Franzi Fuchs, Anke Taube

DNBLab: Daten bereinigen und zusammenführen

Inhalt

1. Bezugswege im DNBLab
2. Datenselektion mit SRU-Schnittstelle
3. Datenbezug mit dem SRU Query Tool
4. Datenbereinigung und Anreicherung
5. Datenanalyse

dnb.de / metadatenbezugswege

Titeldaten

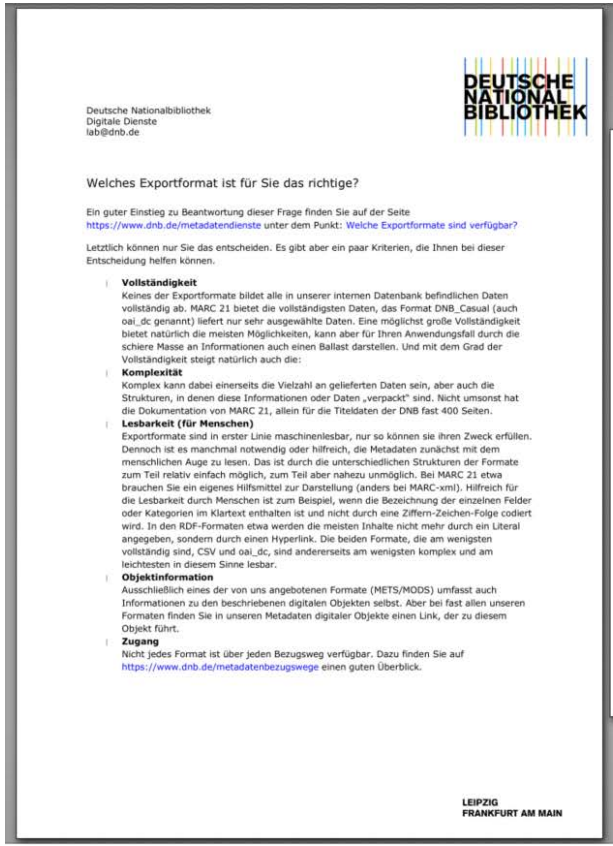
Titeldaten				
Format	**Datenshop**	**SRU**	**OAI-PMH**	**WWW-/SFTP-Server**
MARC 21	x			x
MARC21-xml	x	x	x	x
RDF (RDF/XML)	x	x	x	x
RDF (JSON-LD)	x			x
RDF (Turtle)	x			x
CSV	x			
DNB Casual (oai_dc)	x	x	x	
MODS-xml	x	x	x	
PDF	x			x

Normdaten

Normdaten				
Format	**Datenshop**	**SRU**	**OAI-PMH**	**WWW-/SFTP-Server**
MARC 21	x			x
MARC21-xml	x	x	x	x
RDF (RDF/XML)	x	x	x	x
RDF (JSON-LD)	x			x
RDF (Turtle)	x			x

Welches Exportformat ist für Sie das Richtige?

Handout unter dnb.de/dnblabtutorials



Deutsche Nationalbibliothek
Digitale Dienste
lab@dnb.de

Welches Exportformat ist für Sie das richtige?

Ein guter Einstieg zu Beantwortung dieser Frage finden Sie auf der Seite https://www.dnb.de/metadatendienste unter dem Punkt: Welche Exportformate sind verfügbar?

Letztlich können nur Sie das entscheiden. Es gibt aber ein paar Kriterien, die Ihnen bei dieser Entscheidung helfen können.

- **Vollständigkeit**
  Keines der Exportformate bildet alle in unserer internen Datenbank befindlichen Daten vollständig ab. MARC 21 bietet die vollständigsten Daten, das Format DNB_Casual (auch oai_dc genannt) liefert nur sehr ausgewählte Daten. Eine möglichst große Vollständigkeit bietet natürlich die meisten Möglichkeiten, kann aber für Ihren Anwendungsfall durch die schiere Masse an Informationen auch einen Ballast darstellen. Und mit dem Grad der Vollständigkeit steigt natürlich auch die:
- **Komplexität**
  Komplex kann dabei einerseits die Vielzahl an gelieferten Daten sein, aber auch die Strukturen, in denen diese Informationen oder Daten „verpackt" sind. Nicht umsonst hat die Dokumentation von MARC 21, allein für die Titeldaten der DNB fast 400 Seiten.
- **Lesbarkeit (für Menschen)**
  Exportformate sind in erster Linie maschinenlesbar, nur so können sie ihren Zweck erfüllen. Dennoch ist es manchmal notwendig oder hilfreich, die Metadaten zunächst mit dem menschlichen Auge zu lesen. Das ist durch die unterschiedlichen Strukturen der Formate zum Teil relativ einfach möglich, zum Teil aber nahezu unmöglich. Bei MARC 21 etwa brauchen Sie ein eigenes Hilfsmittel zur Darstellung (anders bei MARC-xml). Hilfreich für die Lesbarkeit durch Menschen ist zum Beispiel, wenn die Bezeichnung der einzelnen Felder oder Kategorien im Klartext enthalten ist und nicht durch eine Ziffern-Zeichen-Folge codiert wird. In den RDF-Formaten etwa werden die meisten Inhalte nicht mehr durch ein Literal angegeben, sondern durch einen Hyperlink. Die beiden Formate, die am wenigsten vollständig sind, CSV und oai_dc, sind andererseits am wenigsten komplex und am leichtesten in diesem Sinne lesbar.
- **Objektinformation**
  Ausschließlich eines der von uns angebotenen Formate (METS/MODS) umfasst auch Informationen zu den beschriebenen digitalen Objekten selbst. Aber bei fast allen unseren Formaten finden Sie in unseren Metadaten digitaler Objekte einen Link, der zu diesem Objekt führt.
- **Zugang**
  Nicht jedes Format ist über jeden Bezugsweg verfügbar. Dazu finden Sie auf https://www.dnb.de/metadatenbezugswege einen guten Überblick.

LEIPZIG
FRANKFURT AM MAIN

Handout unter dnb.de/dnblabtutorials

Deutsche Nationalbibliothek
Digitale Dienste
lab@dnb.de

Übersicht: Inhalte in MARC21-Feldern

Die folgende Tabelle bietet einen Überblick über die am häufigsten benötigten und für die Arbeit mit den bibliografischen Daten unserer Titeldaten relevantesten MARC21-Felder[1]. Sie ist **nicht** vollständig! Eine Gesamtübersicht zu MARC21, allen Feldern und deren Bedeutungen für unsere Titeldaten, Bestandsdaten sowie für die Normdaten der GND finden Sie unter https://www.dnb.de/marc21.

Ich suche		Art des Feldes[2]	Tag	Subfield/ Position
Identifikatoren	IDN	C	001	-
	ISBN	D	020	a
	ISSN	D	022	a
	URN (und weitere Identifikatoren)	D	024	a
	Quelle der Nummer oder des Codes	D	024	2
Verantwortliche und Mitwirkende (bspw. Autoren/Herausgeber)[3]	Personenname (Haupteintragung)	D	100	a
	zugehörige Funktionsbezeichnung	D	100	e
	Personenname (Nebeneintragung – bspw. Mitverantwortliche)	D	700	a
	zugehörige Funktionsbezeichnung	D	700	e
	Körperschaftsname (Haupteintragung)	D	110	a
	Kongressname (Haupteintragung)	D	111	a
	Körperschaftsname (Nebeneintragung)	D	710	a
	Kongressname (Nebeneintragung)	D	711	a
Titelangabe	Titel	D	245	a
	Zusatz zum Titel	D	245	b
	Verfasserangabe etc.	D	245	c
Titelvarianten	Hauptsachtitel / Kurztitel	D	246	a
	Zusatz zum Titel	D	246	b
Gesamttitel (bspw. bei mehrbändigen Werken)	Gesamttitelangabe	D	490	a
	Bandnummer/Bandbezeichnung	D	490	v
	Gesamttitel – Einheitstitel (Nebeneintragung)	D	830	a
Einheitstitel	Einheitstitel (Haupteintragung)	D	240	a
	Einheitstitel (Nebeneintragung)	D	730	a
Ausgabebezeichnung	Ausgabe / Auflage	D	250	a
Verlag	(Erzeuger-, Verlags-, Vertriebs-, Herstellername)	D	264	b
Publikationsort	(Entstehungs-, Verlags-, Vertriebs-, Herstellungsort)	D	264	a
Datumsangaben	Erscheinungsdatum (Entstehungs-, Verlags-, Vertriebs-, Herstellungs- oder Copyrightdatum)	D	264	c
	Erscheinungs-/Entstehungsdatum	C	008	07-10
	Abschließendes Erscheinungs-/Entstehungsdatum[4]	C	008	11-14
	Datum der Ersterfassung	C	008	00-05
Sprache	des vorliegenden Werkes	D	041	a
	des Originals (bei Übersetzungen)	D	041	h
Physische Merkmale	Umfang (bspw. Seitenzahl)	D	300	a
	Andere physische Details	D	300	b
	Dimension (bspw. Höhe in cm)	D	300	c
Klassifikation/Sachgruppe	DDC-Notation	D	082	a
	Vergabestelle	D	082	q
	Ausgabenummer	D	082	2
	Zusätzliche DDC-Notation (alle Felder wie DDC-Hauptsachgruppe – nur, wenn 082 bereits vorhanden)	D	083	Alle Subfields wie 082
	Andere Notationen	D	084	a
	Vergabestelle der Notation	D	084	q
	Quelle der Nummer	D	084	2
Schlagwörter[5]	Schlagwort – Personenname	D	600	a
	Schlagwort – Körperschaftsname	D	610	a
	Schlagwort – Kongressname	D	611	a
	Schlagwort – Einheitstitel	D	630	a
	Schlagwort – Zeitschlagwort	D	648	a
	Schlagwort – Sachschlagwort	D	650	a
	Schlagwort – Geografischer Name	D	651	a
	Schlagwort – Genre/Formschlagwort	D	655	a
Hochschulschriftenvermerk oder	alle Informationen als Freitext	D	502	a
	Typ (Dissertation, Masterarbeit etc.)	D	502	b
	Universität	D	502	c
	Jahr	D	502	d
Links	URN[6]	D	856	u
	DOI	D	856	u
	Inhaltsverzeichnis[7]	D	856	u
	Inhaltstext[8]	D	856	u
	Weitere URIs	D	856	u
	Spezifische Materialangaben	D	856	3
	Allgemeine Anmerkungen	D	856	z

[1] Einige MARC-Felder sind wiederholbar und können mehrfach in einem Datensatz enthalten sein. Alle Informationen hierzu finden sich in der Feldbeschreibung der Titeldaten.
[2] „C" steht für Controlfeld, „D" für Datafeld.
[3] Relevant sind hier auch jeweils die Unterfelder $0 (IDN des zugehörigen Normdatensatzes) sowie $2 (Quelle).
[4] bei Zeitschriften und Schriftenreihen, alternativ Datum des Originals in Sortierform. Vgl. hierzu die Feldbeschreibung der MARC-Titeldaten unter https://www.dnb.de/marc21.
[5] Relevant sind auch hier jeweils die Unterfelder $0 (IDN des zugehörigen Normdatensatzes) sowie $2 (Quelle).
[6] in Verbindung mit der Angabe „urn" im Unterfeld $3
[7] in Verbindung mit der Angabe „Inhaltsverzeichnis" im Unterfeld $3
[8] in Verbindung mit der Angabe „Inhaltstext" im Unterfeld $3

LEIPZIG
FRANKFURT AM MAIN

Welcher Zugang ist für Sie der Richtige?

Handout unter dnb.de/dnblabtutorials

Deutsche Nationalbibliothek
Digitale Dienste
lab@dnb.de

Welcher Zugang ist für Sie der richtige?

Ein guter Einstieg zu Beantwortung dieser Frage finden Sie auf der Seite https://www.dnb.de/metadatendienste unter dem Punkt: Wie kann ich Daten beziehen?

Letztlich können nur Sie das entscheiden. Es gibt aber ein paar Kriterien, die Ihnen bei dieser Entscheidung helfen können.

- **Technische Voraussetzungen**
  Welche technischen Rahmenbedingungen haben Sie? Denn die verschiedenen Zugänge erfordern unterschiedliche Voraussetzungen. Wollen Sie den Zugang auf die Metadaten der DNB in Ihr eigenes System oder Ihre Bibliothekssoftware einbinden? Dann muss geprüft werden, was Ihr System zulässt und kann. Auch die Größe der Datenpakete, die Sie von uns übernehmen können, stellen verschieden Anforderungen an Ihre technischen Voraussetzungen. Sowohl SRU- als auch OAI-Schnittstelle sind dafür gedacht über einen entsprechenden Client bzw. Harvester Daten zu übernehmen. Beide können aber in kleinem Rahmen auch mit einem Browser sozusagen von Hand genutzt werden.

- **Vollständig oder selektiv?**
  Ein entscheidendes Kriterium ist die Frage vollständig oder selektiv? Was ist damit gemeint: Wollen Sie die Metadaten der DNB vollständig oder Teile der Sammlung in sich vollständig in Ihren Katalog oder System übernehmen? Wollen Sie z. B. den Bestand Ihrer Bibliothek mit Hilfe der GND erschließen, dann könnten Sie diese vollständig in Ihren Katalog übernehmen und dort auch laufend aktuell halten oder wollen Sie gezielt nach bestimmten Suchkriterien zusammengestellte Daten für sich nutzbar machen? Zum Beispiel die Publikationen eines Autors, zu einem Thema, eines bestimmten Verlages. Oder haben Sie eine eigene kleine Bibliothek und wollen jeweils für die Medien in Ihrem Bestand die Metadaten bei der DNB prüfen und übernehmen? Dann benötigen Sie einen selektiven Zugang.

- **Laufend oder einmalig**
  Ein Teil unserer Angebote ist darauf ausgerichtet, dass Sie, gerade wenn Sie einmal die Metadaten vollständig übernommen haben, diese Daten immer aktuell halten. Dazu können Sie laufend von uns die Änderungen erhalten. Leisten kann das insbesondere die OAI-Schnittstelle. Aber auch die Zugänge, die Sie unter dem Punkt „Laufende Aktualisierung von Metadaten" finden. Natürlich können auch die genannten Zugänge einmalig genutzt werden, angelegt sind Sie aber eher für eine laufende Nutzung. Umgekehrt ist auch vorstellbar, dass Sie etwa über den Datenshop oder SRU regelmäßig nach Neuerscheinungen zu einem bestimmten Thema suchen. Die Regelmäßigkeit muss aber quasi durch Sie von Hand geleistet werden.

- **Zugang**
  Nicht jedes Format ist über jeden Bezugsweg verfügbar. Dazu finden Sie auf https://www.dnb.de/metadatenbezugswege einen guten Überblick

LEIPZIG
FRANKFURT AM MAIN

	Schnittstellen	Online-Verarbeitung	Immer aktuell
Vorkonfiguriert			
Offline-Verarbeitung			
Aktualität entspricht dem Datum des Abzugs			
Bereitstellung auf WWW- und SFTP-Servern			

Datenshop dnb.de/datenshophilfe

Technische Voraussetzungen	Wenig, z.B. einfach Importmöglichkeit in eigene Anwendung		
Format	csv, oai_dc, MARC 21, MARC21-xml, RDF		
vollständig	⊗	selektiv	☺
einmalig	☺	laufend	☹

Gesamtabzüge dnb.de/gesamtabzuege

Technische Voraussetzungen	Sehr große Datenmengen		
Format	MARC 21, MARC21-xml, RDF		
vollständig	☺	selektiv	⊗
einmalig	☺	laufend	☹ (regelmäßig aktualisiert)

SRU-Schnittstelle dnb.de/sru

Technische Voraussetzungen	Über Browser oder eingebunden in eigene Anwendung		
Format	oai_dc, MARC21-xml, RDF, METS/MODS		
vollständig	⊗	selektiv	☺
einmalig	☺	laufend	☹

Nationalbibliografie, GND- und ZDB-Änderungsdienst
dnb.de/metadatenbezug und dnb.de/gndaenderungsdienst

Technische Voraussetzungen	Einfacher Download		
Format	MARC 21, MARC21-xml		
vollständig	☹ aufsetzend auf Grundbestand	selektiv	⊗
einmalig	⊗	laufend	☺ Regelmäßig bereitgestellt

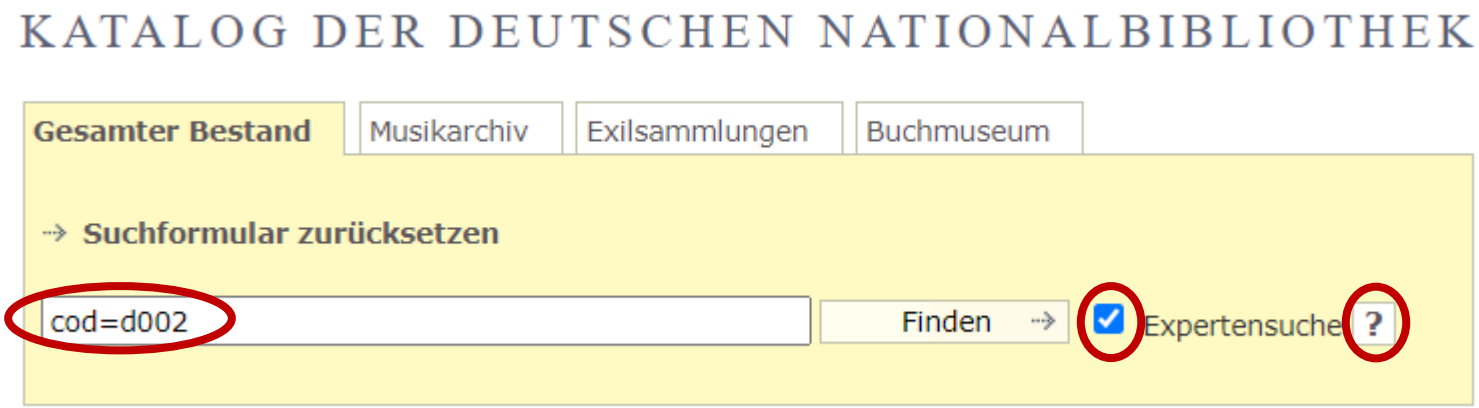
OAI-Schnittstelle
dnb.de/oai bzw. dnb.de/dnblabapi (OAI2 für Mets/Mods)

Technische Voraussetzungen	OAI-Harvester, nur testweise über Browser		
Format	oai_dc, MARC21-xml, RDF, METS/MODS		
vollständig	☺ Aufbauend auf Grundbestand	selektiv	☹ Nach vorgegebenen Sets
einmalig	⊗	laufend	☺

SRU dnb.de/sru

Technische Voraussetzungen	➲ Über Browser oder eingebunden in eigene Anwendung		
Format	oai_dc, MARC21-xml, RDF, METS/MODS		
vollständig	☹	selektiv	☺
einmalig	☺	laufend	😐

portal.dnb.de/



DEUTSCHE

Ergebnis der Suche nach: *cod=d002*
im Bestand: Gesamter Bestand

1 - 10 von 108 Datum (neuestes zuerst) sortieren

1. Egmont
   Goethe, Johann Wolfgang von. - Leipzig : Dt. Nationalbibliothek, [1946], [Online-Ausg.]
2. Das Amulett
   Meyer, Conrad Ferdinand. - Leipzig : Dt. Nationalbibliothek, [1939], [Online-Ausg.]
3. Der Struwwelpeter oder lustige Geschichten und drollige Bilder für Kinder von 3 - 6 Jahren
   Hoffmann, Heinrich. - Leipzig : Dt. Nationalbibliothek, [1939], [Online-Ausg.]
4. Der Zweikampf
   Kleist, Heinrich von. - Leipzig : Dt. Nationalbibliothek, 1939, [Online-Ausg.]
5. Heidi
   Spyri, Johanna. - Leipzig : Dt. Nationalbibliothek, 1939, [Online-Ausg.]
6. Goethes Werke / Bd. 8. Die Wahlverwandschaften
   [1938]
7. Der Schwierige
   Hofmannsthal, Hugo von. - Leipzig : Dt. Nationalbibliothek, [ca. 1935], [Online-Ausg.]
8. Frau Jenny Treibel
   Fontane, Theodor. - Leipzig : Dt. Nationalbibliothek, [1935?], [Online-Ausg.]

Alle Materialarten

- Online Ressourcen (108)

Alle Standorte

- Frankfurt (108)
- Leipzig (108)
- Online (frei zugänglich) (105)

Aktionen

- In meine Auswahl übernehmen
- Druckansicht
- Versenden
- News-Feed

dnb.de/expertensuche

Hilfe: Expertensuche

Nummernsuche +

Thematische Suche +

Räumliche Suche +

Spezielle Indizes für Musik +

Weitere Indizes −

cod

Codes

abk

Abkürzungen

hss

Suche mit Stichwörtern (Hochschule, Abschlussjahr) aus dem **Hochschulschriftenvermerk**. Beispiel: hss=Freiburg AND jhr=2005

Syntax einer SRU-Anfrage	
Beispiel	**Erklärung**
https://services.dnb.de/sru	Anfrage an SRU-Server der DNB
/dnb	Festlegen des Kataloges (z. B. DNB)
?version=1.1	Angabe der SRU-Version, Standard
&operation=searchRetrieve	Befehl an den Server
&query=	Anfrage formulieren
WOE%3DGoethe	WOE ist die Indexbezeichnung, %3D ist die URL-Kodierung für =, Suchbegriff hier: Goethe
%20sortby%20tit/sort.ascending	Sortierung der Ergebnismenge nach Titel aufsteigend von A-Z
&recordSchema=MARC21-xml	gewünschtes Format der SRU-Antwort

URL-Kodierung

In der **URL-Kodierung** werden Umlaute nach UTF-8 und mit vorangestelltem %-Zeichen kodiert (Prozentkodierung).

Beispiel: Kleinbuchstabe ö in UTF-8 = C3 B6 und mit vorangestellten %-Zeichen = %C3%B6.

URL-Kodierung		
Zeichen		**UTF-8 hexadezimal**
Ä	=	%C3%84
Ö	=	%C3%96
Ü	=	%C3%9C
ä	=	%C3%A4
ö	=	%C3%B6
ü	=	%C3%BC
ß	=	%C3%9F
(Leerschritt, Spatium)	=	%20
" (Anführungszeichen)	=	%22
/ (Slash)	=	%2F
\ (Backslash)	=	%5C
Liste weiterer Zeichen		

SRU-Abfrage im Browser

https://services.dnb.de/sru/dnb?version=1.1&operation=searchRetrieve&query=cod%3Dd002&recordSchema=MARC21-xml

Beispiel SRU-Antwort

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<searchRetrieveResponse xmlns="http://www.loc.gov/zing/srw/">
  <version>1.1</version>
  <numberOfRecords>108</numberOfRecords>
 ▼<records>
  ▼<record>
    <recordSchema>MARC21-xml</recordSchema>
    <recordPacking>xml</recordPacking>
   ▼<recordData>
    ▼<record xmlns="http://www.loc.gov/MARC21/slim" type="Bibliographic">
      <leader>00000nam a2200000 c 4500</leader>
      <controlfield tag="001">1003104487</controlfield>
      <controlfield tag="003">DE-101</controlfield>
      <controlfield tag="005">20240215152134.0</controlfield>
      <controlfield tag="007">cr|||||||||||||</controlfield>
      <controlfield tag="008">100601s1946 gw |||||o|||| 00||||ger </controlfield>
     ▼<datafield tag="015" ind1=" " ind2=" ">
        <subfield code="a">10,O10</subfield>
        <subfield code="2">dnb</subfield>
      </datafield>
     ▼<datafield tag="016" ind1="7" ind2=" ">
        <subfield code="2">DE-101</subfield>
        <subfield code="a">1003104487</subfield>
      </datafield>
     ▼<datafield tag="024" ind1="7" ind2=" ">
        <subfield code="2">urn</subfield>
        <subfield code="a">urn:nbn:de:101:2-2010060101</subfield>
      </datafield>
```

Beispiel SRU-Antwort


```
▼<datafield tag="856" ind1="4" ind2="0">
   <subfield code="u">https://nbn-resolving.org/urn:nbn:de:101:2-2010011375</subfield>
   <subfield code="x">Resolving-System</subfield>
 </datafield>
▼<datafield tag="856" ind1=" " ind2="0">
   <subfield code="u">https://d-nb.info/999490184/34</subfield>
   <subfield code="x">Langzeitarchivierung Nationalbibliothek</subfield>
 </datafield>
▼<datafield tag="889" ind1=" " ind2=" ">
   <subfield code="w">(DE-101)1005985162</subfield>
 </datafield>
▼<datafield tag="925" ind1="r" ind2=" ">
   <subfield code="a">ro</subfield>
   <subfield code="a">rb</subfield>
 </datafield>
▼<datafield tag="925" ind1="d" ind2=" ">
   <subfield code="a">d002</subfield>
 </datafield>
▼<datafield tag="927" ind1=" " ind2=" ">
   <subfield code="a">40960003</subfield>
 </datafield>
</record>
</recordData>
<recordPosition>2</recordPosition>
</record>
```

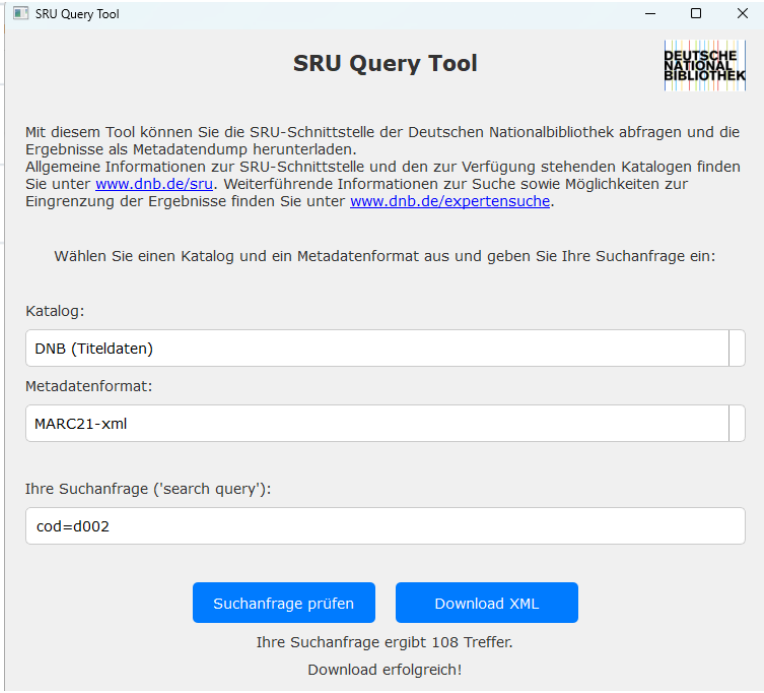
DNB SRU Query Tool

https://github.com/deutsche-nationalbibliothek/SRUQueryTool

DNB_SRUQueryTool-x86_64.AppImage

DNB_SRUQueryTool.dmg

DNB_SRUQueryTool.exe



Tutorial (Jupyter Notebook)

https://www.dnb.de/dnblabtutorials

Wie können Daten mittels Open Refine aufbereitet und angereichert werden?

Bitte verwenden Sie die statische Version zum Lesen der Tutorials. Um den Code in den Tutorials aktiv ausführen und bei Bedarf anpassen zu können, nutzen Sie bitte die interaktive Version.

- Statische Version: **Anzeige im Browser**
- Interaktive Version: **Starten in Binder**
- Dateien im ipynb Format: **GitHub**

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt:

f.fuchs@dnb.de

a.taube@dnb.de

oder

lab@dnb.de