IT'S TOOL TIME!
EINBLICKE IN DEN MASCHINENRAUM DER GND-AGENTUR LEO-BW-REGIONAL

24. MusIS-Jahrestreffen in Konstanz
19./20. März 2024

Jens M. Lill (BSZ)

Agenda

- Thema „Datenqualität & Datenaufbereitung“ beim MusIS-Jahrestreffen 2023
- Motto: „Pimp my Data for GND“
- Datenanalyse – Datenaufbereitung (Pre-processing) – Datenqualität
  - Daten harmonisieren, bereinigen und formal korrigieren (Feldinhalte, Datenstruktur)
  - Daten mit der GND abgleichen → Dubletten ausschließen
  - Daten mit Informationen und IDs aus der GND anreichern
- Kandidaten für GND-Import identifizieren → Eignungs-/Relevanzkriterien
- Datentransformation ins GND-Zielformat (MARC-XML)
- Einspielen der Daten durch GND-Zentrale
- Redaktion & Datenpflege durch GND-Agentur

Tools…

GND4C-Toolbox

OpenRefine
GREL

DPT

Python-Skripte

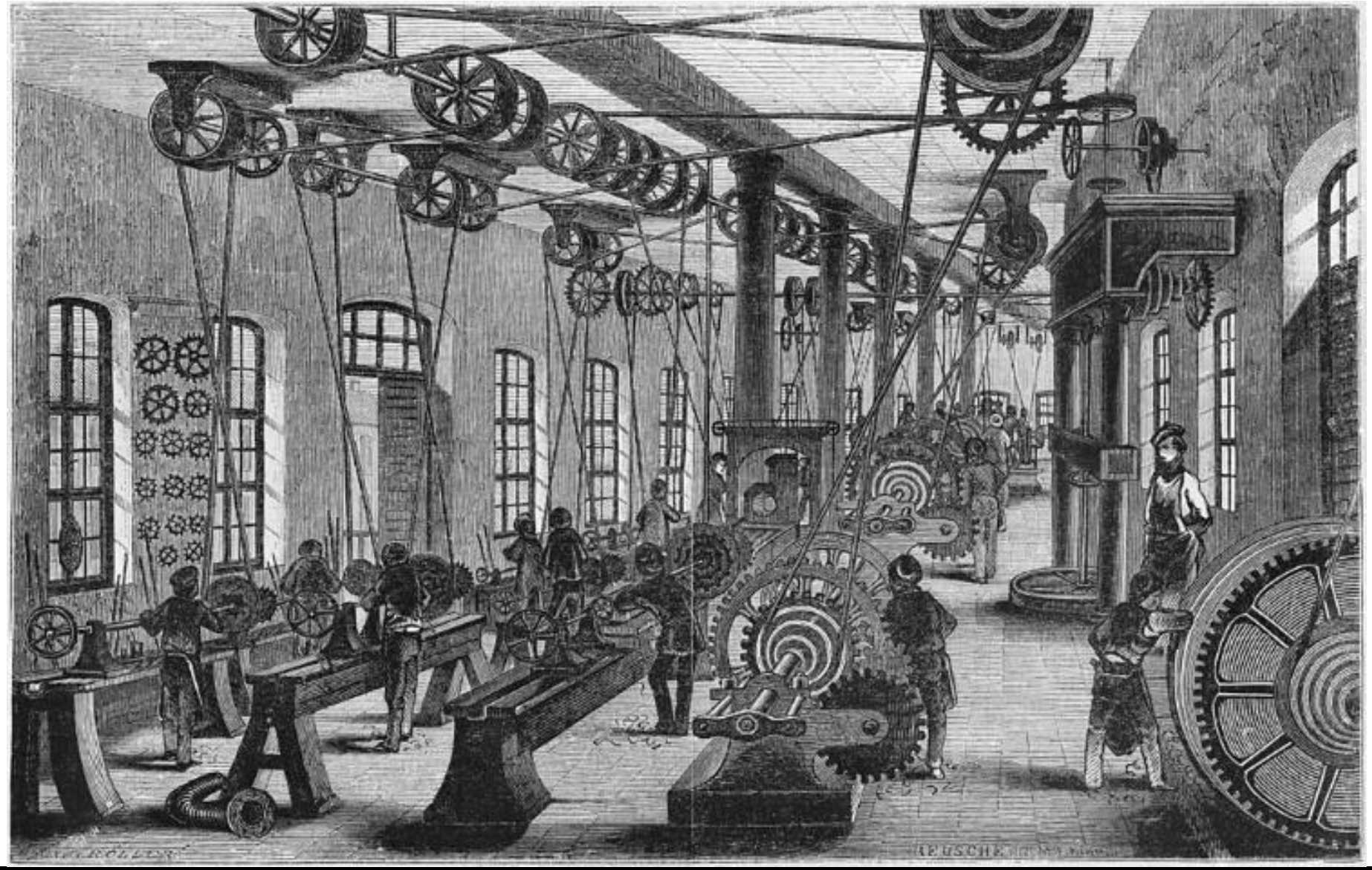
XSLT

entityXML

Notepad++

swb.trafo
ExpoDB 2.0

ChatGPT
NER



„Die Montirungswerkstatt in der Maschinenbauanstalt von Maffei in Hirschau bei München“ (Leipziger Illustrierte Zeitung, um 1849); Wikimedia Commons (Public Domain)

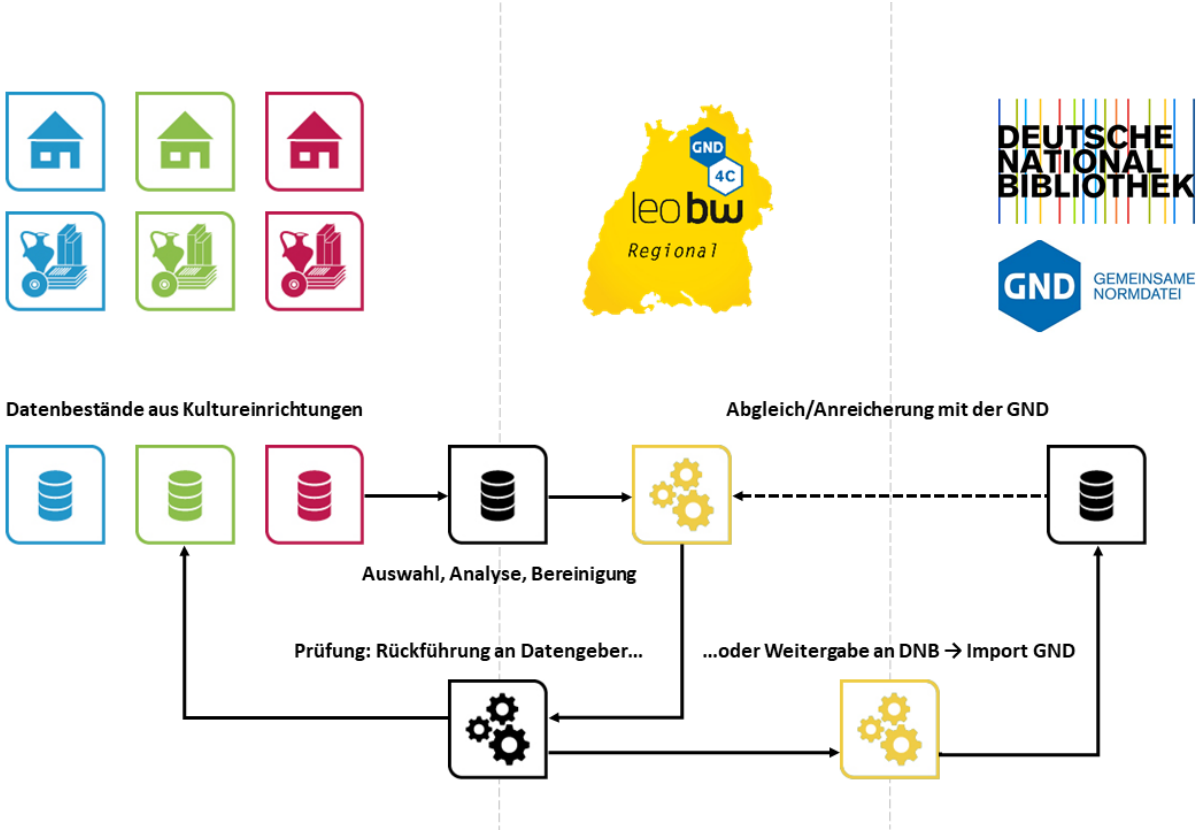
Datentransformation: imdas pro → „GND4C-XML“

- Projektinternes XML-Format, angelehnt an GND-Ontology
- Ausgerichtet auf GND4C-Toolbox v1 (NDS-XML)
- Allerdings obsolet → GND4C-Toolbox v2 basiert auf neuem Software-Stack (JSON)
- Abhilfe: skriptbasierter Workflow „NDS-XML-2-OR-2-NDS-XML“
  - „Lösung ist nicht elegant, aber effektiv“
  - „Durch die vielen Einzelschritte sehr transparent und dadurch wenig fehleranfällig“


```xml
<item>
<recordID>9006</recordID>
<person>
<nameOfThePerson role="preferred">
<surname>Gesner</surname>
<forename>Conrad</forename>
</nameOfThePerson>
<gender>
<label lang="de" role="preferred">männlich</label>
</gender>
<academicDegree>
<label>Professor</label>
</academicDegree>
<professionOrOccupation>
<label lang="de" role="preferred">Physiker</label>
</professionOrOccupation>
<professionOrOccupation>
<label lang="de" role="preferred">Arzt</label>
</professionOrOccupation>
<functionOrRole>
<label lang="de" role="preferred">abgebildete Person</label>
</functionOrRole>
</person>
</item>
```

„Interimslösung“ als Alternative zur GND4C-Toolbox

- Ziel: Verknüpfung, Anreicherung, Integration von Datenbeständen mit / in der GND
  - Datenbereinigung, Reconciliation
  - Datenanreicherung (GND-Nummern)
  - Datentransformation für GND-Import (MARC-XML)
  - Matching u. Deduplizierung seitens DNB
- Offene Software-Komponenten
  - **OpenRefine** inkl. lobid-API
  - Python-Skripte
  - **Data Preparation Tool** der DDB-Fachstelle Archiv



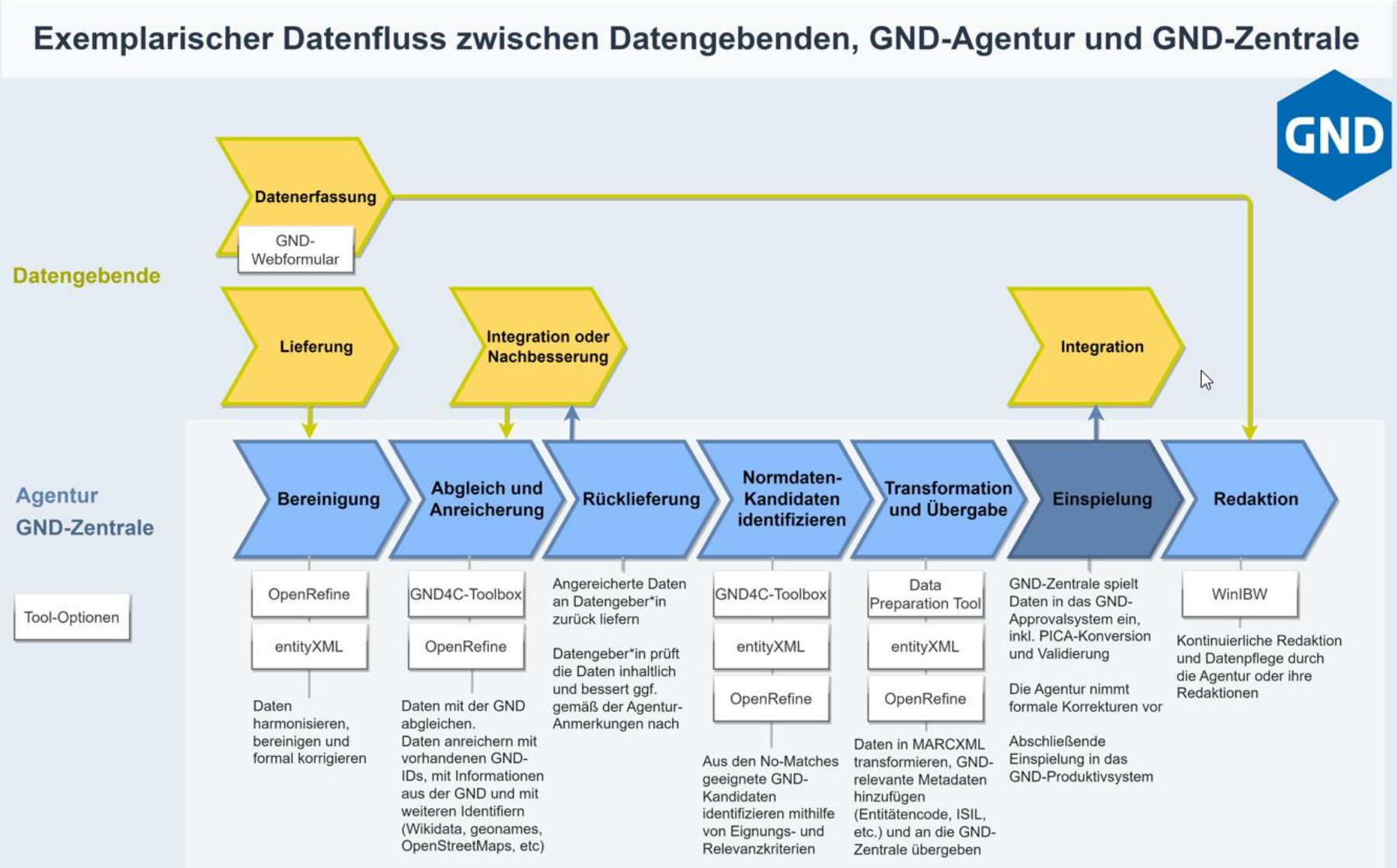
Test mit Datenset aus Gotha:

- 1.157 abgebildete Personen auf Kunstwerken, Medaillen
- 945 Personen nach Daten-bereinigung an DNB übergeben
- 485 Personen von DNB als valide ins Testsystem importiert
- Echtimport steht noch aus…

Zeitlicher Verlauf:

- Start des Testlaufs: 02/2023
- Übergabe an DNB: 07/2023
- GND-Testimport: 09/2023
- GND-Echtimport: XX/2024

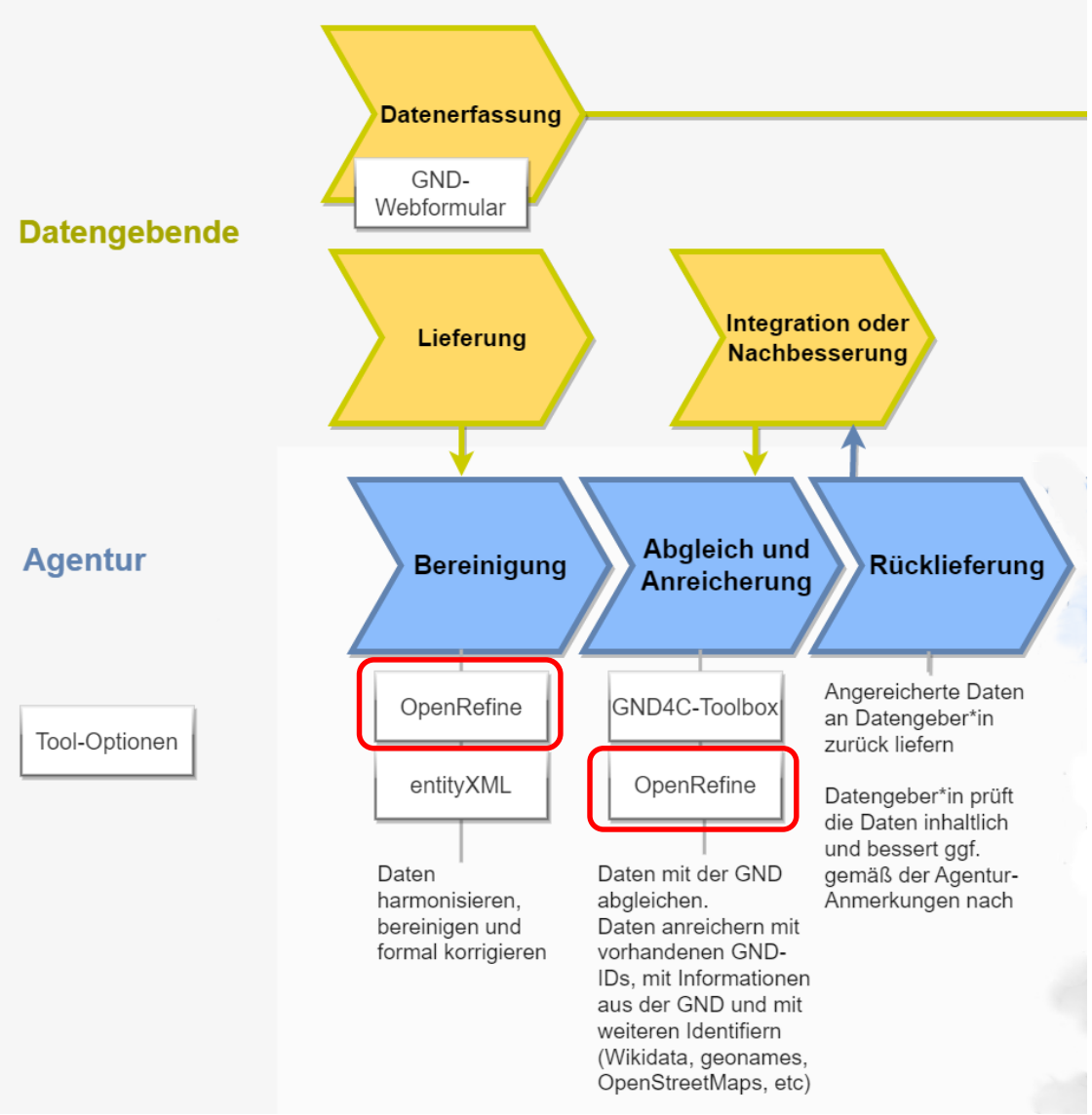
- Arbeitszeit (netto): 2-3 Wochen



Grafik: Chantal Köppl (DNB), CC BY-SA 4.0, Jan. 2024

Pre-processing

- Ziel: Verbesserung der Datenqualität
- Auch losgelöst von GND-Import zu sehen
- Daten analysieren & aufräumen (Feldinhalte, Datenstruktur)
  - Schreibfehler korrigieren
  - Feldbelegungen vereinheitlichen bzw. auftrennen → Titel vs. Stand
  - Schreibweisen vereinheitlichen
  - Datumsangaben vereinheitlichen
- Daten mit GND abgleichen & anreichern



Beispiele für notwendige Datenaufbereitungen

- Personennamen RDA-gerecht auftrennen oder zusammenführen
- Datumsformate überprüfen und ggf. angleichen
- Datumsangaben aufspalten/duplizieren in exakte Lebensdaten und nur Jahresangaben ($4-Codes datl und ggf. datx)
- Ungefähre Lebens-/Wirkungsdaten → einleitende Wendung "ca." ergänzen
- Datenelement "functionOrRole" → da objektbezogene Angaben und keine Berufsangaben im GND-Sinn, Angaben wie Bestimmer, Finder, Hersteller, Dargestellte Person etc. herausfiltern
  → Folge mitunter: es können keine Berufsangaben übergeben werden...

Beispiele für notwendige Datenaufbereitungen

- Angaben zur Nationalität von Personen im Datenelement "geographicAreaCode" in GND-Ländercode wandeln (Pflichtfeld!)
- Mehrfachfelder auftrennen, z.B. Geografika nach Geburts-, Sterbe-, Wirkungsort
- Geschlechtsspezifische Anpassung von Berufsangaben – oftmals Sammelbezeichnungen wie "Künstler/in", "KünstlerIn", "Künstler*in" oder nur männliche Benennungen
- GND-Reconciliation: Ermittlung der GND-Identifier für 5XX-Felder wie Personen, Orte, Sachbegriffe → Voraussetzung für GND-Import

Datenbereinigung, -abgleich, -anreicherung

Anforderungen Pre-processing

- [x] Angaben aus <academicDegree> vereinheitlichen bzw. in <titleOfNobility> oder <professionOrOccup
- [x] Angaben in <gender> in GND-Form überführen
- [x] Angaben von <miscellaneousNote> in <biographicalOrHistoricalInformation> überführen oder ggf. ir
- [x] Angaben in <functionOrRole> nach Möglichkeiten überführen in <professionOrOccupation>
- [x] Geschlechtsspezifische Anpassung und Vereinheitlichung von Berufsangaben
- [x] ggf. Zusatzangaben in spitzwinkligen Klammern aus dem Gemeindeverzeichnis bei Geografika eliminie
- [x] ⚠️ Angaben mit einem kaufmännischen & müssen auch aufgelöst werden → &
- [x] Namensspalten zusammenführen und als neue Spalte "new_nameOfThePerson" anlegen bzw. in <pers
- [x] Datumsangaben bereinigen bzw. aufspalten in datl und datx – mittels GREL-Funktionen → ggf. datx b
  - [x] neue Spalten: new_dateOfBirth_datl, new_dateOfBirth_datx, new_dateOfDeath_datl, new_dateO
- [x] ungefähre Datumsangaben in <primitive_value> mit Präfix "ca." ergänzen
- [x] Reconciling und Nachladen der GND-Nummern bei 5xx-Feldern → Berufe, Orte, Adelstitel
  - [x] <professionOrOccupation> → GND-Nummern in neue Spalte "gnd_professionOrOccupation_b
  - [x] <titleOfNobility> → GND-Nummern in neue Spalte "gnd_titleOfNobility_adel" nachladen
  - [x] <associatedPlace> → GND-Nummern in neue Spalte "gnd_associatedPlace_ortw" nachladen
  - [x] <placeOfBirth> → GND-Nummern in neue Spalte "gnd_placeOfBirth_ortg" nachladen
  - [x] <placeOfDeath> → GND-Nummern in neue Spalte "gnd_placeOfDeath_orts" nachladen
  - [x] ⚠️ ISIL der GND als Präfix zur GND-Nummer ergänzen → (DE-588)4033423-5 → via DPT
- [x] Ländercodes anhand von Landesangaben in z.B. Wirkungsort ergänzen sowie von reconciled Geburts-
  → 1. Reconcile / Use values as identifiers → 2. Edit column / Add column from reconciled values...
- [x] leere Ländercodes mit ZZ auffüllen → im row-Modus und nur bei ausgefüllter <recordID>
  - [x] Arbeitsspalte "temp_geographicAreaCode" mit <geographicAreaCode> zusammenführen
- [x] mit xx in <recordID> markierte Datensätze löschen
- [x] Dokumentation der neu angelegten Spalten → Anpassungen Python-Skript & DPT

item - person - professionOrOccupation - label	gnd_professionOrOccupation_beru	item - person - titleOfNobility - label	gnd_titleOfNobility_adel
		Graf Choose new match	4021778-4
		Königin Choose new match	4164379-3
Großadmiral Choose new match	4204637-3		
Münzherr Choose new match	1051614252	Herzog Choose new match	4159676-6
		Herzog Choose new match	4159676-6
Bischof Choose new match	4006949-7		
		Herzog Choose new match	4159676-6
Maler Choose new match	4037215-7		

Je nach Datenlage → 20-30 separate Arbeitsschritte

Angereichertes NDS-XML

```xml
<item>
<recordID>9006</recordID>
<person>
<nameOfThePerson role="preferred">
<surname>Gesner</surname>
<forename>Conrad</forename>
</nameOfThePerson>
<gender>
<label lang="de" role="preferred">männlich</label>
</gender>
<academicDegree>
<label>Professor</label>
</academicDegree>
<professionOrOccupation>
<label lang="de" role="preferred">Physiker</label>
</professionOrOccupation>
<professionOrOccupation>
<label lang="de" role="preferred">Arzt</label>
</professionOrOccupation>
<functionOrRole>
<label lang="de" role="preferred">abgebildete Person</label>
</functionOrRole>
</person>
</item>
```

```xml
<item>
  <recordID>9006</recordID>
  <person>
    <nameOfThePerson role="preferred">
      <surname>Gesner</surname>
      <forename>Conrad</forename>
      <NEWnameOfThePerson>Gesner, Conrad</NEWnameOfThePerson>
    </nameOfThePerson>
    <gender>
      <label>m</label>
    </gender>
    <academicDegree>
      <label>Prof.</label>
    </academicDegree>
    <professionOrOccupation>
      <label role="preferred" gnd="4045968-8">Physiker</label>
    </professionOrOccupation>
    <professionOrOccupation>
      <label role="preferred" gnd="4003157-3">Arzt</label>
    </professionOrOccupation>
    <functionOrRole>
      <label lang="de" role="preferred">abgebildete Person</label>
    </functionOrRole>
    <geographicAreaCode>
      <label lang="de">ZZ</label>
    </geographicAreaCode>
  </person>
</item>
```

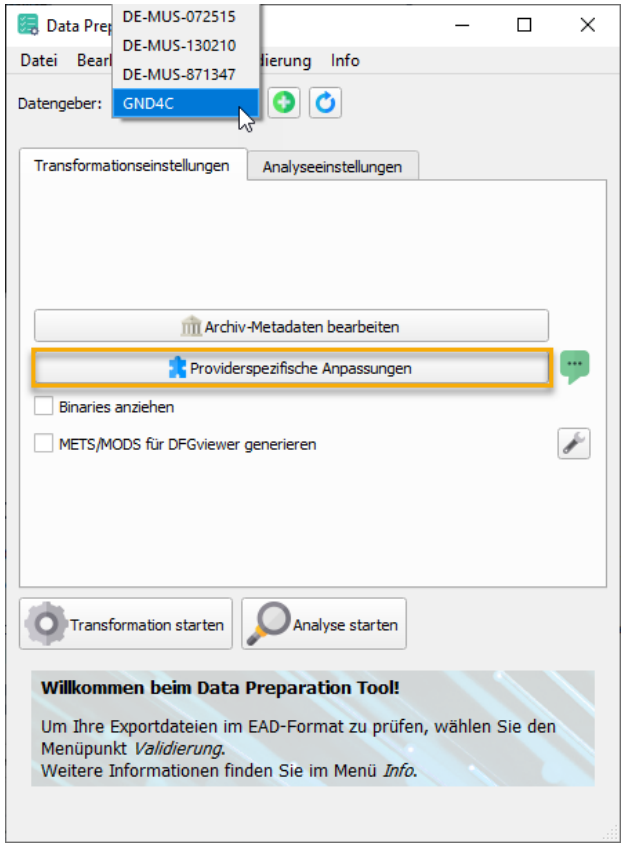
Bausteine / Werkzeuge

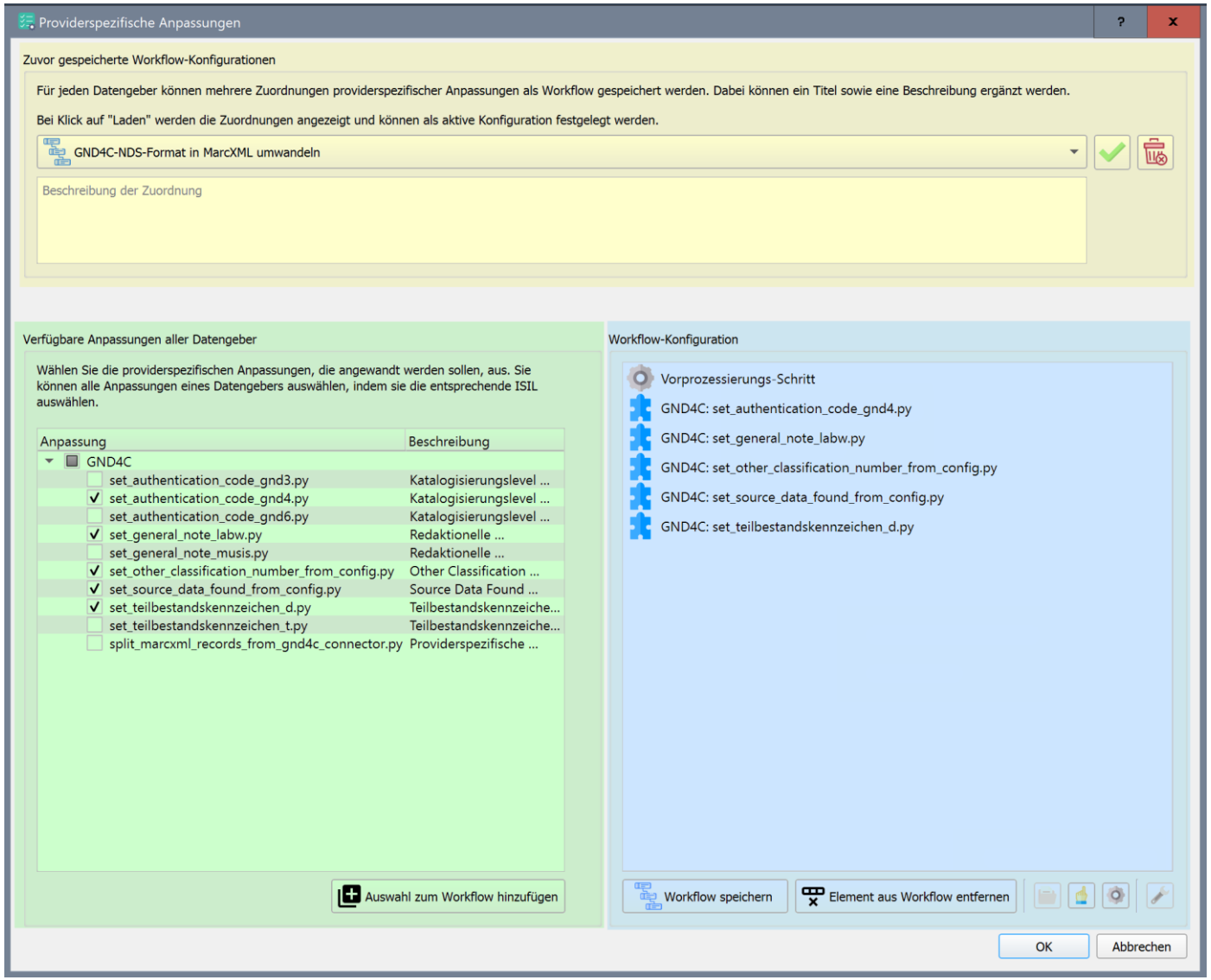
- OpenRefine für Datenbereinigung, -abgleich, -anreicherung…
  - …um Datenqualität zu verbessern, d.h. weniger redaktionelle Nacharbeit vonnöten
  - …um potenzielle GND-Dubletten vorab auszuschließen
  - …um Anforderungen der DNB zu begegnen (GND-IDs liefern anstelle von Literalen in 5XX-Feldern)
  - …um nachgelagerte Transformation im DPT nach MARC-XML zu erleichtern → Zielformat für GND-Import

Bausteine / Werkzeuge

- Data Preparation Tool (DDB-Fachstelle Archiv, LABW)
  - Python-basierte Desktopanwendung, die Daten aus beliebigen Quellformaten in andere Zielformate transformiert
  - Webanwendung in Vorbereitung
  - Transformationsroutine für Personen nach MARC-XML umgesetzt
  - Voraussetzung: Konkordanz vom NDS-Lieferformat auf MARC-XML
  - Anpassungsbedarf (iterativ) in Rücksprache mit Entwickler
  - Ermöglicht v.a. Ergänzung GND-interner Felder

Data Preparation Tool





Transformation nach MARC-XML (Zielformat GND)


```
<record type="Authority">
<leader>00000nz  a2200000o  4500</leader>
<controlfield tag="001">9006</controlfield>
<controlfield tag="003">DE-MUS-871347</controlfield>
<controlfield tag="008">230715n||azznnaabn           | aaa    |c</controlfield>
<datafield tag="040" ind1=" " ind2=" ">
<subfield code="a">DE-MUS-871347</subfield>
<subfield code="b">ger</subfield>
<subfield code="c">DE-MUS-871347</subfield>
<subfield code="9">r:DE-576-3</subfield>
</datafield>
<datafield tag="042" ind1=" " ind2=" ">
<subfield code="a">gnd4</subfield>
</datafield>
<datafield tag="043" ind1=" " ind2=" ">
<subfield code="c">ZZ</subfield>
</datafield>
<datafield tag="075" ind1=" " ind2=" ">
<subfield code="b">p</subfield>
<subfield code="2">gndgen</subfield>
</datafield>
<datafield tag="075" ind1=" " ind2=" ">
<subfield code="b">piz</subfield>
<subfield code="2">gndspec</subfield>
</datafield>
<datafield tag="079" ind1=" " ind2=" ">
<subfield code="a">g</subfield>
<subfield code="q">t</subfield>
</datafield>
<datafield tag="100" ind1="1" ind2=" ">
<subfield code="a">Gesner, Conrad</subfield>
</datafield>
<datafield tag="375" ind1=" " ind2=" ">
<subfield code="a">1</subfield>
<subfield code="2">iso5218</subfield>
</datafield>
<datafield tag="550" ind1=" " ind2=" ">
<subfield code="a">Prof.</subfield>
<subfield code="4">akad</subfield>
</datafield>
<datafield tag="550" ind1=" " ind2=" ">
<subfield code="0">(DE-588)4045968-8</subfield>
<subfield code="a">Physiker</subfield>
<subfield code="4">berc</subfield>
</datafield>
<datafield tag="550" ind1=" " ind2=" ">
<subfield code="0">(DE-588)4003157-3</subfield>
<subfield code="a">Arzt</subfield>
<subfield code="4">beru</subfield>
</datafield>
<datafield tag="667" ind1=" " ind2=" ">
<subfield code="a">MusIS</subfield>
<subfield code="5">DE-641</subfield>
</datafield>
<datafield tag="670" ind1=" " ind2=" ">
<subfield code="a">Museumsdokumentation Stiftung Schloss Friedenstein Gotha</subfield>
</datafield>
</record>
```

Input vs. Outcome

- Datenumfang ↔ Datenqualität
- Mindestanforderungen an identifizierenden Merkmalen für GND-Sätze gegeben?
- GND-Ausschuss muss Import zustimmen (ggf. Verzögerung)
- Abhängigkeit vom Flaschenhals DNB-Zentrale bei Import!!
- Abwägen, ob Aufwand lohnt oder nicht schneller Eingabe per GND-Webformular

Modernisierung des Maschinenraums…?

- Abstimmung auf GND4C-Toolbox und ggf. Austauschformat entityXML (NFDI-Kontext)
- NER-Komponente via ExpoDB 2.0?
  - Extraktion von identifizierenden Angaben aus biografischen Freitextfeldern (Berufe, Orte, Datumsangaben, Personenbeziehungen)
  - Ansatz: KI-gestützte Named Entity Recognition

app
Über GND4C
Personen
Nachrichten
Blog Posts
Tutorials

Username
michael.markert@uni-jena.de

Password
••••••••••••

Login

Logout

Features Sub-Menu
- Toolbox
- Matching
- Data Import
- Ergebnisse
- Reporting
- Personal Liste
- Data Profiling
- Data Visualisierung
- Blogpost
- Benutzer Verwaltung

Update

default_title	gndIdentifier	GND-URL	new_total_score	total_score	new_param_max_name_...	new_param_lstMx_x_Anz
FactGrid-DublettenPers...	1034900420		4.19	3.34	5.192	0
FactGrid-DublettenPers...	1029305587		4.19	3.34	5.192	0
FactGrid-DublettenPers...	124604609		-0.49	-1.34	5.192	0
FactGrid-DublettenPers...	131773852		-2.76	-2.76	3.894	0
FactGrid-DublettenPers...	1015785050		-3.82	-3.82	2.596	0
FactGrid-DublettenPers...	1159095906		-4.52	-4.52	3.894	0
FactGrid-DublettenPers...	1147498156		-4.52	-4.52	3.894	0
FactGrid-DublettenPers...	1053491611		-4.52	-4.52	3.894	0
FactGrid-DublettenPers...	140685707		-4.52	-4.52	3.894	0
FactGrid-DublettenPers...	1211645231		-4.52	-4.52	3.894	0
FactGrid-DublettenPers...	1036422674		-4.68	-4.68	2.596	0
FactGrid-DublettenPers...	11644312X		-4.77	-4.77	1.298	0
FactGrid-DublettenPers...	1037506359		-4.77	-4.77	1.298	0
FactGrid-DublettenPers...	1072306484		-4.9	-4.9	1.1682	0
FactGrid-DublettenPers...	1053387660		-5.38	-5.38	3.894	0

1 to 15 of 18 Page 1 of 2

NDS Source Data

	0
Entity Type	Person
Forename	Daniel Georg
Surname	von der Schulenburg
Personal Name	None
Name Addition	None
Counting	None

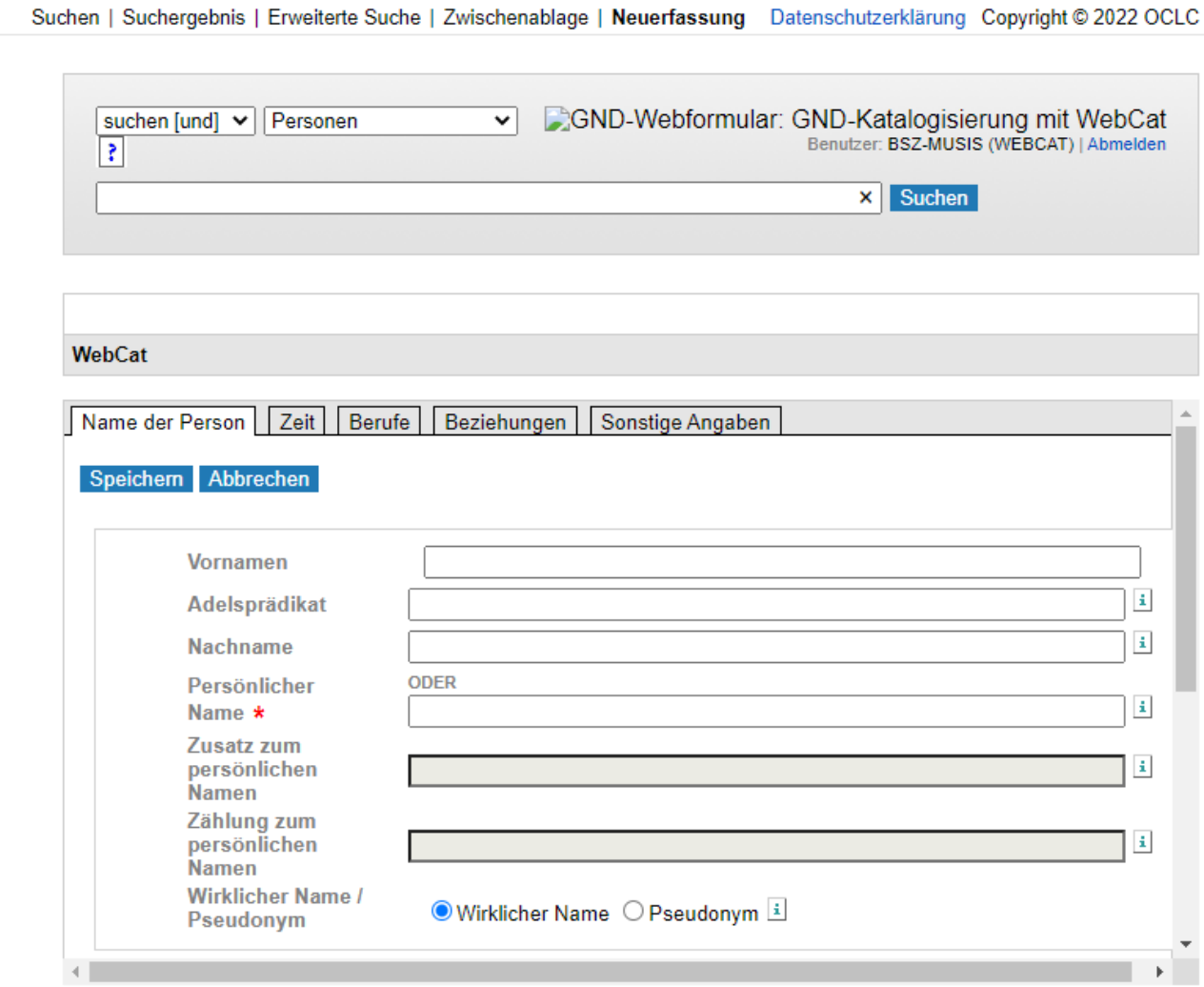
GND Target Data

	0	1
gndIder	1034900420	1029305587
id	https://d-nb.info/gnd/1034900420	https://d-nb.info/gnd/1029305587
preferre	Schulenburg-Hehlen, Daniel Christc	Schulenburg, Daniel Christoph Geor
variantN	nan	nan
dateOfE	['1716-04-17']	['1716']
dateOfD	['1772-11-25']	['1772']

Grafik „Personen-Matching in der Toolbox “: Michael Markert (ThULB), CC 0, Jan. 2024

Weitere Infos / Hinweise

- Blogpost: [It's Tool Time! Datenprozesse in den neuen GND-Agenturen]
- Nächste Schulungstermine im GND-Webformular:
  - 5./7. Juni 2024 (online)
  - 6./8. November 2024 (online)



Kontakt

Vielen Dank für Ihre Aufmerksamkeit!

gnd4c@leo-bw.de
https://www.leo-bw.de/gnd-agentur