

FAIRe Forschungsdaten

Aufbereitung von Daten mit Hilfe von RDF-Datenmodellen

Arnošt Štanzel, Anna-Lena Körfer, Ingo Frank, Sandra König



Gefördert durch
DFG Deutsche
Forschungsgemeinschaft

Workshopreihe „Digitales Praxislabor Geschichtswissenschaft“
der AG Digitale Geschichtswissenschaft im VHD

24. Mai 2022

Überblick

- ① Einführung
- ② Vorstellungsrunde
- ③ Theorie: Semantische Datenmodellierung, Normdaten und Wissensorganisationssysteme
- ④ Kaffeepause
- ⑤ Praxis: Tutorial Datenaufbereitung und Erstellung kontrollierter Vokabulare
- ⑥ Abschlussdiskussion / Q & A

Einführung

Problemstellung Modellierung von FAIRen Daten für Historiker*innen I

Problemstellung

Problem von unvollständig oder unpassend aufbereiteten Daten:
Daten sind nicht ohne weitere Aufbereitung nachnutzbar bzw.
interoperabel.

Interoperabilität

Die Daten sollten so vorliegen, dass sie ausgetauscht, interpretiert und in einer (semi-)automatisierten Weise mit anderen Datensätzen von Menschen sowie Computersystemen kombiniert werden können.¹

Problemstellung Modellierung von FAIRen Daten für Historiker*innen II

Nachnutzbarkeit

Eine gute Beschreibung von Daten und Metadaten sorgt dafür, dass die Daten für die zukünftige Forschung wiederverwendet werden können und mit anderen, kompatiblen Datenquellen vergleichbar sind.²

¹<https://blogs.tib.eu/wp/tib/2017/09/12/die-fair-data-prinzipien-fuer-forschungsdaten/#i>

²<https://blogs.tib.eu/wp/tib/2017/09/12/die-fair-data-prinzipien-fuer-forschungsdaten/#r>

Ansatz

Ansatz

Um Forschungsdaten **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) zu machen, bietet sich an, diese als **Linked (Open)** **Data** mit **RDF-Datenmodellen** zu modellieren.

Vorgehensweise

Im Workshop werden wir anhand von ausgewählten Daten aus Projekten und veröffentlichten Datensätzen der OstData-Partnerinstitute zeigen, wie Forschende mit Hilfe von RDF, semantischer Technologien und Normdaten die Aufbereitung von Forschungsdaten durchführen können.

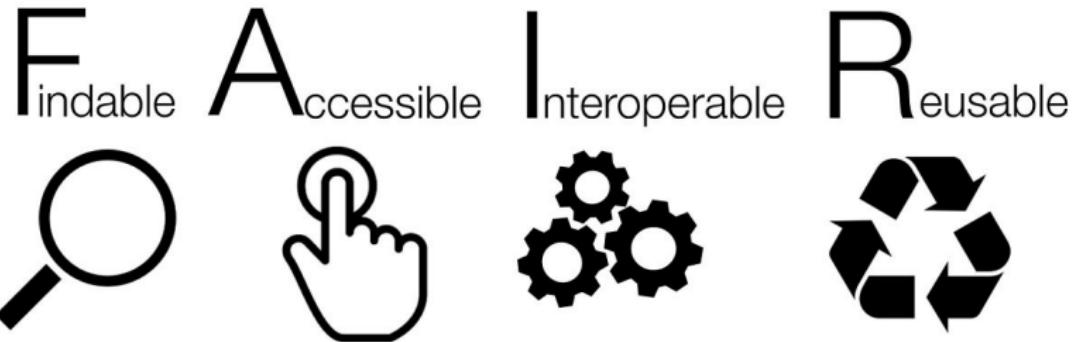


Abbildung 1: Grafik “ FAIR guiding principles for data resources”: https://de.wikipedia.org/wiki/Datei:FAIR_data_principles.jpg

FAIR nicht als Buzzword, sondern als Prinzipien zur FAIRifikation von Forschungsdaten I

Interoperability – FAIR-Prinzipien & Umsetzung

- I1 & I2 Bereitstellung der Daten in gängigen Formaten (Datendistributionen) – ggf. aufgeteilt nach verschiedenen inhaltlichen Kriterien (z. B. Berichtszeitraum)³.
- I3 Einsatz und konsistente Verwendung von möglichst standardisierter Kodierung zur Harmonisierung von Daten in Sammlungen von Datensätzen.

FAIR nicht als Buzzword, sondern als Prinzipien zur FAIRifikation von Forschungsdaten II

Reusability – FAIR-Prinzipien & Umsetzung

- R1.2 Herkunft von Forschungsdaten (besonders relevant bei Sekundärdaten) durch detaillierte Provenienzinformation⁴ nachvollziehbar machen.
- R1.2 Entstehungskontext und Sinnzusammenhang der Forschungsdaten kurz beschreiben.
- R1.3 Möglichst fachspezifisch gängige [geeignete/passende/praktikable!] Datenmodelle und Standard-konforme Metadatenschemata zur Modellierung und Beschreibung von Forschungsdaten verwenden.

³siehe dazu <https://www.w3.org/TR/dcat-ucr/#ID34>

⁴siehe dazu <https://www.w3.org/TR/dcat-ucr/#ID12>

FAIRification von Forschungsdaten

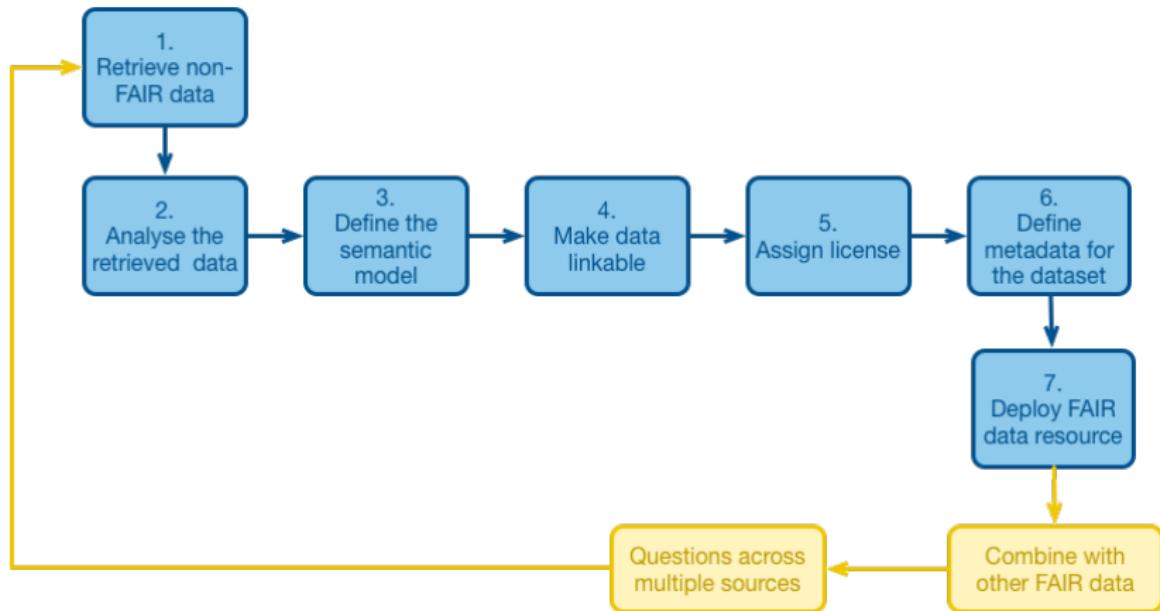


Abbildung 2: Relevanz von RDF-Datenmodellen, Semantic Web und Linked Data in den Arbeitsschritten 3 und 4 im FAIRification Process⁵

⁵<https://www.go-fair.org/fair-principles/fairification-process/>

Ziele des Workshops

Ausgangsfrage

Wie numerische/statistische Daten einheitlich und damit interoperabel und nachnutzbar (im Sinne von FAIR) modellieren?

Theorieteil

- Bewusstsein für Probleme der Interoperabilität und Nachnutzbarkeit von Daten schaffen
- Grundlegendes Verständnis für Modellierung maschinenlesbarer Daten mit RDF-Datenmodellen vermitteln

Praxisteil

- Aufbereitung statistischer Daten mit dem RDF Data Cube Vocabulary demonstrieren
- Anreicherung der aufbereiteten Daten mit Normdaten und kontrollierten Vokabularen durchführen

Vorstellungsrunde

Unser Projektkontext

Forschungsdatendienst OstData

Forschungsdatendienst für die Ost-, Ostmittel- und Südosteuropaforschung

OSTDATA

Im Projekt werden bestehende Daten aus Forschungsprojekten der Partnerinstitute aufbereitet und als Forschungsdaten im Forschungsdatenrepositorium von OstData publiziert.

Gefördert durch



Aufbereitung und Modellierung von Altdaten als 'Linked Research Data'

Historische Zensusdaten aus Ethnodoc

- Das Projekt „Datenbank zur Minderheitenproblematik und zu den ethnischen Gruppen Südosteuropas“ (Ethnodoc)⁶ wurde vom Forschungsverbund Ost- und Südosteuropa (forost) am Südost-Institut München (Vorgängerinstitut des IOS) durchgeführt.
- Die Sammlung bietet “facts and figures” über mehr als 40 Minderheiten in Südosteuropa.
- Kodierung der ethnisch, sprachlich und religiös definierten Gruppen in den Daten.
- Wie kann man tabellarische Daten anders als in Form von Excel-Tabellen aufbereiten, bearbeiten und nicht zuletzt für anderweitige Nachnutzung zur Verfügung stellen?

⁶http://www.forost.lmu.de/forosthope/fo_pro_III1.htm

Vorstellung der Workshop-Teilnehmer*innen

Link zum Etherpad

https://yopad.eu/p/0stData-Workshop_RDF-Datenmodellierung-1day

Theorie: Semantische Datenmodellierung, Normdaten und Wissensorganisationssysteme

Linked Data Intertwingularity (Fortschritt durch Rückblick)

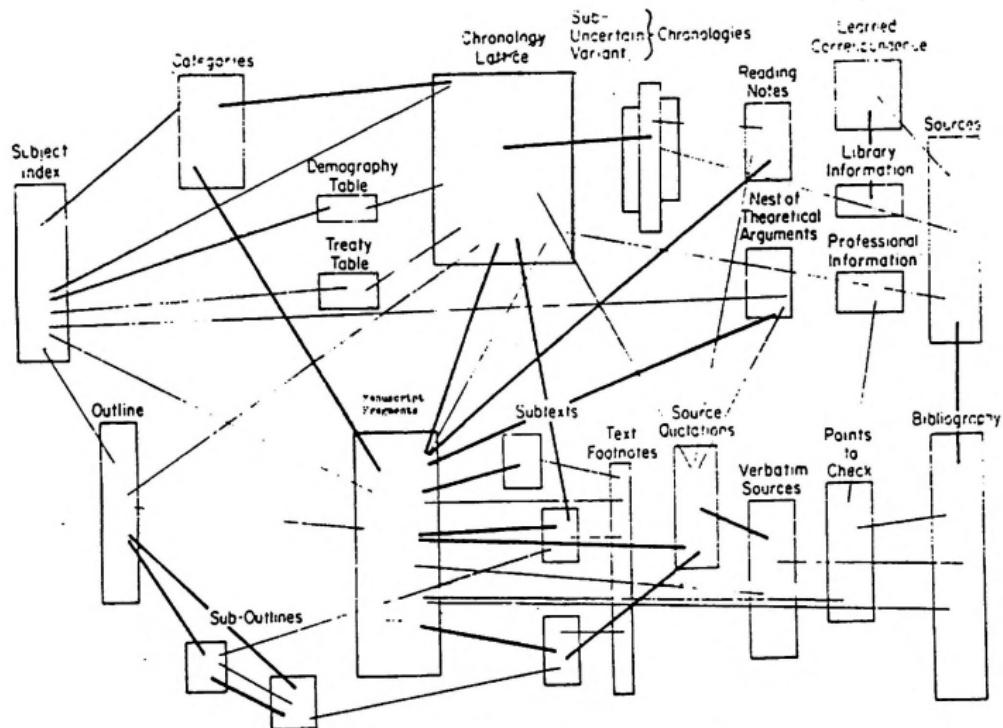


Abbildung 3: Vision einer Hypertext-Arbeitsumgebung (inkl. Text und Daten) für die Digitale Geschichtswissenschaft (aus Nelson 1965, S. 100)

Der Nutzen von RDF-Datenmodellierung für historische Forschung

- ① RDF (Resource Description Framework) als grundlegendes Datenmodell von Semantic Web-Technologien
- ② Linked Data-Paradigma
- ③ Datenmodellierung mit Semantic Web-Technologien gemäß dem Linked Data-Ansatz
- ④ Datenmodellierung ist nicht Wissensrepräsentation:
Anwendungsprofile vs. Ontologien
- ⑤ Möglichkeit von Datenväldierung mit SHACL (Shapes Constraint Language)



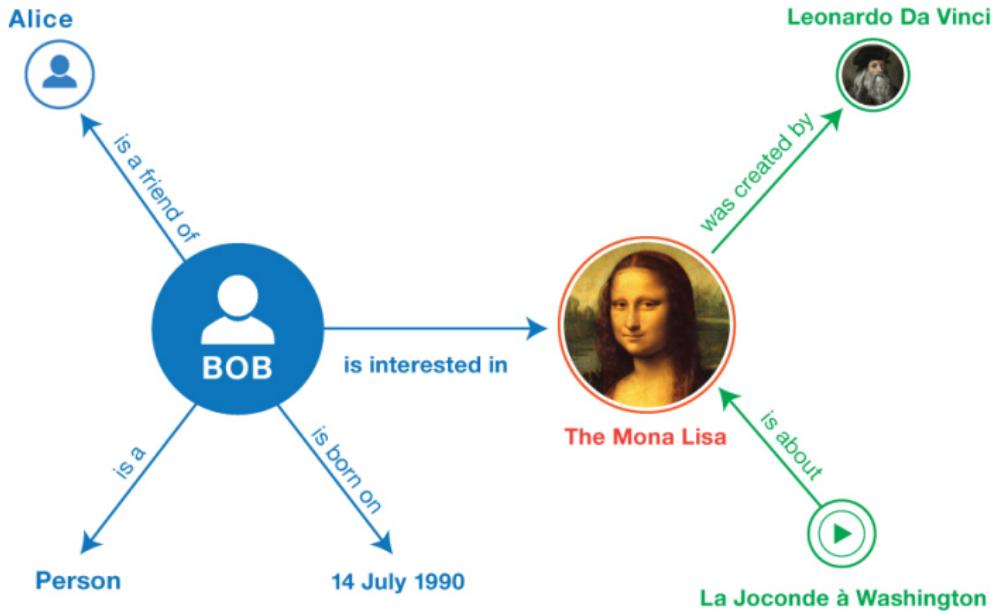


Abbildung 4: Beispiel für Tripel-Struktur von RDF-Daten – aus <https://www.w3.org/TR/rdf11-primer/#fig1>

Alice

<http://example.org/alice#me>



foaf:knows



foaf:topic_interest



The Mona Lisa

<http://www.wikidata.org/entity/Q12418>

dcterms:creator

dcterms:title



"Mona Lisa"

dcterms:subject



La Joconde à Washington

<http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>

Person
foaf:Person

rdf:type

schema:birthDate

"1990-07-04"^^xsd:date

Abbildung 5: RDF-Graph mit URIs, Eigenschaften aus RDF-Vokabularen und Datentypen – aus <https://www.w3.org/TR/rdf11-primer/#fig4>

RDF-Daten in Turtle-Syntax (Terse RDF Triple Language):⁷

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX schema: <http://schema.org/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX wd: <http://www.wikidata.org/entity/>

<bob#me>
  a foaf:Person ;
  foaf:knows <alice#me> ;
  schema:birthDate "1990-07-04"^^xsd:date ;
  foaf:topic_interest wd:Q12418 .

wd:Q12418
  dcterms:title "Mona Lisa" ;
  dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .

<http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
  dcterms:subject wd:Q12418 .
```

⁷<https://www.w3.org/TR/rdf11-primer/#section-turtle>

Linked Data-Paradigma I

Linked Data (als Daten)

In computing, linked data (often capitalized as Linked Data) is structured data which is interlinked with other data so it becomes more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URLs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers.⁸

Linked Data (als Paradigma)

The term *Linked Data* refers to a set of best practices for publishing and interlinking structured data on the Web.⁹

Linked Data-Paradigma II

Gute Charakterisierung des Linked Data-Paradigmas (inkl. Abgrenzung Datenmodellierung/Datenmodelle von Wissensrepräsentation/Ontologien) von Kalampokis, Zeginis und Tarabanis (2019) im Aufsatz “On modeling linked open statistical data”, S. 58:

Linked data paradigm is based on semantic Web philosophy and technologies but it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing [...] A good approach in linked data is to re-use standard vocabularies to encode data and meta-data [...]

Linked Data-Paradigma III

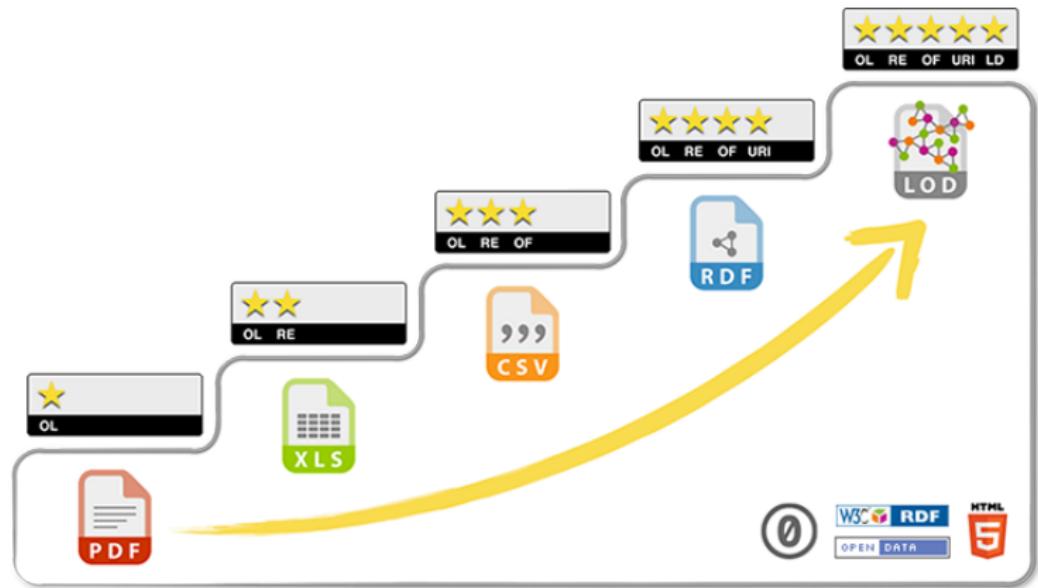


Abbildung 6: Von der Quelle oder den 'Rohdaten, zu Linked Open Data¹¹

⁸https://en.wikipedia.org/wiki/Linked_data

⁹<http://linkeddatabook.com/editions/1.0/#htoc8>

¹⁰<https://5stardata.info/de/>

¹¹<https://5stardata.info/de/>

Datenmodellierung I

Linked Data-Ansatz zur Datenmodellierung

Wie kann das Linked Data-Paradigma angewendet werden, um Forschungsdaten FAIR für Forschung in der Digitalen Geschichtswissenschaft zu machen?

- Das Linked Data-Paradigma stellt eine Lösung für Quellen-nahe Modellierung (vgl. Harvey und Press 1996, p 189 f.) und Integration historischer Quellen – als Forschungsdaten – bereit.
- Der Linked Data-Ansatz ist recht naheliegend für die Aufbereitung von historischen Zensusdaten und Survey-Daten (Boer, Meroño-Peñuela und Ockeloen 2016).
- Schwerpunkt auf Interoperabilität, durch Integration und entsprechende Distribution von FAIRifizierten Daten (z. B. historische Zensusdaten, historische Karten und weitere relevante Geodaten)

RDF Data Cube Vocabulary für statistische Daten

- Das RDF Data Cube Vocabulary (QB) ermöglicht “to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts”.¹²
- Wir verwenden QB zur Modellierung von Beobachtungen in Zensusdaten und Geodaten.
- Durch die einheitliche Modellierung von Daten mit QB wird insbesondere die Nachnutzung und Interoperabilität der Daten verbessert.

¹²<http://www.w3.org/TR/vocab-data-cube/>

- Die Messung repräsentiert das beobachtete Phänomen (z. B. Bevölkerung).
- Zur Messung gibt es ein Attribut zur Angabe der Maßeinheit der Messung (z. B. ‚Person‘ für Messung der Bevölkerung) und ggf. auch ein Attribut für den Status (z. B. ‚geschätzt‘).
- Dimensionen charakterisieren die Messung (typisch sind die Dimensionen für den zeitlichen und räumlichen Bezug).

Verortung von Daten im Datenwürfel I

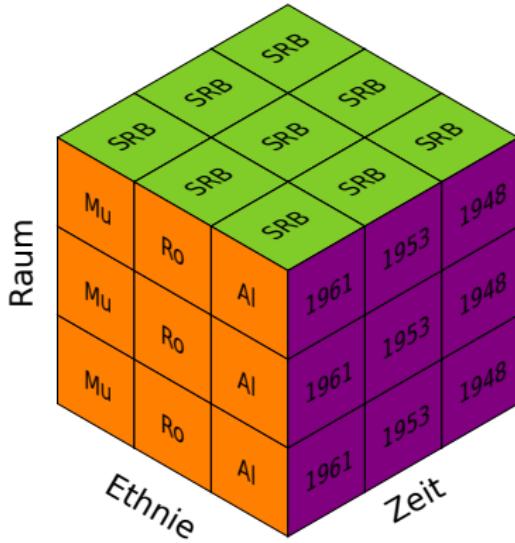


Abbildung 7: Auszug aus historischen Zensusdaten¹¹ in einem intuitiv verständlichen Data Cube mit den Dimensionen Raum (räumlicher Bezug auf Landesebene anhand IOC-Ländercodes), Zeit (zeitlicher Bezug: Jahr der jugoslawischen Volkszählung) und ethnische Zugehörigkeit

Verortung von Daten im Datenwürfel II

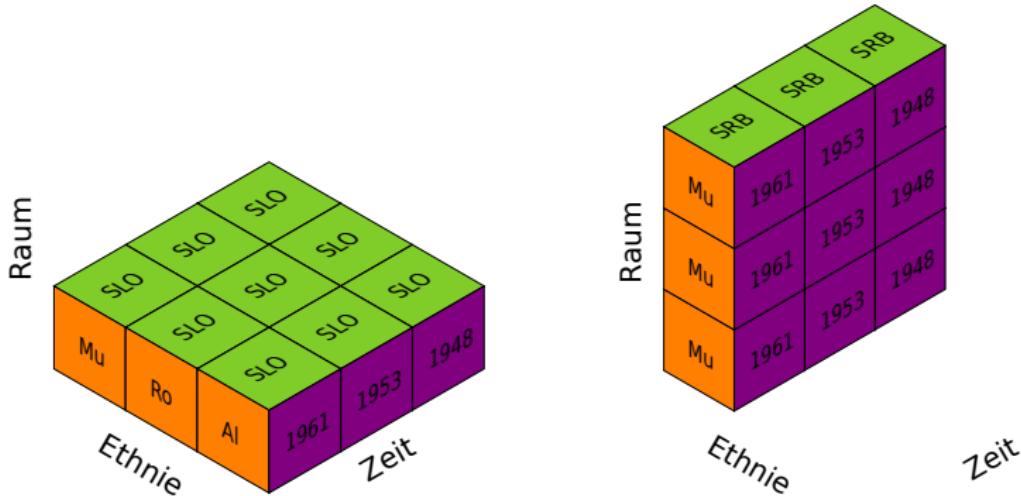
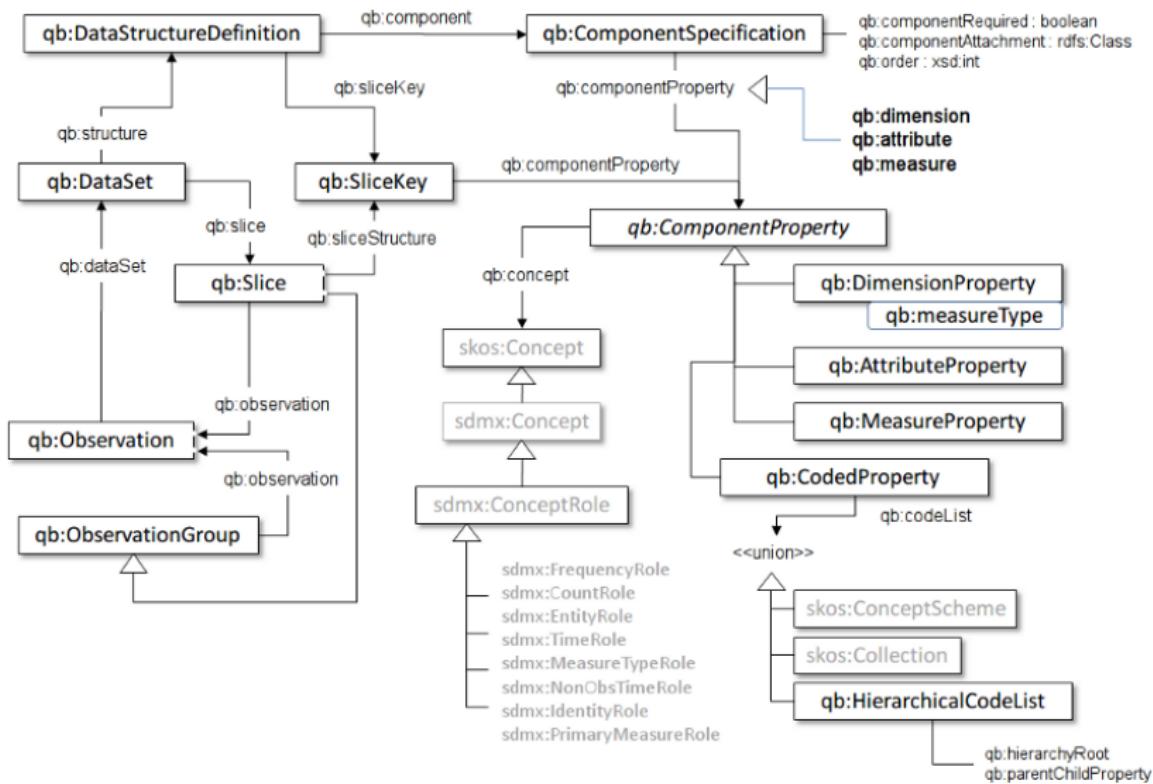


Abbildung 8: Slices im Data Cube – nachvollziehbare Sichten auf die Daten aus dem vorherigen Beispiel durch ‚Herausschneiden‘ von Scheiben aus dem Würfel

¹¹Quellen: https://en.wikipedia.org/wiki/Demographics_of_Serbia, https://en.wikipedia.org/wiki/Demographics_of_Croatia und https://en.wikipedia.org/wiki/Demographics_of_Slovenia

RDF Data Cube Vocabulary in a Nutshell (aus Doku) I



RDF Data Cube Vocabulary in a Nutshell (aus Doku) II

A **statistical data** set comprises a **collection of observations** made at some points across some logical space. The collection can be **characterized by a set of dimensions** that define what the observation applies to (e.g. time, area, gender) along with **meta-data** describing what has been measured (e. g. economic activity, population), how it was measured and how the observations are expressed (e. g. units, multipliers, status). We can think of the statistical data set as a multi-dimensional space, or hyper-cube, indexed by those dimensions.

RDF Data Cube Vocabulary in a Nutshell (aus Doku) III

A **cube** is organized according to a set of **dimensions**, **attributes** and **measures**. We collectively call these **components**.

The **dimension** components serve to identify the observations. A set of values for all the dimension components is sufficient to identify a single observation. Examples of dimensions include the time to which the observation applies, or a geographic region which the observation covers.

The **measure** components represent the phenomenon being observed.

The **attribute** components allow us to qualify and interpret the observed value(s). They enable specification of the units of measure, any scaling factors and metadata such as the status of the observation (e.g. *estimated*, *provisional*).¹²

¹²<https://www.w3.org/TR/vocab-data-cube/#cubes-model>

Ausgangslage tabellarische Daten

Abbildung 9: Typisches Beispiel aus der Ethnodoc-Sammlung: <https://lambda.ios-regensburg.de/dataset/ethnodoc-yu-1948-81-1>

Anforderung: saubere Daten für Aufbereitung und QB-Erstellung

Typisches Problem

Spaltenüberschriften sind Werte und nicht die Namen von Variablen!

Tabelle mit bereinigten Daten ('tidy data' nach Wickham (2014)):

- Jede Zeile entspricht einer Beobachtung (qb:Observation).
- Je Spalte eine Dimension, Messung (gemessene Variable) oder Attribut.

Bereinigung und Aufbereitung der tabellarischen Daten I

Beobachtung

Im Jahr **1948** (Dimension Zeit) gab es in **Jugoslawien** (Dimension Raum) eine Bevölkerung von **55337** (Messung) **Personen** (Attribut/Maßeinheit) mit ethnischer Zugehörigkeit **Deutsch** (Dimension ethnische Zugehörigkeit).

refArea	ethnicity	refPeriod	population
Jugoslawien	Muslime	1948	808921
Jugoslawien	Deutsche	1948	55337
Jugoslawien	Zigeuner	1948	72736
Jugoslawien	Muslime	1953	998698
Jugoslawien	Deutsche	1953	60536
Jugoslawien	Zigeuner	1953	84713
Jugoslawien	Muslime	1961	972960
Jugoslawien	Deutsche	1961	20015
Jugoslawien	Zigeuner	1961	31674

Bereinigung und Aufbereitung der tabellarischen Daten II

refArea	ethnicity	refPeriod	population
Q36704	37	1948	808921
Q36704	15	1948	55337
Q36704	77	1948	72736
Q36704	37	1953	998698
Q36704	15	1953	60536
Q36704	77	1953	84713
Q36704	37	1961	972960
Q36704	15	1961	20015
Q36704	77	1961	31674

Modellierung historischer Zensusdaten in QB I

Diese Art von aggregierten Zensusdaten kann wie folgt in QB modelliert werden:

```
:obs1 a qb:Observation ;
  qb:dataset :cube ;
  sdmx-dimension:refArea wd:Q36704 ;
  sdmx-dimension:refPeriod "1948"^^xsd:gYear ;
  :ethnicity ethnodoc-code:35 ;
  :population "425703"^^xsd:integer ;
  qb:measureType :population ;
  sdmx-attribute:unitMeasure qudt:Person .
```

Best Practices für QB-Anwendungsprofile

Im Aufsatz “On modeling linked open statistical data” liefern Kalampokis, Zeginis und Tarabanis (2019) viele Modellierungsbeispiele und begründete Best Practices (quasi Anwendungsprofile für QB).

Modellierung historischer Zensusdaten in QB II

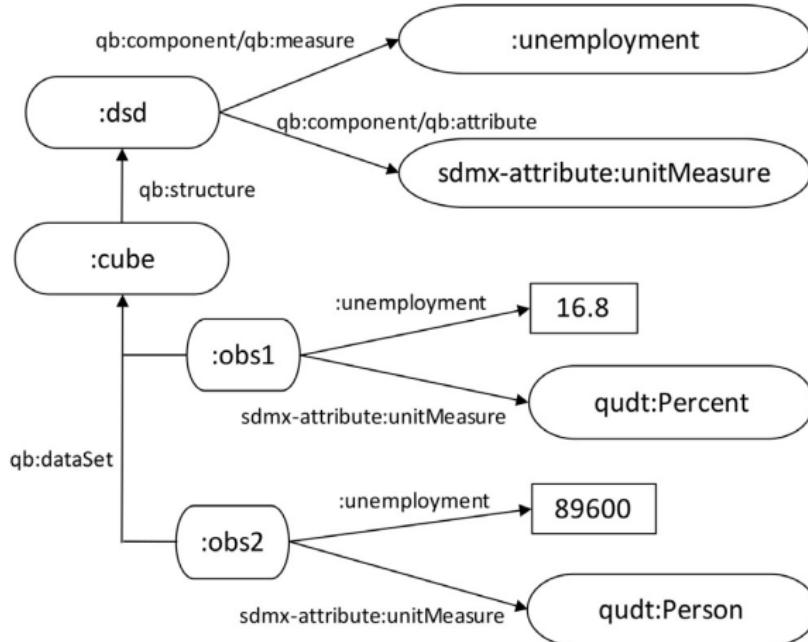


Abbildung 10: Modellierung einer Messung mit verschiedenen Einheiten (z. B. Bevölkerung in Anzahl Personen und in Prozent) (aus Kalampokis, Zeginis und Tarabanis 2019, S. 60)

SKOS (Simple Knowledge Organization System) I

Kurze Charakterisierung von SKOS aus “Key choices in the design of Simple Knowledge Organization System (SKOS)” (Baker et al. 2013, S. 37):

Using SKOS, **concepts** can be identified using URIs, **labeled** with lexical strings in one or more natural languages, assigned **notations** (lexical codes), **documented** with various types of note, **linked to other concepts** and organized into informal hierarchies and association networks, aggregated into **concept schemes**, grouped into labeled and/or ordered **collections**, and **mapped** to concepts in other schemes.

SKOS (Simple Knowledge Organization System) II

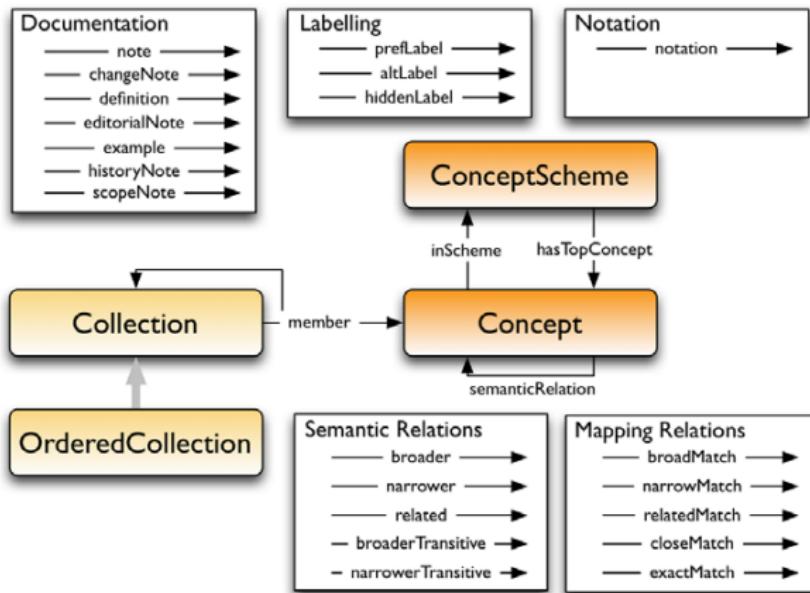


Abbildung 11: Hauptelemente des SKOS-Datenmodells (aus Baker et al. 2013, S. 38)¹⁴

¹⁴ Siehe auch: https://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System#Element_categories

XKOS (eXtended Knowledge Organization System)

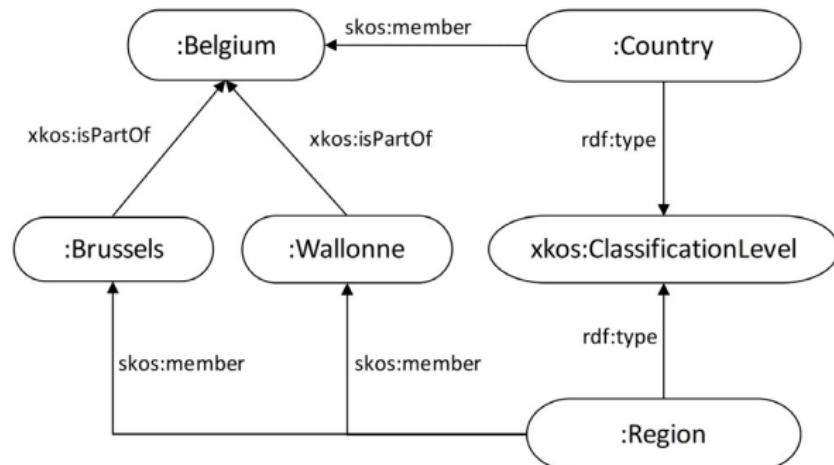


Abbildung 12: XKOS-Modellierungsbeispiel für ein typisches hierarchisches Kodierschema (aus Kalampokis, Zeginis und Tarabanis 2019, S. 65)

NUTS (Nomenclature of territorial units for statistics)

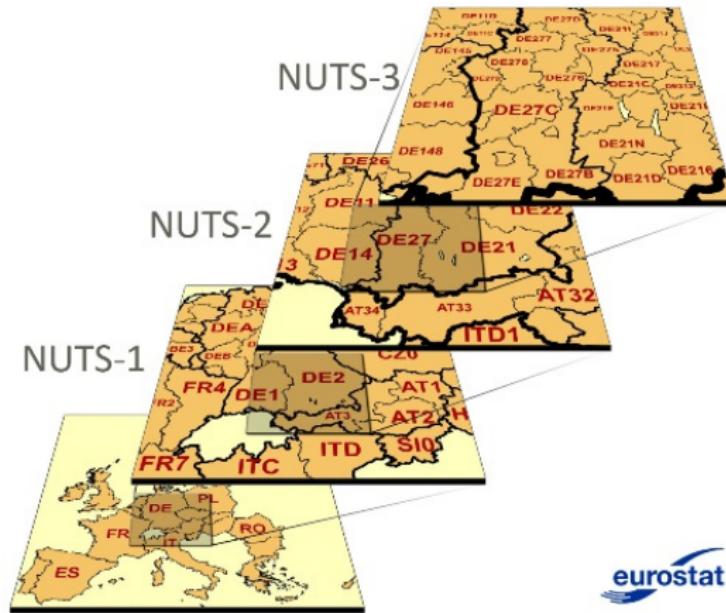


Abbildung 13: NUTS (Nomenclature of territorial units for statistics) zur Veranschaulichung eines hierarchischen Klassifikationssystems zur Kodierung von Regionen auf verschiedenen Ebenen (räumlicher Bezug amtlicher Statistiken):

<https://ec.europa.eu/eurostat/web/nuts/background>

Kodierlisten und statistische Klassifikationssysteme mit SKOS, XKOS und SKOS-XL erstellen

SKOS (Simple Knowledge Organization System) kann mit Klassen und Eigenschaften aus

XKOS (eXtended Knowledge Organization System) erweitert werden, um die für statistische Klassifikationssysteme typischen Ebenen (z. B. Verwaltungseinheiten) oder auch die folgenden Relationen zwischen Begriffen zu modellieren:

- generische and mereologische Relationen
- sequentielle, temporale und kausale Relationen

SKOS-XL (SKOS eXtension for Labels) kann verwendet werden, um die Bezeichner von Begriffen nicht als simple Literale, sondern als Instanzen der Klassen `skosxl:Label` und `skosxl:altLabel` sowie Relationen zwischen ihnen (z. B. `custom:previousLabel` als Untereigenschaften von `skosxl:labelRelation`).

Kaffeepause

WE SHOULD MAKE COFFEE FOR OUR GUESTS.
/ CRAP. I KNOW NOTHING ABOUT COFFEE.
WE'RE BASICALLY FAKE ADULTS. /



WE JUST
POUR THE
COFFEE
GROUNDS...



...ADD
WATER...



NOW WE JUST
HOLD IT OVER
THE BURNERS...



ANNNND... SERVE.

/ NICE!
I'M A REGULAR
/ STARBUCK!



Praxis: Tutorial Datenaufbereitung und Erstellung kontrollierter Vokabulare

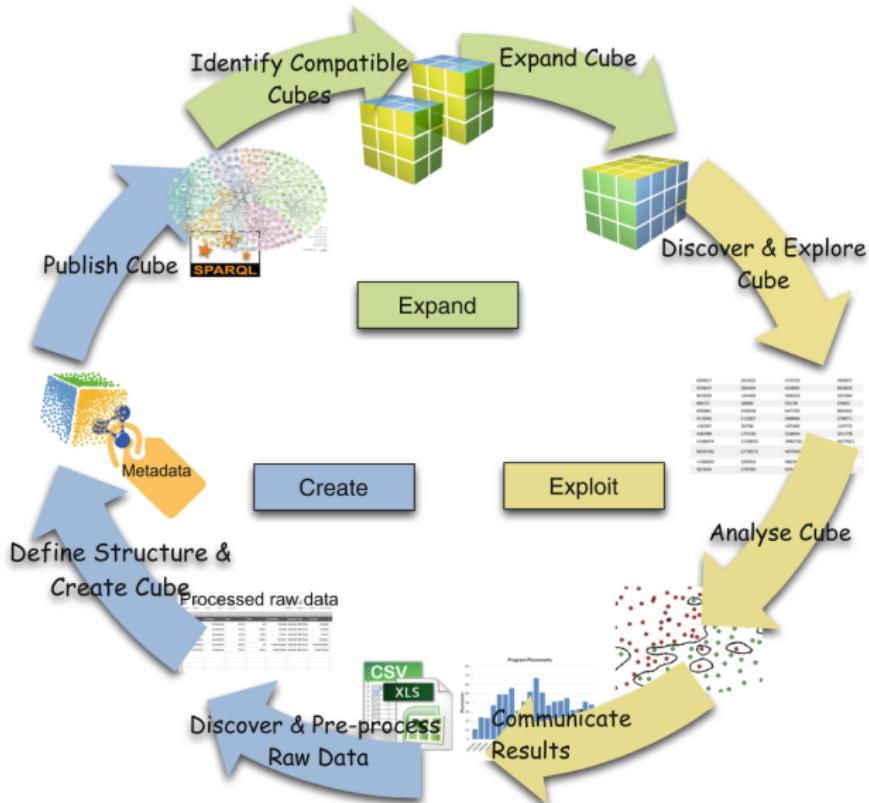


Abbildung 14: (Forschungs-)Datenlebenszyklus zur Erstellung, Verknüpfung und Nachnutzung von statistischen Daten als Linked Open Data (aus Tambouris, Kalampokis und Tarabanis 2015)

Datenaufbereitung mit table2qb

Für den Einsatz des Werkzeugs table2qb werden die folgenden CSV-Dateien gebraucht und müssen auf Basis der Ausgangsdaten vorbereitet werden:

`observations.csv` mit den bereinigten Daten (jede Zeile enthält Angaben einer Observation).

`components.csv` legt die Komponenten des Datenmodells (für die Observations) fest, d. h. die Dimensionen, Messungen und zugehörigen Attribute.

`codes.csv` (z. B. `gender.csv` etc.) beschreibt die Einträge von Kodierlisten.

`columns.csv` beschreibt schließlich wie die Tabelle aus `observations.csv` interpretiert werden soll, d. h. welche Komponenten und Kodierlisten für welche Spalte in der Observations-Tabelle verwendet wird.

Bereinigung der vorliegenden tabellarischen Daten I

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Land: Jugoslawien													
2	Bezeichnung: Amtliche Volkszählungen													
3	Zeitraum: 1948-1981													
4	Kompetenzbereich: nach ethnischer Zugehörigkeit; auf Landesebene													
5	Code: 59002170													
6	Quelle: Dubravka Velt: Stanovništvo Jugoslavije u posleratnom periodu. Beograd 1988, S. 139.													
7														
8														
9	Ethnische Struktur der Bevölkerung Jugoslawiens 1948-1981 (in absoluten Zahlen)						Ethnische Struktur der Bevölkerung Jugoslawiens 1948-1981 (in Prozent)							
10														
11		1948	1953	1961	1971	1981			1948	1953	1961	1971	1981	
12	Insgesamt	15.772.098	16.936.573	18.549.291	20.552.972	22.427.585	Insgesamt	100,0	100,0	100,0	100,0	100,0		
13	Montenegriner	425.703	466.093	513.832	508.843	579.043	Montenegriner	2,7	2,8	2,8	2,5	2,6		
14	Kroaten	3.784.353	3.975.550	4.293.809	4.526.782	4.428.043	Kroaten	24,0	23,5	23,1	22,1	19,8		
15	Makedonen	810.126	893.247	1.045.518	1.194.784	1.341.598	Makedonen	5,1	5,3	5,6	5,8	6,0		
16	Muslims	808.921	998.698	972.960	1.729.932	1.999.890	Muslims	5,1	5,9	5,2	8,4	8,9		
17	Slowenen	1.415.432	1.487.100	1.589.211	1.678.032	1.753.571	Slowenen	9,0	8,8	8,6	8,2	7,8		
18	Serben	6.547.117	7.065.923	7.806.152	8.143.246	8.140.507	Serben	41,5	41,7	42,1	39,7	36,3		
19	Albaner	750.431	754.245	914.733	1.309.523	1.730.878	Albaner	4,8	4,5	4,9	6,4	7,1		
20	Bulgaren	61.140	61.708	62.624	58.627	36.189	Bulgaren	0,4	0,4	0,3	0,3	0,2		
21	Tschechen	39.015	34.517	30.331	24.620	19.624	Tschechen	0,2	0,2	0,2	0,1	0,1		
22	Italiener	79.575	35.874	25.615	21.791	15.132	Italiener	0,5	0,2	0,1	0,1	0,1		
23	Magyaren	496.492	502.175	504.369	477.374	426.867	Magyaren	3,2	3,0	2,7	2,3	1,9		
24	Deutsche	55.337	60.538	20.015	12.785	8.712	Deutsche	0,4	0,4	0,1	0,1	0,0		
25	Zigeuner	72.736	84.713	31.674	78.485	168.197	Zigeuner	0,5	0,5	0,2	0,4	0,7		
26	Rumänen	64.095	60.364	60.862	58.570	54.955	Rumänen	0,4	0,4	0,3	0,3	0,2		
27	Russen	20.069	12.426	12.305	7.427	4.467	Russen	0,1	0,1	1,0	0,0	0,0		
28	Rusinen	37.140	37.353	38.619	24.640	23.286	Rusinen	0,2	0,2	0,2	0,1	0,1		
29	Slowaken	83.626	84.999	86.433	83.658	80.344	Slowaken	0,5	0,5	0,5	0,4	0,4		
30	Türken	97.954	259.535	182.964	127.920	101.293	Türken	0,6	1,5	1,0	0,6	0,5		
31	Ukrainer	-	-	-	13.972	12.813	Ukrainer	-	-	-	0,1	0,1		
32	Vlachen	102.953	36.728	9.463	21.990	32.071	Vlachen	0,7	0,2	0,1	0,1	0,1		
33	Andere	19.883	18.400	16.488	31.982	25.117	Andere	0,1	0,1	0,1	0,2	0,1		
34	Jugoslawen	-	-	317.124	273.077	1.219.024	Jugoslawen	-	-	1,7	1,3	5,4		
35	ohne nationale Deklaration	-	-	-	32.774	46.701	ohne nationale Deklaration	-	-	-	0,2	0,2		
36	Regionale Identität	-	-	-	15.002	25.730	Regionale Identität	-	-	-	0,1	0,1		
37	Unbekannt	-	6.389	14.192	67.138	153.545	Unbekannt	-	0,0	0,1	0,3	0,7		

Abbildung 15: Inhalt der alten Excel-Datei aus Ethnodoc: <https://lambda.ios-regensburg.de/dataset/ethnodoc-yu-1948-81-1>

Bereinigung der vorliegenden tabellarischen Daten II

Beobachtungen aufbereiten I

Gemäß den Best Practices (siehe Kalampokis, Zeginis und Tarabanis 2019, 4.8 Modeling single value dimensions, S. 64) sollten in den Beobachtungen auch Dimensionen aufgeführt werden, deren Wert bei allen Beobachtungen im Datensatz gleich ist – in unseren Daten betrifft das den räumlichen Bezug zum Land Jugoslawien:

```
Country ,Ethnicity ,Year ,Value
Jugoslawien ,Montenegriner ,1948 ,425703
Jugoslawien ,Kroaten ,1948 ,3784353
Jugoslawien ,Makedonen ,1948 ,810126
Jugoslawien ,Muslime ,1948 ,808921
Jugoslawien ,Slowenen ,1948 ,1415432
Jugoslawien ,Serben ,1948 ,6547117
Jugoslawien ,Albaner ,1948 ,750431
```

Beobachtungen aufbereiten II

- ① Zur Angabe des Landes wird der entsprechende GND-Normdateneintrag verwendet.
- ② Für die Angabe der ethnische Gruppen werden die Freitexteinträge durch die entsprechenden Notationen aus der Kodierliste für die ethnischen Gruppen ersetzt.
- ③ Für die Weiterverarbeitung mit table2qb wird die zusätzliche Spalte Measure Type benötigt:

```
Country ,Ethnicity ,Year ,Measure Type ,Value ,Unit
4028966-7 ,99 ,1948 ,Population ,15772098 ,Person
4028966-7 ,35 ,1948 ,Population ,425703 ,Person
4028966-7 ,27 ,1948 ,Population ,3784353 ,Person
4028966-7 ,33 ,1948 ,Population ,810126 ,Person
4028966-7 ,37 ,1948 ,Population ,808921 ,Person
4028966-7 ,55 ,1948 ,Population ,1415432 ,Person
4028966-7 ,49 ,1948 ,Population ,6547117 ,Person
4028966-7 ,05 ,1948 ,Population ,750431 ,Person
```

Komponenten der Beobachtungen definieren

Konfiguration der Datenstruktur (Data Structure Definition) für die Komponenten (Dimensionen, Messung und Attribut) in den QB-Daten (CSV-Datei `components.csv`):

```
Label,Description,Component Type,Codelist
Ethnicity,The ethnicity of the measured population,Dimension,http://ios-regensburg.de/et
Population,The population measured in persons,Measure,

table2qb exec components-pipeline \
--base-uri http://ios-regensburg.de/ethnodoc/ \
--input-csv components.csv \
--output-file components.ttl

<ethnodoc/def/dimension/ethnicity> a qb:DimensionProperty ;
  dcterms:description "The ethnicity of the measured population" ;
  qb:codeList <ethnodoc/def/concept-scheme/group> ;
  rdfs:label "Ethnicity" ;
  rdfs:range <ethnodoc/def/Ethnicity> ;
  skos:notation "ethnicity" .
```

Kodierliste für ethnische Gruppen aufbereiten

Aufbereitung der alten Ethnodoc-Kodierliste für die ethnischen Gruppen in der Datei groups.csv:

```
Label,Notation,Parent Notation
Ägypter,04,
Albaner,05,
Armenier,07,
Aromunen,09,
Ashkali,08,
Bosniaken,10,
Bulgaren,11,
Bunjewatzen,13,
Deutsche,15,
Gagausen,17,
```

```
table2qb exec codelist-pipeline \
--codelist-csv groups.csv \
--codelist-name "Ethnic Groups" \
--codelist-slug "group" \
--base-uri http://ios-regensburg.de/ethnodoc/ \
--output-file groups.ttl
```

Umgang mit aggregierten Werten in Tabellen

Aus "On modeling linked open statistical data":

Challenge 9.3: Should aggregate (e. g., "Total") values be included as dimension values? A common practice suggests that aggregate values should be included in a dimension (e. g., male, female and total). (Kalampokis, Zeginis und Tarabanis 2019, S. 65)

Label	Notation	Parent	Notation
Insgesamt	,99,		
Ägypter	,04,99		
Albaner	,05,99		
Armenier	,07,99		
Aromunen	,09,99		
Ashkali	,08,99		
Bosniaken	,10,99		
Bulgaren	,11,99		
Bunjewatzen	,13,99		
Deutsche	,15,99		
Gagausen	,17,99		

Zuordnung von Spalten zu Komponenten für QB-Datengenerierung konfigurieren

In der Datei `columns.csv` wird die Zuordnung der Spalten zu den zugehörigen Komponenten konfiguriert.

```
title ,name ,component_attachment ,property_template ,value_template ,datatype ,value_transformer  
Ethnicity ,ethnicity ,qb:dimension ,http://ios-regensburg.de/ethnodoc/def/dimension/ethnicity  
Year ,year ,qb:dimension ,http://purl.org/linked-data/sdmx/2009/dimension#refPeriod ,http://purl.org/linked-data/cube#refPeriod  
Country ,country ,qb:dimension ,http://purl.org/linked-data/sdmx/2009/dimension#refArea ,http://purl.org/linked-data/cube#refArea  
Measure Type ,measure_type ,qb:dimension ,http://purl.org/linked-data/cube#measureType ,http://purl.org/linked-data/cube#measureType  
Unit ,unit ,qb:attribute ,http://purl.org/linked-data/sdmx/2009/attribute#unitMeasure ,http://purl.org/linked-data/cube#unitMeasure  
Value ,value ,http://ios-regensburg.de/ethnodoc/def/measure/{measure_type} ,string ,  
Population ,population ,qb:measure ,http://ios-regensburg.de/ethnodoc/def/measure/population
```

```
table2qb exec cube-pipeline \  
--input-csv VZ_YU1948-1981_numbers_molten.csv \  
--dataset-name "VZ YU1948-1981 Dataset" \  
--dataset-slug "ethnodoc_yu_1948-81_1" \  
--column-config columns.csv \  
--base-uri http://ios-regensburg.de/ethnodoc/ \  
--output-file VZ_YU1948-1981_numbers.ttl
```

Modellierungsentscheidung Datentyp vs. Begriff

Zeitliche Dimension aus columns.csv:

title	Year
name	year
component_attachment	qb:dimension
property_template	http://purl.org/linked-data/sdmx/2009/dimension#refPeriod
value_template	http://reference.data.gov.uk/id/year/{year}
datatype	string
value_transformation	

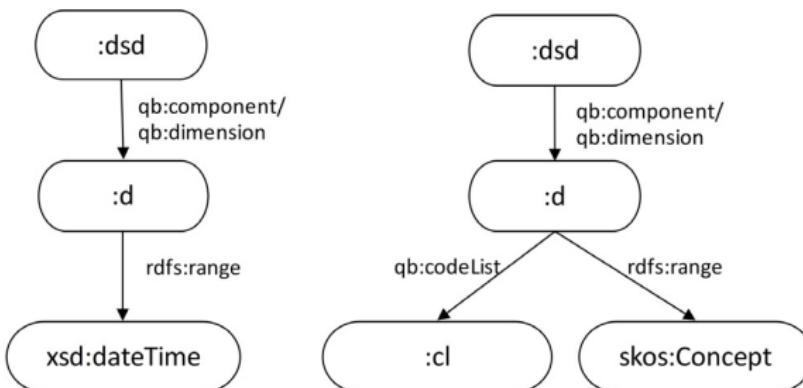


Abbildung 17: Ansätze zur Modellierung des zeitlichen Bezugs
(Dimension): mit Datumsangabe/Zeitintervall oder anhand eines Begriffs
aus einem Wissensorganisationssystem (aus Kalampokis, Zeginis und
Tarabanis 2019, S. 63)

table2qb components-pipeline, codelist-pipeline und cube-pipeline zur Datenaufbereitung und QB-Erstellung

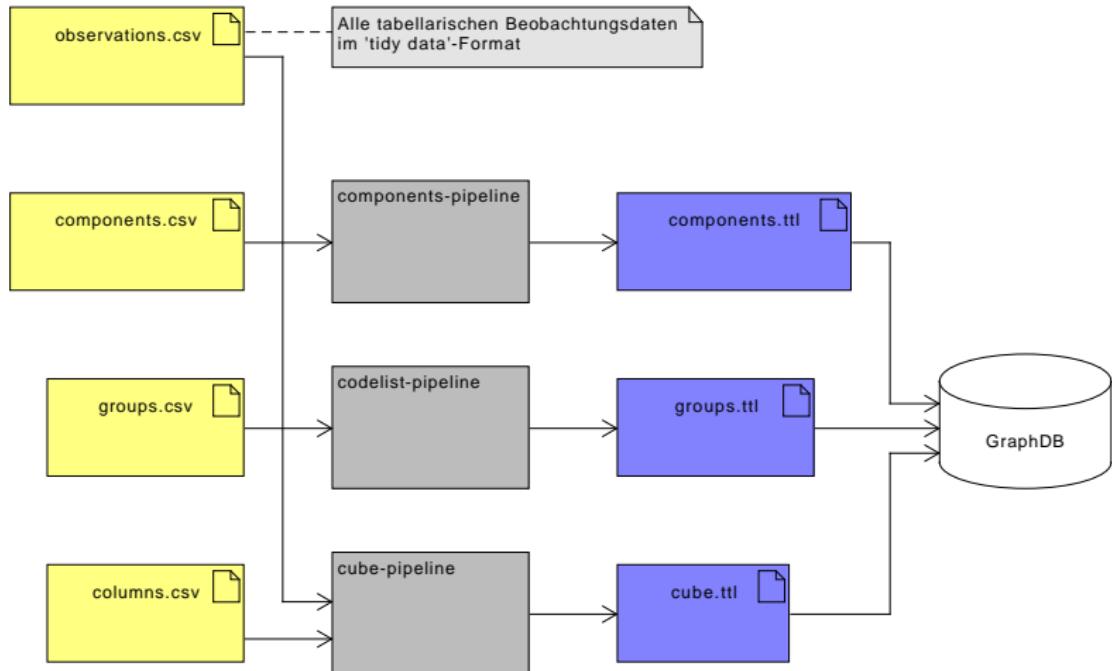


Abbildung 18: Pipelines zur Erstellung von QB-Daten mit table2qb

QB-Datensatz mit Definition der Datenstruktur

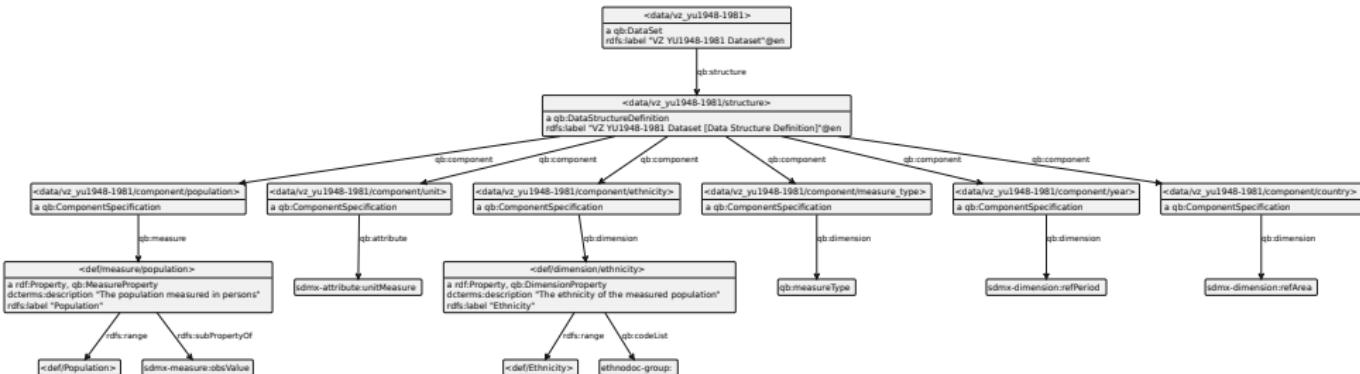


Abbildung 19: rdfpuml-Diagramm¹⁶ der von table2qb generierten RDF-Daten zur Repräsentation des QB-Datensatzes und dessen Datenstrukturdefinition (vereinfachte Darstellung)

¹⁶Zur Überprüfung der automatisch generierten RDF-Daten eignet sich das Visualisierungswerkzeug rdfpuml von Alexiev (2016) sehr gut:
<https://github.com/VladimirAlexiev/rdf2rml>

Manuelle Anpassungen der Datenstrukturdefinition I

Neue Dimension (als Unterklasse von sdmx-dimension:refPeriod) für die Angabe von Jahren als Zeitintervall (Klasse aus der Time Intervals Ontology) (siehe Kalampokis, Zeginis und Tarabanis 2019, Ansatz 7.1b, S. 63) definieren:

```
<def/dimension/year> a qb:DimensionProperty ;
  rdfs:subPropertyOf sdmx-dimension:refPeriod ;
  rdfs:range <http://reference.data.gov.uk/def/intervals/Year> .
```

Komponente mit der neuen Dimension zur Datenstrukturdefinition hinzufügen:

```
<data/vz_yu1948-1981/component/year>
  a qb:ComponentSpecification ;
  qb:dimension <def/dimension/year> .
```

Manuelle Anpassungen der Datenstrukturdefinition II

Neue Dimension (als Unterklasse von sdmx-dimension:refPeriod) zur Verwendung von Normdateneinträgen (Orte aus der Gemeinsamen Normdatei (GND)) für den räumlichen Bezug definieren (vgl. Kalampokis, Zeginis und Tarabanis 2019, Ansatz 7.2, S. 63f.):

```
<def/dimension/country> a qb:DimensionProperty ;  
    rdfs:subPropertyOf sdmx-dimension:refArea ;  
    rdfs:range gndo:PlaceOrGeographicName .
```

```
<data/vz_yu1948-1981/component/country>  
    a qb:ComponentSpecification ;  
    qb:dimension <def/dimension/country> .
```

Ergänzungen in QB-Komponenten und Datentrukturdefinition



Abbildung 20: rdfpuml-Diagramm mit manuell angepassten Komponenten (year, country, ethnicity und population) und ergänzten Dimensionen (siehe insbesondere die `rdfs:range`-Angaben)

Ergebnis der Datenaufbereitung im RDF-Datenmodell

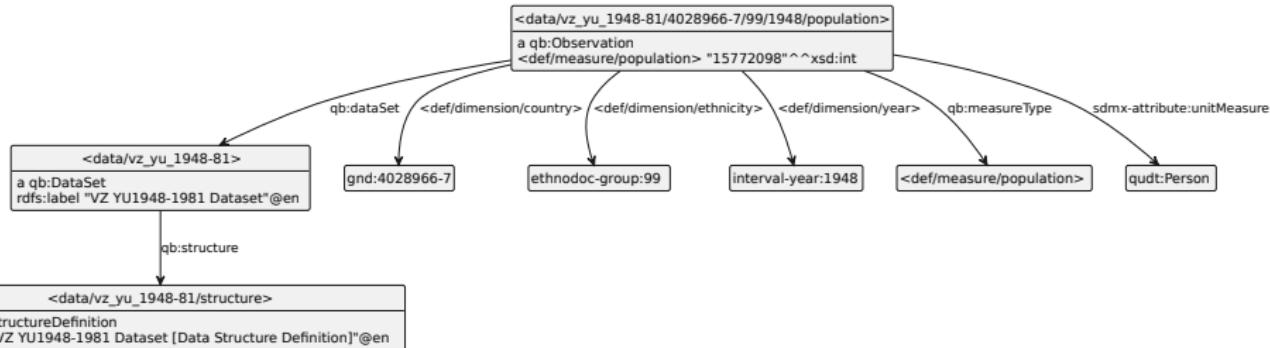


Abbildung 21: rdfpuml-Diagramm mit einer fertigen Beobachtung (qb:Observation) im QB-Datensatz (qb:DataSet)

Verfeinerungen mit Tarql I

Tarql ist ein Werkzeug zur Aufbereitung von Daten aus CSV-Dateien mit der Abfragesprache SPARQL:

<https://github.com/tarql/tarql>

CSV-Datei `groups_enriched.csv` (Auszug) für die Ethnodoc-Kodierliste aus dem `table2qb`-Beispiel mit zusätzlicher Spalte zur Angabe von beudeutungsgleichen Einträgen in Wikidata, sowie synonyme Bezeichner und Verweise auf verwandte Begriffe:

Insgesamt ,99,,Q41710,,
Ägypter ,04,99,Q805039,,
Albaner ,05,99,Q179248,,
Bunjewatzen ,13,99,Q591964,,
Deutsche ,15,99,Q42884,,
Gagausen ,17,99,Q180361,,
Goranci ,18,99,Q851126,Gorani ,
Ruthenen ,47,99,Q690869,,73
Sonstige ethnische Gruppen ,79,99,,
Ukrainer ,73,99,Q44806,,47
Zigeuner ,77,99,Q8060,Roma ,

Verfeinerungen mit Tarql II

Mit SPARQL CONSTRUCT-Abfragen kann man mit Tarql QB- oder SKOS-Daten aus entsprechenden CSV-Dateien generieren:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

CONSTRUCT {
    ?concept a skos:Concept ;
        rdfs:label ?Label ;
        skos:inScheme <http://ios-regensburg.de/ethnodoc/def/concept-scheme/group> ;
        skos:notation ?Notation ;
        skos:prefLabel ?Label ;
        skos:altLabel ?AltLabel ;
        skos:broader ?broader ;
        skos:related ?related ;
        skos:exactMatch ?wd .
}
FROM <groups_enriched.csv> WHERE {
    BIND (URI(CONCAT('http://ios-regensburg.de/ethnodoc/def/concept/group/', ?Notation))
        AS ?concept)
    BIND (URI(CONCAT('http://ios-regensburg.de/ethnodoc/def/concept/group/', ?ParentNotati
        AS ?broader)
    BIND (URI(CONCAT('http://ios-regensburg.de/ethnodoc/def/concept/group/', ?Related))
        AS ?related)
    BIND (URI(CONCAT('http://www.wikidata.org/entity/', ?Wikidata)) AS ?wd)
}

tarql construct-codelist.sparql \
    groups_enriched.csv > groups_enriched.ttl
```

Verfeinerungen mit Tarql III

Ergebnis SKOS-Kodierliste in RDF-Turtle (Auszug):

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://ios-regensburg.de/ethnodoc/def/concept/group/99>
    rdf:type      skos:Concept ;
    rdfs:label    "Insgesamt" ;
    skos:inScheme <http://ios-regensburg.de/ethnodoc/def/concept-scheme/group> ;
    skos:notation "99" ;
    skos:prefLabel "Insgesamt" ;
    skos:exactMatch "Q41710" .

<http://ios-regensburg.de/ethnodoc/def/concept/group/04>
    rdf:type      skos:Concept ;
    rdfs:label    "Ägypter" ;
    skos:inScheme <http://ios-regensburg.de/ethnodoc/def/concept-scheme/group> ;
    skos:notation "04" ;
    skos:prefLabel "Ägypter" ;
    skos:exactMatch "Q805039" .

<http://ios-regensburg.de/ethnodoc/def/concept/group/05>
    rdf:type      skos:Concept ;
    rdfs:label    "Albaner" ;
    skos:inScheme <http://ios-regensburg.de/ethnodoc/def/concept-scheme/group> ;
    skos:notation "05" ;
    skos:prefLabel "Albaner" ;
    skos:exactMatch "Q179248" .
```

Fertiges SKOS-Klassifikationssystem für ethnische Gruppen

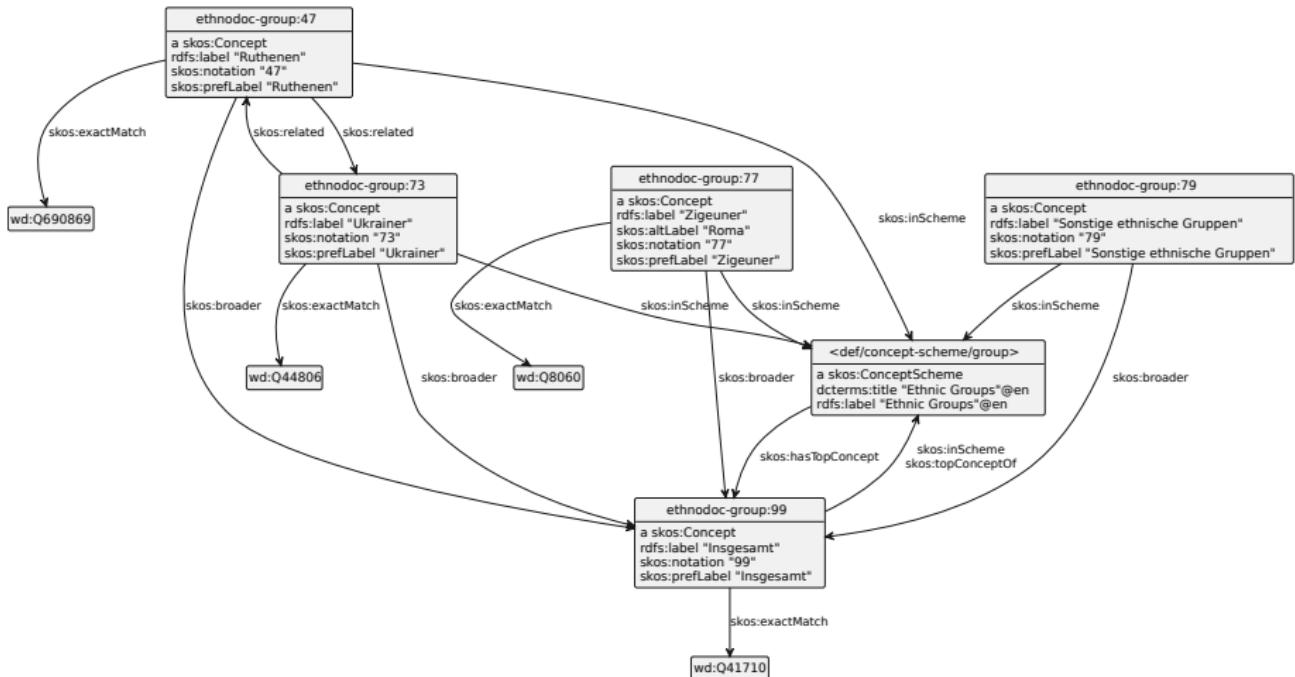


Abbildung 22: rdfpuml-Diagramm mit den Begriffen (Auswahl) zur Kodierung der ethnischen Gruppen

Geodaten zur räumlichen Verbreitung ethnischer Gruppen

- Wie am Beispiel der historischen Zensusdaten gezeigt, werden mit dem QB-Datenmodell typischerweise statistische Daten, d. h. numerische Werte als Messungen modelliert.
- QB ist jedoch nicht auf statistische Daten beschränkt, sondern eignet sich allgemein für mehrdimensionale Daten.
- Exemplarisch werden wir den Datensatz GREG (Geo-referencing of Ethnic Groups)¹⁷ mit QB und der Erweiterung QB4ST (RDF Data Cube extensions for spatio-temporal components)¹⁸ aufbereiten.
- Die geographische Verbreitung einer ethnischen Gruppe kann im QB-Datenmodell als Messung aufgefasst und modelliert werden.
- Die Geodaten (Polygone) in den QB-Messungen repräsentieren also das beobachtete Phänomen (das Siedlungsgebiet einer ethnischen Gruppe).

¹⁷<https://icr.ethz.ch/data/greg/>

¹⁸<https://www.w3.org/TR/qb4st/>

Ziel GREG-LD

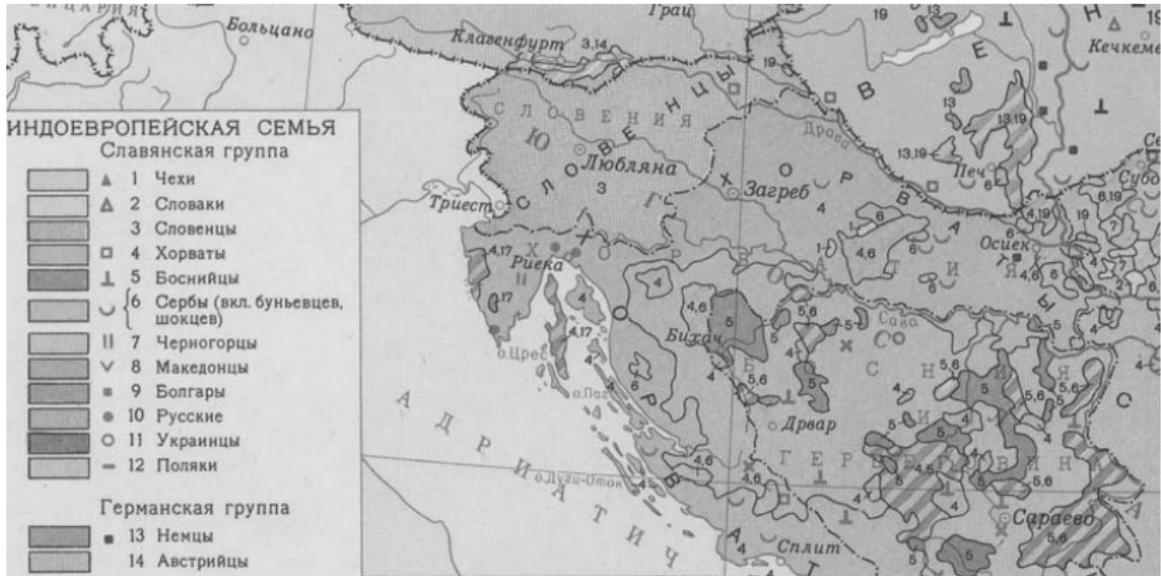


Abbildung 23: Polygone können für Analyse und Visualisierung aus dem integrierten Linked Data-Bestand abgefragt werden (Scan aus *Atlas Narodov Mira* von (Weidmann, Rød und Cederman 2010))

GREG Shapefile

Tabellarische Daten im GREG Shapefile GREG.shp:

GREG :: Features total: 8969, filtered: 8969, selected: 0																
	FIPS_CTRY	GROUP1	GROUP2	GROUP3	G1ID	G2ID	G3ID	G1SHORTNAME	G2SHORTNAME	G3SHORTNAME	G1LONGNAME	G2LONGNAME	G3LONGNAME	FeatureID	AREA	COW
1	AA		12	0	0	312	0	0	Curaao Isla...		Curaao Isla...			0	200779464...	0
2	AC		16	0	0	354	0	0	English-spe...		English-spe...			1	539856985...	58
3	AF		33	53	0	117	202	0	Baloch	Brahui	Baloch (Bal...	Brahui		2	118978120...	700
4	AF		24	34	0	898	12	0	Persians	Afghans	Persians	Afghans (P...		3	165361024...	700
5	AF		34	41	0	12	1051	0	Afghans	Tajiks	Afghans (P...	Tajiks (Tadz...		4	325101081...	700
6	AF		41	0	0	1051	0	0	Tajiks		Tajiks (Tadz...			5	633642253...	700
7	AF		33	0	0	117	0	0	Baloch		Baloch (Bal...			6	166250366...	700
8	AF		20	0	0	1149	0	0	Turkmens		Turkmens (...)			7	617558569...	700
9	AF		33	0	0	117	0	0	Baloch		Baloch (Bal...			8	634629108...	700
10	AF		41	0	0	1051	0	0	Tajiks		Tajiks (Tadz...			9	196693749...	700
11	AF		40	0	0	1084	0	0	Teymurs		Teymurs			10	490436898...	700
12	AF		34	0	0	12	0	0	Afghans		Afghans (P...			11	112719747...	700
13	AF		40	0	0	1084	0	0	Teymurs		Teymurs			12	401146984...	700
14	AF		36	0	0	493	0	0	Jamshidis		Jamshidis			13	601267828...	700
15	AF		34	41	0	12	1051	0	Afghans	Tajiks	Afghans (P...	Tajiks (Tadz...		14	204029638...	700
16	AF		41	0	0	1051	0	0	Tajiks		Tajiks (Tadz...			15	969151201...	700

Abbildung 24: Shapefile aus dem GREG-Datensatz in QGIS

Linked Data-Paradigma als Ausweg aus den Einschränkungen gängiger Datenmodelle für Geodaten¹⁹:

- metadata and schemata are separate entities
- the semantics of terms remains implicit or hard to share and reason with
- data are seen as provider-independent truths, though they often contradict
- global, unique identifiers are hard to obtain and not encouraged
- valuable data sets remain isolated and hard to integrate

¹⁹Auszug aus "Linked Data – A Paradigm Shift for Geographic Information Science" von Kuhn, Kauppinen und Janowicz (2014)

Aufbereitung des GREG-Datensatzes mit Python pandas & RDFLib

Auszug aus dem Python-Skript im Jupyter-Notebook:

```
from rdflib import Graph, Literal, BNode, Namespace, RDF, URIRef, XSD
from rdflib.namespace import DC, DCTERMS, SKOS, RDF, RDFS, OWL
import geopandas as gpd
import pandas as pd
import numpy as np

g = Graph()

greg = gpd.read_file("GREG.shp")

ds_uri = "http://ios-regensburg.de/greg/data/greg-ld/"
ds = URIRef(ds_uri)

for index, row in greg.iterrows():
    greg_id = row['FeatureID']
    obs_uri = "http://ios-regensburg.de/greg-ld/data/greg-ld/%s" % greg_id
    obs = URIRef(obs_uri)

    g.add( (obs, RDF.type, QB.Observation) )
    g.add( (obs, QB.dataSet, ds) )

    g.add( (obs, GREG_DIM.group,
            URIRef('http://ios-regensburg.de/greg/def/concept/group/%s' % row['GROUP1'])) )
    g.add( (obs, GREG_MEA.geom, Literal(row['geometry']), datatype=GEO.wktLiteral) )

g.serialize(destination='greg_qb.ttl', format='turtle')
```

GREG-LD in QB

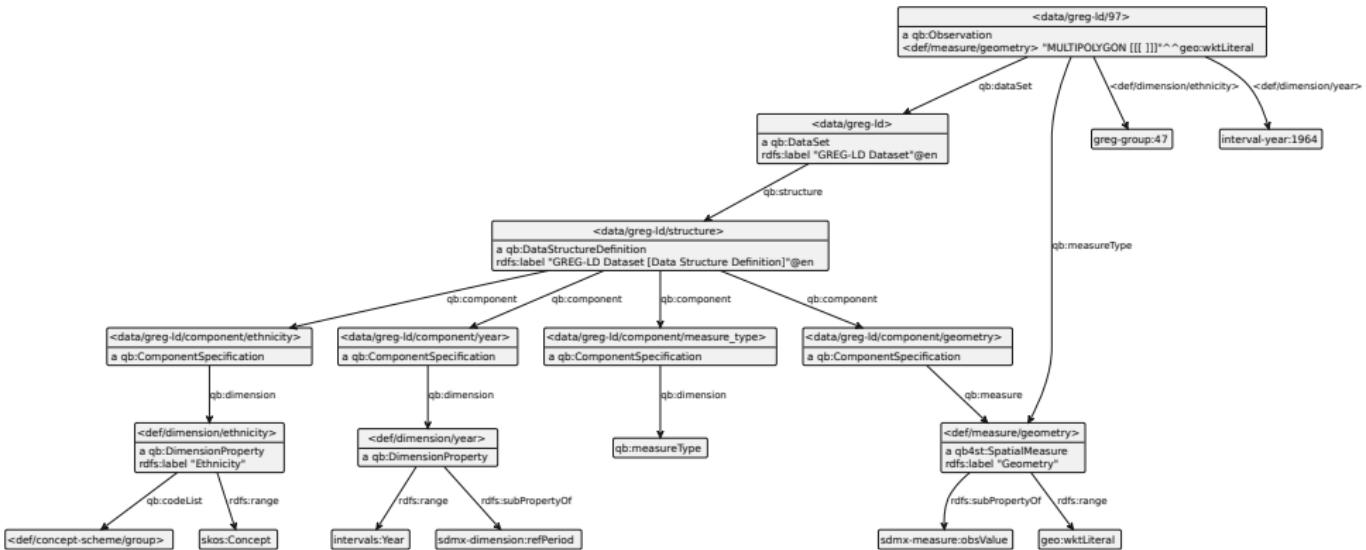


Abbildung 25: rdflpuml-Diagramm des GREG-Datensatzes als QB-Daten mit -Datenstruktur

GREG-Kodierliste für ethnische Gruppen in SKOS

Auszug aus der groups.csv für den GREG-Datensatz:

```
GID , SHORTNAME , LONGNAME ,
398 , Germans , Germans ,
401 , Gipsies , Gipsies (Romanies) ,
1255 , Jews , Jews , General group for Jews when there is no further specification given .
1256 , Arabs , Arabs , General group for Arabs when there is no further specification given .
1257 , Chini , Chini , Translation missing in the Atlas .
1258 , Yugoslavs , Yugoslavs , Combined group for all Yugoslavs . Used for the USSR .
1251 , Others and Unknown , Others and Unknown - dummy group ,
```

Vorbereitung der CSV-Daten für die ‚SKOSifikation‘:

```
notation , prefLabel , altLabel , scopeNote , wd
398 , Germans , Germans , ,
401 , Gipsies , Gipsies (Romanies) , ,
1255 , Jews , Jews , General group for Jews when there is no further specification given . ,
1256 , Arabs , Arabs , General group for Arabs when there is no further specification given . ,
1257 , Chini , Chini , Translation missing in the Atlas . ,
1258 , Yugoslavs , Yugoslavs , Combined group for all Yugoslavs . Used for the USSR . ,
1251 , Others and Unknown , Others and Unknown - dummy group , ,
```

Statistische Daten und Geodaten testen und weiter anreichern in Ontotext GraphDB



GraphDB Free Edition²⁰

GraphDB Workbench Die Workbench ist das Web-basierte Administrationswerkzeug von GraphDB – zum Importieren, Explorieren, Abfragen und Exportieren von RDF-Daten.

OntoRefine OntoRefine²¹ basiert auf OpenRefine²² und ist Teil der GraphDB Workbench) zur Datenaufbereitung und Datenintegration.

²⁰<https://www.ontotext.com/products/graphdb/graphdb-free>

²¹<https://graphdb.ontotext.com/documentation/free/loading-data-using-ontorefine.html>

²²<https://openrefine.org/>

Anreicherung und Aufbereitung in OntoRefine I

GraphDB FREE

Import

RDF

Tabular (OntoRefine)

Explore

SPARQL

Monitor

Setup

Help

OntoRefine ⓘ

groups_selection.csv ⚪️ 🔍

Open... Export Help

Facet / Filter Undo / Redo 0 / 0

7 rows

Show as: rows records Show: 5 10 25 50 rows first previous 1 - 7 next last

All	notation	prefLabel	altLabel	scopeNote	wd
☆	1.	398	Germans	Germans	
☆	2.	401	Gipsies	Gipsies (Romanies)	
☆	3.	1255	Jews	Jews	General group for Jews when there is no further specification given.
☆	4.	1256	Arabs	Arabs	General group for Arabs when there is no further specification given.
☆	5.	1257	Chini	Chini	Translation missing in the Atlas.
☆	6.	1258	Yugoslavs	Yugoslavs	Combined group for all Yugoslavs. Used for the USSR.
☆	7.	1251	Others and Unknown	Others and Unknown - dummy group	

Anreicherung und Aufbereitung in OntoRefine II

GraphDB FREE

Import

RDF

Tabular (OntoRefine)

Explore

SPARQL

Monitor

Setup

Help

OntoRefine ⓘ

groups_selection.csv 7 rows

Open... Export Help

Facet / Filter Undo / Redo ↺ ↻

Reconcile column "prefLabel"

Reconcile each cell to an entity of one of these types:

Name	Count	Score
ethnic group Q41710	4	3.33
painting Q3305213	3	1.17
scholarly article Q13442814	2	0.83
language Q34770	1	1
ethnic group Q2531956	1	1
ethnoreligious group Q11197007	1	1
symptom Q169872	1	1
...

Also use relevant details from other columns:

Column	Include? As Property
notation	<input type="checkbox"/>
altLabel	<input type="checkbox"/>
scopeNote	<input type="checkbox"/>
wd	<input type="checkbox"/>

Reconcile against type:

Reconcile against no particular type

Auto-match candidates with high confidence

Maximum number of candidates to return ↕

Add Standard Service...

Start Reconciling Cancel

Anreicherung und Aufbereitung in OntoRefine III

The screenshot shows the OntoRefine interface with the following components:

- Left Sidebar:** Contains links for Import, Explore, SPARQL, Monitor, Setup, Repositories, Users and Access, My Settings, Connectors, Namespaces, Autocomplete, RDF Rank, JDBC, and SPARQL Templates.
- Header:** Shows "OntoRefine" with a help icon, a search bar with "greg-id", and navigation buttons for Open..., Export, and Help.
- Middle Left:** A teal box titled "Using facets and filters" provides instructions on how to use facets and filters to select subsets of data. It includes a link to "Watch these screencasts".
- Middle Right:** A data table titled "groups_selection.csv" with 7 rows. The table columns are All, notation, prefLabel, altLabel, scopeNote, and wd. The data is as follows:

All	notation	prefLabel	altLabel	scopeNote	wd
1. 398	Germans	Germans			Q42884
2. 401	Gipsies	Gipsies (Romanies)			Q8060
3. 1255	Jews	Jews		General group for Jews when there is no further specification given.	Q7325
4. 1256	Arabs	Arabs		General group for Arabs when there is no further specification given.	Q35323
5. 1257	Chini	Chini		Translation missing in the Atlas.	Q24040442
6. 1258	Yugoslavs	Yugoslavs		Combined group for all Yugoslavs. Used for the USSR.	Q236807
7. 1251	Others and Unknown	Others and Unknown - dummy group			

GraphDB FREE

OntoRefine ⓘ

Import

Explore

SPARQL

Monitor

Setup

Repositories

Users and Access

My Settings

Connectors

Namespaces

Autocomplete

RDF Rank

JDBC

SPARQL Templates

Help

Configuration Preview Mapping has unsaved changes Both Save Download JSON Upload JSON RDF SPARQL

notation prefLabel altLabel scopeNote wd

http://ios-regensburg.de/greg/def/dimension/group

Use the current repository prefixes or add new using the Turtle or SPARQL syntax, i.e PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

greg-group ✎ rdfs ✎ skos ✎ wd ✎

greg-group: notation	<IRI> 	a	<IRI> 	skos: Concept <IRI>
greg-group: notation	<IRI> 	skos: prefLabel <IRI> 	prefLabel en "Literal" @Language	
greg-group: notation	<IRI> 	skos: altLabel <IRI> 	altLabel en "Literal" @Language	
greg-group: notation	<IRI> 	skos: scopeNote <IRI> 	scopeNote en "Literal" @Language	
greg-group: notation	<IRI> 	skos: exactMatch	<IRI> 	wd: wd <IRI>
greg-group: notation	<IRI> 	skos: notation	<IRI> 	notation "Literal"

Abgleich der beiden SKOS-Kodierlisten über Normdaten

The screenshot shows the GraphDB interface with the following details:

- Left Sidebar:** Includes "Import", "Explore", "SPARQL" (selected), "Monitor", "Setup", and "Help".
- Top Bar:** Search bar with "greg-ld", tabs for "Editor only", "Editor and results" (selected), "Results only", and a refresh icon.
- SPARQL Editor:** A code editor containing the following query:

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
CONSTRUCT {
    ?group_ethnodoc skos:exactMatch ?group_greg .
    ?group_greg skos:exactMatch ?group_ethnodoc .
}
WHERE {
    ?group_ethnodoc a skos:Concept ;
        skos:inScheme <http://ios-regensburg.de/ethnodoc/def/concept-scheme/group> ;
        skos:exactMatch ?group_wikidata .
    ?group_greg a skos:Concept ;
        skos:inScheme <http://ios-regensburg.de/greg/def/concept-scheme/group> ;
        skos:exactMatch ?group_wikidata .
}
FILTER (strstarts(str(?group_wikidata), "http://www.wikidata.org/entity/"))
```
- Right Sidebar:** Icons for saving, opening, linking, sharing, and more.
- Buttons:** "Run" (red), "Keyboard shortcuts", "Table", "Raw Response", "Pivot Table", "Google Chart", "Download as", and "Visual".
- Result Area:** A table showing the results of the query, with 8 rows displayed from 1 to 8. The columns are "subject", "predicate", and "object".

	subject	predicate	object
1	ethnodoc-group:15	skos:exactMatch	greg-group:398
2	greg-group:398	skos:exactMatch	ethnodoc-group:15
3	ethnodoc-group:77	skos:exactMatch	greg-group:401
4	greg-group:401	skos:exactMatch	ethnodoc-group:77
5	ethnodoc-group:23	skos:exactMatch	greg-group:1255
6	greg-group:1255	skos:exactMatch	ethnodoc-group:23

Statistische Daten und Geodaten bereitstellen mit Apache Jena Fuseki



GeoSPARQL Fuseki SPARQL-Server²³

Inbetriebnahme GeoSPARQL Fuseki

Aktuelle Version von GeoSPARQL Fuseki herunterladen:

<https://repo1.maven.org/maven2/org/apache/jena/jena-fuseki-geosparql/>

GeoSPARQL Fuseki mit QB-Geodaten starten:

```
java -jar jena-fuseki-geosparql-4.5.0.jar \
      -dataset greg-ld -rf greg_qb.ttl -i
```

SPARQL-Endpoint nutzen:

<http://localhost:3030/greg-ld>

²³<https://jena.apache.org/documentation/geosparql/geosparql-fuseki.html>

Beschreibung der Daten mit Metadaten für die Publikation

OstData-Leitfaden „Forschungsdaten publikationsfähig aufbereiten“: <https://doi.org/10.5281/zenodo.6517433>

The screenshot shows the Zenodo interface with the following details:

- Title:** OstData Leitfaden: Forschungsdaten publikationsfähig aufbereiten
- Views:** 54
- Downloads:** 47
- Indexed in:** OpenAIRE
- Publication date:** May 4, 2022
- DOI:** [10.5281/zenodo.6517433](https://doi.org/10.5281/zenodo.6517433)
- Keywords:** Forschungsdaten, Forschungsdatenmanagement, Publikation, Datenaufbereitung, Deterioration
- Subject(s):** Forschungsdaten
- Published in:** Materialien zum Forschungsdatenmanagement in der Ost-, Ostmittel- und Südosteuropaforschung.
- Related identifiers:**
 - Cites: [10.5281/zenodo.5566592](https://doi.org/10.5281/zenodo.5566592) (Data management plan)
 - [10.5281/zenodo.6365341](https://doi.org/10.5281/zenodo.6365341) (Journal article)
 - [10.5281/zenodo.6420798](https://doi.org/10.5281/zenodo.6420798) (Journal article)
- Communities:** OstData – Research Data Service for Central, Eastern and South-Eastern Europe

Mit diesem Leitfaden stellt OstData ein Instrumentarium an Fragen und Arbeitsschritten bereit, um Forschungsdaten für die Publikation aufzubereiten. Ziele einer qualitativ hochwertigen Aufbereitung von Forschungsdaten sind in den FAIR-Prinzipien zum Umgang mit Forschungsdaten beschrieben, denen sich OstData verpflichtet. Nach den FAIR-Prinzipien sollen Forschungsdaten

- Findable („auffindbar“), d. h. über Repositorien und mit persistenter Adresse auf-

SPARQL-Abfragen

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ethnодoc-dimension: <http://ios-regensburg.de/ethnодoc/def/
PREFIX ethnодoc-measure: <http://ios-regensburg.de/ethnодoc/def/me
PREFIX ethnодoc-group: <http://ios-regensburg.de/ethnодoc/def/conc

SELECT ?year ?ethnicity_code ?ethnicity_label ?population
WHERE {
    ?obs a qb:Observation ;
        ethnодoc-dimension:year ?year ;
        ethnодoc-dimension:ethnicity ?ethnicity ;
        ethnодoc-measure:population ?population .
    FILTER (?population < 1000)
}
ORDER BY ?year desc(?population)
```

Verwendung der Daten für Visualisierungen

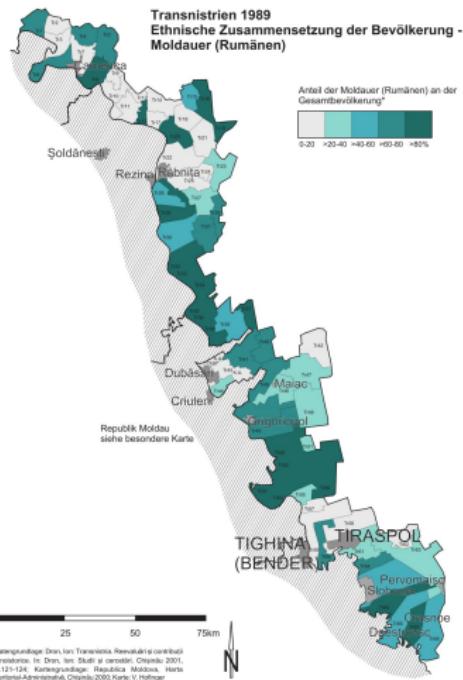


Abbildung 26: Projektergebnis ethnographische Karte von Transnistrien basierend auf aggregierten Daten aus Ethnodoc:
<http://geoportal.ios-regensburg.de/catalog/BV044865267>

Verschneidung von Kartenmaterial und Geodaten



Abschlussdiskussion / Q & A

Beispielprojekt CEDAR (Census Data Research) I



Data

The dataset consists of about 2000 "raw" data cubes integrated into 1 "harmonized" data cube. All the data is accessible via a SPARQL endpoint.

[Take a look »](#)

Statistics

Integrating all the data involves manual mappings to standardised values. As we continue working on it, the data set gets regenerated and evaluated.

[Take a look »](#)

Information

Do you want to know more about how we produce the data? Or do you want to get in touch with us? We try to answer all the questions in one page.

[Take a look »](#)

Beispielprojekt CEDAR (Census Data Research) II

Das niederländische Projekt CEDAR (Census Data Research) verwendet eine Reihe von Kodierschemas und Klassifikationssystemen, wie etwa HISCO (Historical International Standard Classification of Occupations), bei der Aufbereitung tabellarischer historischer Zensudaten als Linked Data (Meroño-Peña et al. 2017):

```
cedar:example-observation a qb:Observation ;
    maritalstatus:maritalStatus maritalstatus:single ;
    sdmx-dimension:sex sdmx-code:sex-M ;
    sdmx-dimension:age "12"^^xml:integer ;
    cedar:yearOfBirth "1878"^^xml:integer ;
    sdmx-dimension:refArea gg:Amsterdam ;
    cedar:occupation hisco:88030 ;
    cedar:occupationPosition cedar:job-D ;
    cedar:population "128"^^xml:integer .
```

Beispielprojekt CEDAR (Census Data Research) III

CEDAR

Home

About

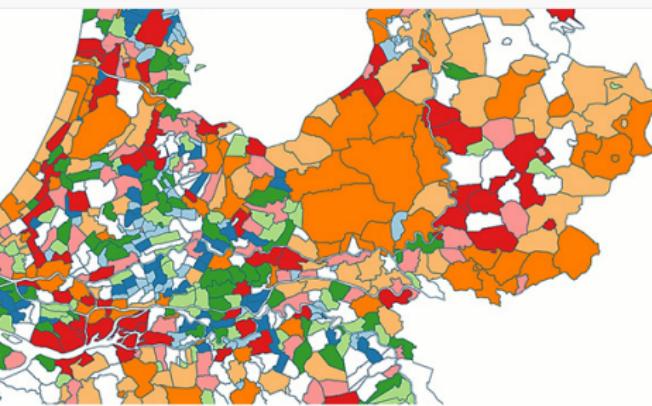
Harmonized Data

Publications

RDF ▾

GIS

Contact us



Download data

Download a dump of the harmonized data in one big table. Different formats are provided to suite different needs..

[View details »](#)



Variable Search

In this section we guide the users to make the correct combinations in order to query for specific variables across the years..

[View details »](#)



SPARQL-Endpoint

Query the data directly from our SPARQL-endpoint. This option is for expert users with clear insights on the data and its structure..

[View details »](#)

Implikationen für Forschung in der Digitalen Geschichtswissenschaft

- Allgemein: Vorteile eines einheitlichen Datenmodells für statistische Daten und darauf aufbauende Analysen und Visualisierungen (Zapilko und Mathiak 2011; Meroño-Peñuela und Ashkpour 2016).
- Digitale Quellenkritik wird durch den Zugriff auf integriertes Quellenmaterial bzw. einheitlich modellierten Datenbasis unterstützt und gefördert.²⁴
- Aus Sicht des Forschungsdatenmanagements: Der Linked Data-Ansatz erlaubt sozusagen einen nahtlosen Übergang von den Metadaten in die Daten selbst – und dort von den Makrodaten zu den Mikrodaten ...

²⁴Die aufbereiteten Daten könnten z. B. eine Grundlage für folgende Untersuchung bilden: "What Does the Atlas Narodov Mira Measure?" (Bridgman 2008) über die Definition von 'ethnicity'

Vielen Dank für Ihre Aufmerksamkeit

Kontakt

Web	www.ostdata.de
Email	ostdata@bsb-muenchen.de
Twitter	@ostdata

Literatur |

-  Alexiev, Vladimir (Nov. 2016). "RDF by Example: rdfpuml for True RDF Diagrams, rdf2rml for R2RML Generation". In: *Semantic Web in Libraries 2016 (SWIB 16)*. Bonn, Germany.
-  Baker, Thomas et al. (2013). "Key choices in the design of Simple Knowledge Organization System (SKOS)". In: *Journal of Web Semantics* 20, S. 35–49. DOI:
<https://doi.org/10.1016/j.websem.2013.05.001>.
-  Boer, Victor de, Albert Meroño-Peñuela und Niels Ockeloen (2016). "Linked Data for Digital History: Lessons Learned from Three Case Studies". In: *Historiografía digital. Proyectos para almacenar y construir la Historia*. Hrsg. von Mirella Romero Recio und Jesús Colmenero Ruiz. Anejos de la Revista de Historiografía. Madrid.
-  Bridgman, Benjamin (2008). "What Does the Atlas Narodov Mira Measure?" In: *Economics Bulletin* 10.6, S. 1–8. URL: <https://ideas.repec.org/a/eb1/ecbull/eb-08j10005.html>.

Literatur II

-  Corcho, Oscar, María Poveda-Villalón und Asunción Gómez-Pérez (2015). "Ontology Engineering in the Era of Linked Data". In: *Bulletin of the Association for Information Science and Technology* 41.4, S. 13–17. DOI: 10.1002/bult.2015.1720410407. URL: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/bult.2015.1720410407>.
-  Harvey, Charles und Jon Press (1996). "Source-Oriented Database Systems". In: *Databases in Historical Research: Theory, Methods and Applications*. Palgrave Macmillan. Kap. 7, S. 190–217.
-  Heery, Rachel und Manjula Patel (2000). "Application Profiles: Mixing and Matching Metadata Schemas". In: *Ariadne* 25. URL: <http://www.ariadne.ac.uk/issue/25/app-profiles/>.

Literatur III

-  Kalampokis, Evangelos, Dimitris Zeginis und Konstantinos Tarabanis (2019). "On modeling linked open statistical data". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 55, S. 56–68.
-  Kuhn, Werner, Tomi Kauppinen und Krzysztof Janowicz (2014). "Linked Data – A Paradigm Shift for Geographic Information Science". In: *Geographic Information Science*. Hrsg. von Matt Duckham et al. Bd. 8728. Lecture Notes in Computer Science. Springer International Publishing, S. 173–186.
-  Meroño-Peña, Albert und Ashkan Ashkpour (2016). "Historical Quantitative Reasoning on the Web". In: *European Social Science History Conference (ESSHC)*.
-  Meroño-Peña, Albert et al. (2017). "CEDAR: The Dutch Historical Censuses as Linked Open Data". In: *Semantic Web* 8.2, S. 297–310. DOI: 10.3233/SW-160233.

Literatur IV

-  Nelson, Theodor H. (1965). "Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate". In: *Proceedings of the 1965 20th National Conference*. ACM '65. ACM, S. 84–100. DOI: [10.1145/800197.806036](https://doi.org/10.1145/800197.806036).
-  Tambouris, Efthimios, Evangelos Kalampokis und Konstantinos Tarabanis (2015). "Processing Linked Open Data Cubes". In: *Electronic Government*. Hrsg. von Efthimios Tambouris et al. Cham: Springer International Publishing, S. 130–143.
-  Weidmann, Nils B., Jan Ketil Rød und Lars-Erik Cederman (2010). "Representing Ethnic Groups in Space: A New Dataset". In: *Journal of Peace Research* 47.4, S. 491–499.

Literatur V

- ❑ Wickham, Hadley (2014). "Tidy Data". In: *Journal of Statistical Software* 59.10, S. 1–23. DOI: 10.18637/jss.v059.i10. URL: <https://www.jstatsoft.org/index.php/jss/article/view/v059i10>.
- ❑ Zapilko, Benjamin und Brigitte Mathiak (2011). "Performing Statistical Methods on Linked Data". In: *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*. DCMI'11. Dublin Core Metadata Initiative, S. 116–125.

FDM-Workflow zur Integration von Forschungsdaten

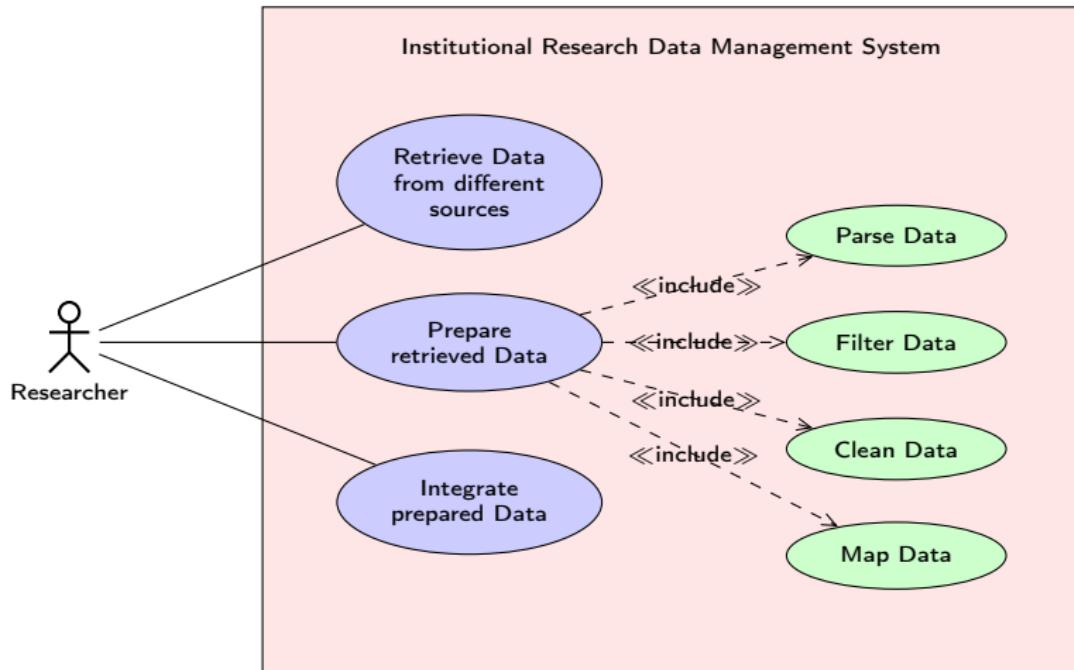


Abbildung 27: Linked Open Data-Workflow zur Sammlung, Bereinigung und Harmonisierung statistischer Daten aus verschiedenen Quellen (vgl. Zzapilko und Mathiak 2011, p. 116)

Ontologien vs. Datenmodelle?

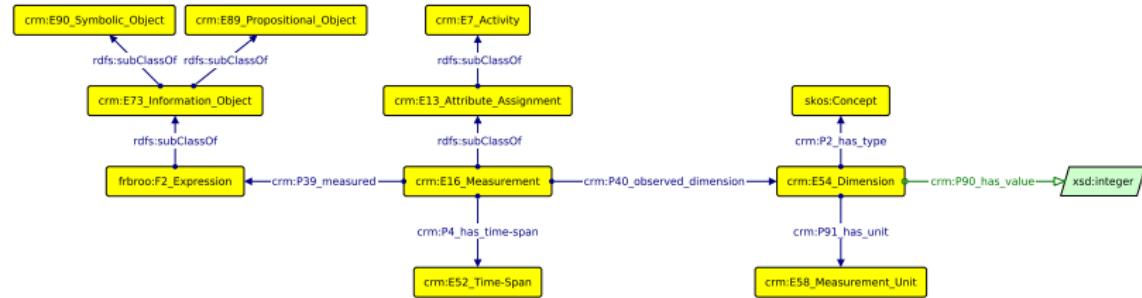


Abbildung 28: Repräsentation einer Messung (Beobachtung) und der dabei beobachteten Größe in CIDOC CRM



[T]here is a worrying practice associated with the creation of linked data that we may call the use of *Frankenstein ontologies*, in which linked data publishers decide on the vocabularies to be used for the annotation of their data items and select concepts and properties from diverse ontologies [...] without checking whether they are actually compatible or whether the original semantics of the reused terms are preserved in the ontology being developed.

— Corcho, Poveda-Villalón und Gómez-Pérez (2015)
“Ontology Engineering in the Era of Linked Data”

Datenmodelle und Anwendungsprofile

- Um beim Aufbau eines Datenmodells keine ‚Frankenstein-Ontologie‘ (vgl. Corcho, Poveda-Villalón und Gómez-Pérez 2015) zu konstruieren, empfehlen wir den Entwurf von Datenmodellen als Anwendungsprofile.
- Anwendungsprofile werden definiert als Metadatenschemas “which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application” (Heery und Patel 2000).
- SHACL (Shapes Constraint Language) eignet sich sehr gut, um die *node shapes* und *property shapes* für ein Anwendungsprofil festzulegen.
- SHACL ermöglicht auch die automatische Validierung der gemäß dem im Anwendungsprofil definierten Datenmodell modellierten Daten.