artikels müssen als solche erkannt werden, damit sie gezählt werden können.

Artikel-Ordner
PDF-Repräsentation
XML-Repräsentation
Unterordner
Grafik 1
Grafik 2

```
APA13048
├── apa13048.pdf
├── apa13048.xml
└── image_n
    ├── apa13048-fig-0001.png
    └── apa13048-toc-0001.png
```

Annahme:
PDF-Datei im Article-Root
ist die PDF-Repräsentation
des Artikels → Zahlen stimmen nicht

```
APA13084
├── apa13084_am.pdf
├── apa13084.pdf
├── apa13084.xml
└── image_n
    ├── apa13084-fig-0001.png
    ├── apa13084-fig-0002.png
    ├── apa13084-fig-0003.png
    ├── apa13084-fig-0004.png
    └── apa13084-fig-0005.png
```

?

Neue Annahme:
PDF-Datei im Article-Root
ist die PDF-Repräsentation
des Artikels, wenn der
Dateiname der XML-Datei
entspricht ...genügt das?

Globaler und detaillierter Abgleich

Skript: Wiley Dealer

I Globaler Abgleich

Ziele:

- Grobe Einschätzung der Lieferung „Backissues“
- Identifizierung von Journals, die wahrscheinlich fehlerfrei sind (für den Ingest)

Begrenzter Detailgrad:

- nur einige Metadaten (Anzahl Artikel pro ZS, Datum,...)
- keine Auswertung der Artikel-XMLs

Einbezogene Datenquellen:

- Dateinamen der Lieferung
- Journal Inventory

II Detaillierter Abgleich

Ziele:

- kein Ingest von fehlerhaften Journals
- Aufspüren von Fehlern zur Rückmeldung an Wiley

Begrenzter Umfang:

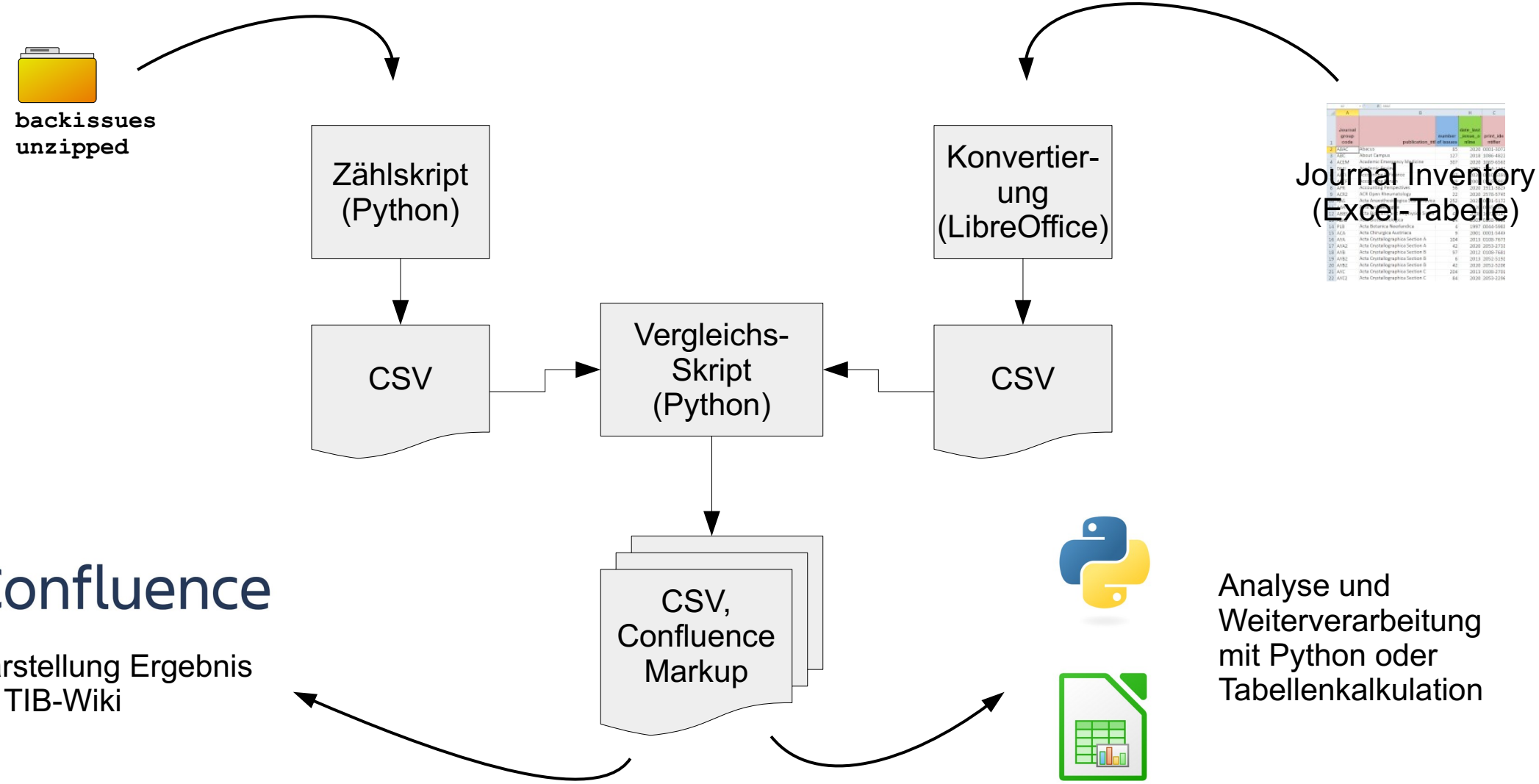
- Teil des Workflows, wird pro Zeitschrift ausgeführt

Einbezogene Datenquellen:

- Dateinamen der Lieferung
- Artikel-XMLs
- Journal Inventory
- Issue Inventory

Skript: Ingest Material Provider

I Globaler Abgleich · Prozess

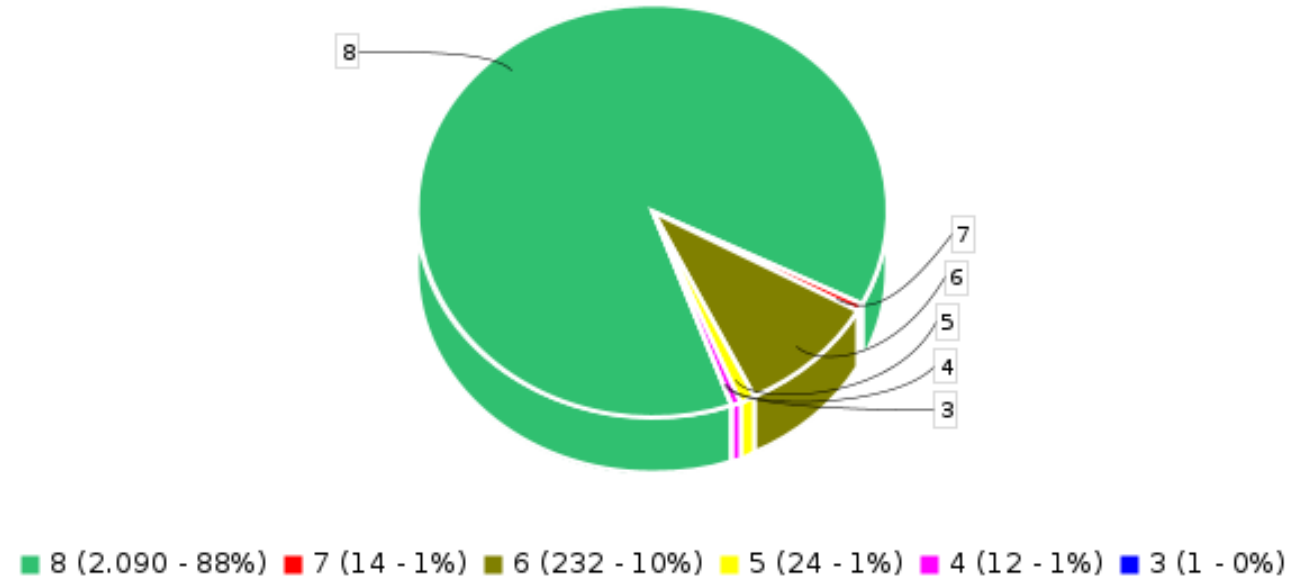


I Globaler Abgleich · Ergebnis

Match Level

Der Match Level ist ein Zähler, der bei einem erfolgreich abgeglichenen Wert hochgezählt wurde. Er ist ein Indikator für den Grad der Übereinstimmung eines Journals. Da nur je 8 Werte verglichen wurden bedeutet ein Match-Level von 8 die maximale mögliche Übereinstimmung innerhalb dieses Abgleichs. Abweichungen jenseits des Abgleichs – zum Beispiel bei den Issues – sind theoretisch möglich und müssen beim Pre-Ingest geprüft werden.

Verteilung Journals pro Match Level



2373 Zeitschriften berücksichtigt
(Screenshot aus internem Wiki)

Abgeglichene Daten

- Journal Group Code
- Print ISSN
- Date First Issue
- Date Last Issue
- Number of Issues
- Number of Articles
- Number of Articles with PDF-Representation
- Number of Articles without PDF-Representation

= 8 Elemente

II Detaillierter Abgleich

IMP · Ingest Material Provider

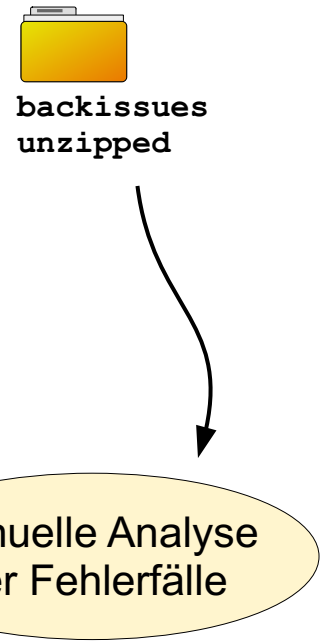
Features

- Abgleich verschiedener (Meta-)Daten untereinander
- Kontrolle der Ordnerstruktur
- Kontrolle des obligatorischen Inhalts (XML/PDF)
- Warnung bei leeren oder sehr kleinen Dateien
- Anzahl der Dateien, Größe des Journals (MiB)

Ergebnisse

- Report-File (Überblick)
- Log-File (chronologisch, detaillierte Fehlermeldungen)
- merkt sich verarbeitete ISSNs (keine doppelte Verarbeitung)
- Bei fehlerfreiem Durchlauf: Journal wird kopiert, MD5-Prüfsummen generiert, Weiterverarbeitung (XSLT, CSV-Ingest,...)

...wenn Fehler



II Detaillierter Abgleich

IMP · Ingest Material Provider

IMPlementierung

- Überwiegend Python, Teile als Bash-Skript
- Input: ISSN
- Iteriert über Issue-Ordner
- Hierarchisches Modell
  - Artikel-Ebene
    - alle Dateien des Artikel-Ordners
    - Metadaten aus XML-Repräsentation extrahiert
    - Metadaten aus Dateiname abgeleitet
  - Issue-Objekt erbt Informationen aus Artikel-Ebene
  - Journal-Objekt erbt Informationen aus Issue-Ebene
- Wenn keine Fehler: kopiere Journal und erzeuge Prüfsummen

- Output: Log, Report, kopiertes Journal + Prüfsummen

II Detaillierter Abgleich · Ebenen der Datenkontrolle

Ebene des Abgleichs	Quelle A	Quelle B
Journal-Ebene	Verdichtete Metadaten aus den XML-Repräsentationen der Artikel Summe der Artikel pro Journal	Metadaten und statistische Informationen aus dem Wiley Journal Inventory
Issue-Ebene	Summe der Artikel pro Issue (Anzahl der Artikelordner)	Summe der Artikel im Wiley Issue Inventory
Artikel-Ebene	Metadaten aus deskriptiven Dateinamen Anzahl Artikel-PDFs im Article-Root	Metadaten aus Artikel-XML Soll: 1 XML-Repräsentation 1 PDF-Repräsentation (idR)
Datei-Ebene	Dateipfad Dateiname Dateigröße	Vorgegebene Ordnerstruktur? Dateigröße null?

II Detaillierter Abgleich · Akkumulation der Daten

Beispiel: Number of Articles

Journal Level

Anzahl Artikel im Journal
= Summe Artikel der Issues

Issue Level

Anzahl Artikel pro Issue
= Summe Artikel-Ordner

Article Level

1 Artikel-Ordner = 1 Artikel

Regeln

Addiere die Summe der Artikel des Issues zu der Summe der bisher gezählten Artikel

Zähle den Artikel zu der Summe der bisherigen Artikel

II Detaillierter Abgleich · Akkumulation der Daten

Beispiel: Last Issue Online
(Jg. 2011, Heft Nr. 3)

Journal Level

Bezeichnung des ältesten Issues
Bisher spätestes Datum

Information ableitbar

Issue Level

Bezeichnung des ältesten Issues
Bisher spätestes Datum

Information existiert nicht

Article Level

XML: Issue
XML: Date

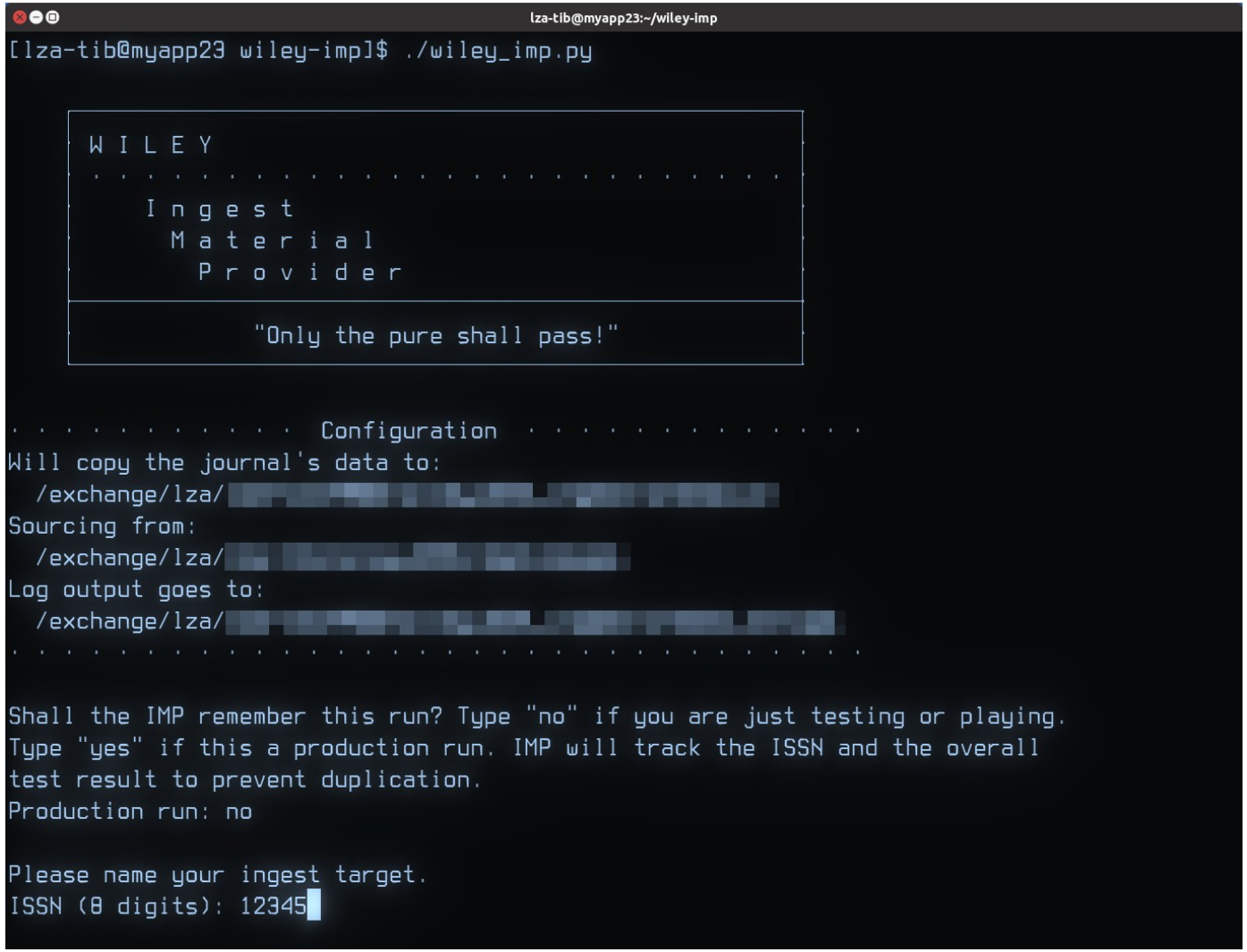
Information existiert nicht

Regel

Wenn Datum später als bisher gesehen:

- setze neues spätestes Datum
- setze aktuell verarbeiteten Issue als ältesten Issue

II Detaillierter Abgleich · IMP User Input



Screenshot: Der Ingest Material Provider wird auf einem Linux-Server ausgeführt

II Detaillierter Abgleich · IMP Auszug Report-File

```
IMP Report · Wiley Journals · 2022-02-24

Journal Title: Acta Paediatrica, Acta Pædiatrica
ISSN: 0803-5253
Issues: 344
Bad Issues: 0
Articles: 10284
Tainted Articles: 0
Files: 31884
Size: 4475.46 MiB


General evaluation: FAILURE.
Copy- and MD5-job: skipped.

There have been WARNINGS.
There have been ERRORS.
Check the log file.


---

Journal Level Metadata Check:

VALUE                    MATCH    WILEY JOURNAL INVENTORY              ACTUAL DATA (XML/FILES)

Journal group code:        ✓      APA                                  APA
publication_title:         ✗      Acta Paediatrica                     Acta Paediatrica, Acta Pædiatrica
print_identifier:          ✓      0803-5253                            0803-5253
date_first_issue_ online:  ✓      1997                                 1997
num_first_vol_online:      ✓      86                                   86
num_first_issue_ online:   ✓      1                                    1
date_last_issue_online:    ✓      2021                                 2021
num_last_vol_online:       ✓      110                                  110
num_last_issue_online:     ✓      1                                    1
number of issues:          ✓      344                                  344
number of articles:        ✓      10284                                10284
number of PDF:             ✓      10284                                10284
number without PDF:        ✓      0                                    0
```

In den XML-Metadaten der Artikel variiert die Schreibweise des Titels.

Da dies nicht identisch mit der Information aus dem Journal Inventory ist, scheitert der Abgleich.

Schluss

Positiv

- Wir finden Fehler, es ist nicht umsonst
- leere Dateien
- selten: Metadatenfehler in den XMLs
- Fehler in den Inventories

Mögliche Mängel

- Wir vergleichen Informationen von Wiley (Datenlieferung) mit Informationen von Wiley (Inventory)
- Wir wissen nichts über die Entstehung der Inventories

Thanks!

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

MEHR INFORMATIONEN

www.tib.eu

Kontaktdaten
Peter Eisner
peter.eisner@tib.eu

Leibniz Gemeinschaft