Panagiotis Kitmeridis, Lars G. Svensson

Datensetbeschreibungen in DCAT - Modell und Implementierung

DEUTSCHE NATIONAL BIBLIOTHEK

“Data is the New Oil”

Picture by KamiPhuc (CC BY): https://flic.kr/p/in4qSv

“The data-driven economy stimulates research and innovation on data, and increases business opportunities and availability of knowledge and capital across Europe.”

(*Elements of the European data economy strategy*)

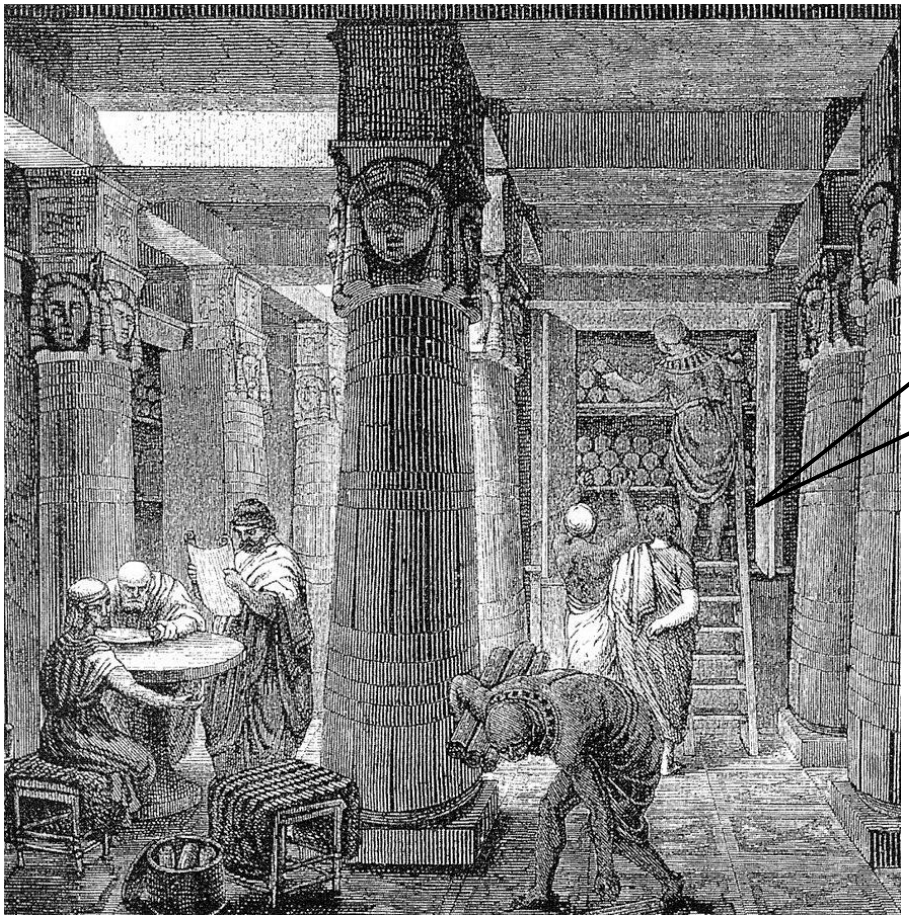
Zitat wahrsch. von Clive Humby. Vgl. https://medium.com/project-2030/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294

DEUTSCHE NATIONAL BIBLIOTHEK

Für uns (Meta-)Datenexperten ist das natürlich eine Binsenweisheit…

DEUTSCHE NATIONAL BIBLIOTHEK

Damit die Datenindustrie friktionsfrei läuft, muss man erst die Daten finden



"The Great Library of Alexandria" (Public Domain): https://commons.wikimedia.org/wiki/File:Ancientlibraryalex.jpg#/media/File:Ancientlibraryalex.jpg

Sind die Daten da oben?

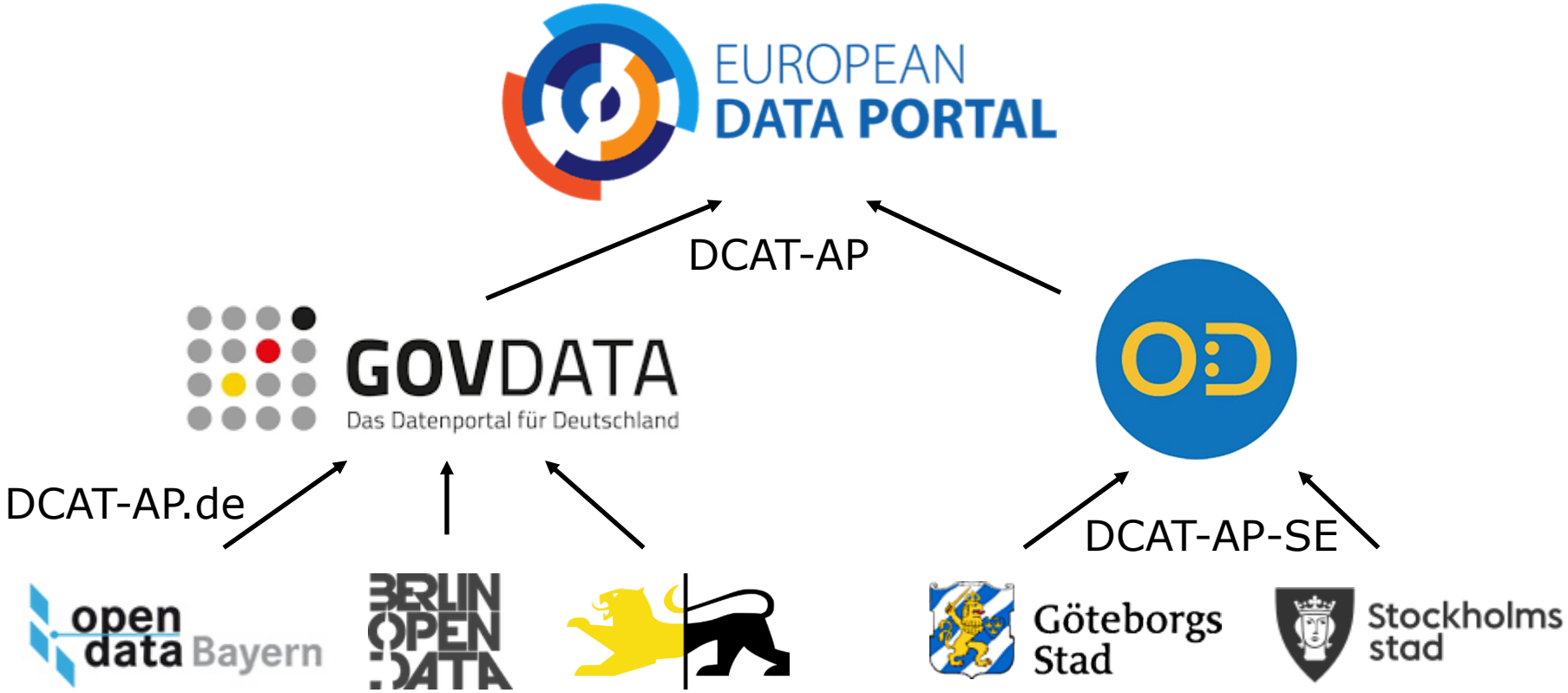
Dafür gibt es dedizierte Datenportale

https://www.europeandataportal.eu/

https://www.govdata.de/

DEUTSCHE NATIONAL BIBLIOTHEK

DCAT sorgt dafür, dass der Datenaustausch zwischen den Portalen interoperabel ist



DEUTSCHE NATIONAL BIBLIOTHEK

DCAT ist ein RDF-Vokabular um Datensets zu beschreiben

Eine **Datenstruktur (Dataset)** ist eine sinnvolle Sammlung von zusammenhängenden Daten, die von einer einzelnen Quelle veröffentlicht oder kuratiert wird und in einem oder mehreren Formaten erreichbar ist oder als Download zur Verfügung steht.

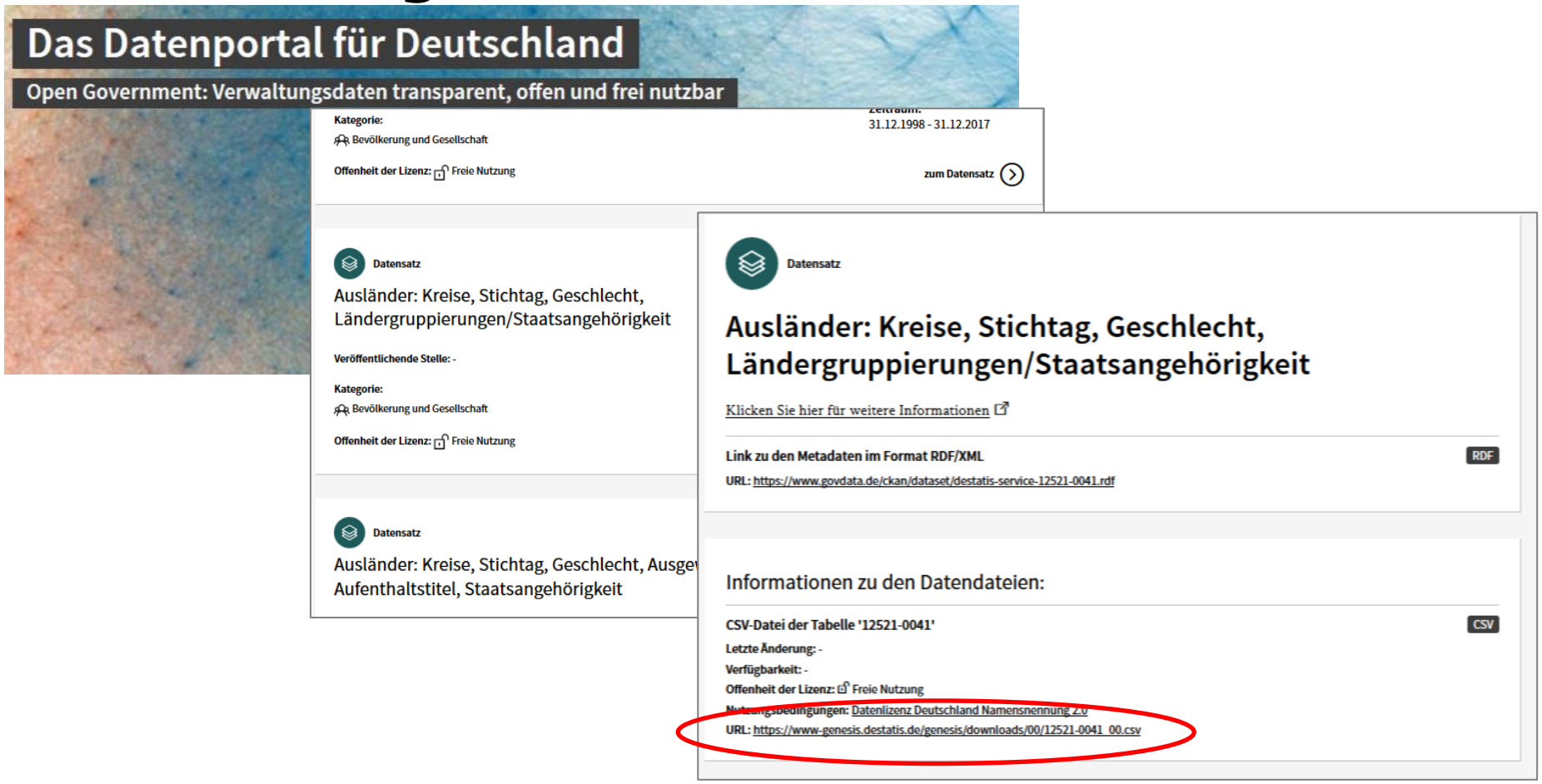
Ein Datenportal ist ein Web-basiertes System, welches einen Datenkatalog enthält, in dem Datenstrukturen mittels Metadaten beschrieben werden. Des Weiteren stellt ein Datenportal Dienste zur Recherche und Wiederverwendung von Datenstrukturenn [sic!] bereit.
(Quelle: DCAT-AP.de)

Weiter hat **schema.org** auch Klassen und Properties um Datensets zu beschreiben mit dem Ziel der Indexierung in Suchmaschinen.

Und **VoID** bietet ein Vokabular um RDF-Daten zu beschreiben

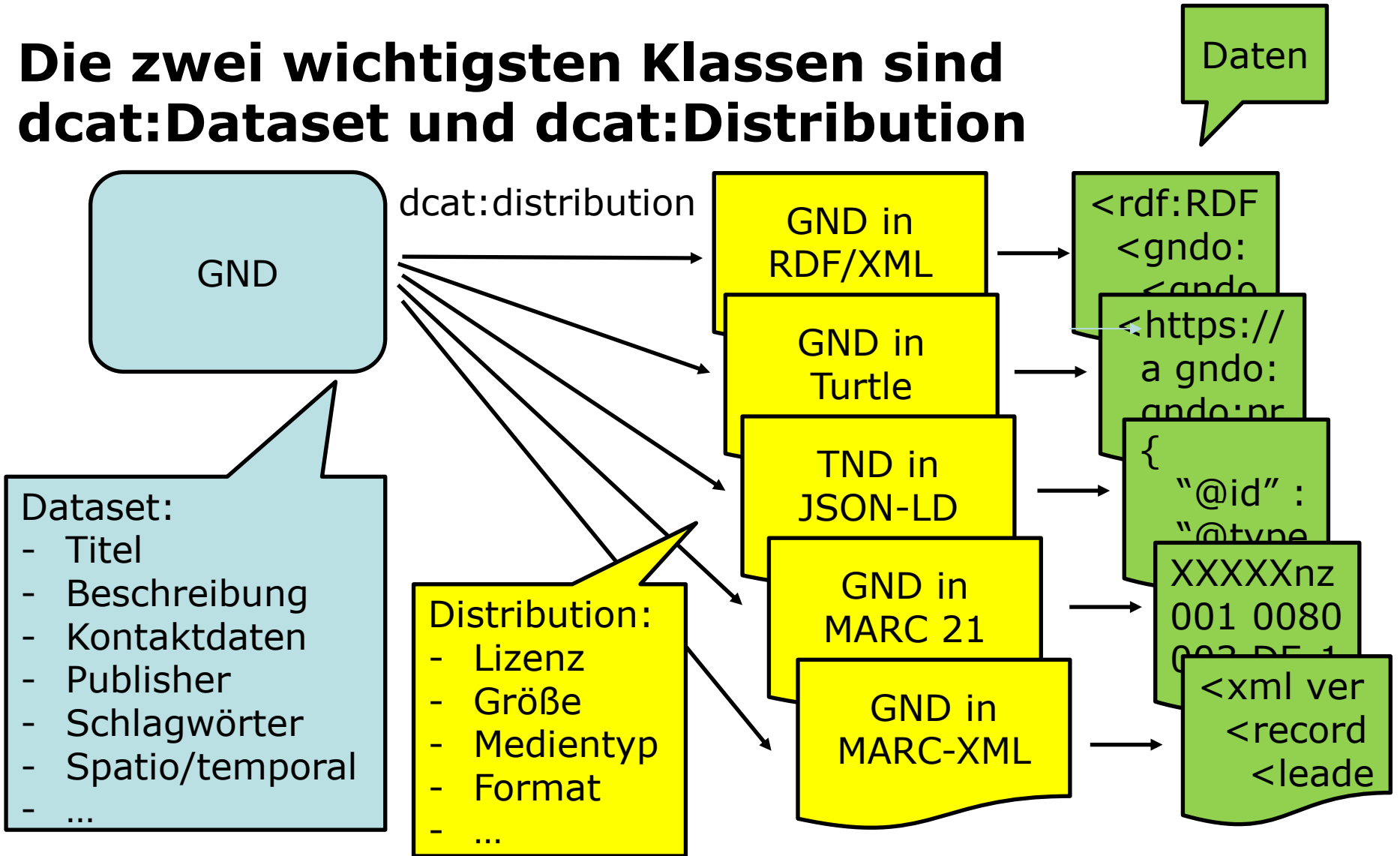
DEUTSCHE NATIONAL BIBLIOTHEK

Die angedachte Suche geht von einem Datenkatalog zu den konkreten Daten



DEUTSCHE NATIONAL BIBLIOTHEK

Die zwei wichtigsten Klassen sind dcat:Dataset und dcat:Distribution



Datenportale haben verschiedene DCAT-Anwendungsprofile für ihre Metadaten

Ein **Application Profile** ist eine Spezifikation, die Begrifflichkeiten bzw. Konzepte eines oder mehrerer grundlegender Standards weiterverwendet. Eine größere Bestimmtheit wird erreicht, indem für eine bestimmte Anwendung Klassen und Klassenattribute (Eigenschaften) als *obligatorisch*, *empfohlen* oder *optional* eingeordnet werden. Zusätzlich werden Empfehlungen für die Verwendung von kontrollierten Vokabularen gegeben.
Quelle: https://www.dcat-ap.de/def/dcatde/1.0.1/spec/specification.pdf

Beispiele

EU
DCAT-AP (EU-Portale)
GeoDCAT-AP (Geodaten, z. B. INSPIRE)
StatDCAT-AP (Statistische Daten)

Nationale
DCAT-AP.de (Deutschland, GovData)
DCAT-AP-SE (Schweden)

DEUTSCHE NATIONAL BIBLIOTHEK

DXWG hat den Auftrag, DCAT zu überarbeiten und weiterzuentwickeln

58 Teilnehmer von 5 Kontinenten
Chairs: Karen Coyle (DCMI),
Peter Winstanley (gov.scot)
Vertreten:

- Bibliotheken (u. a. DNB, OCLC)
- Universitäten und Forschungseinrichtungen
- EU-Organe (JRC)
- Industrie (u.a. Refinitiv, Geonovum)
- Öffentliche Verwaltung

W3C®

Viele Datenportale bieten inzwischen APIs für den Zugriff aber DCAT kennt nur Dateien

authorities-geografikum_lds_20190213.jsonld.gz	2019-03-08 10:44	31M	306280	I
authorities-geografikum_lds_20190213.rdf.gz	2019-03-08 10:54	31M	306280	I
authorities-geografikum_lds_20190213.ttl.gz	2019-03-08 10:41	31M	306280	I
authorities-koerperschaft_lds_20190213.jsonld.gz	2019-03-08 10:44	158M	1486660	
authorities-koerperschaft_lds_20190213.rdf.gz	2019-03-08 10:55	158M	1486660	
authorities-koerperschaft_lds_20190213.ttl.gz	2019-03-08 10:41	154M	1486660	
authorities-kongress_lds_20190213.jsonld.gz	2019-03-08 10:44	74M	806715	I
authorities-kongress_lds_20190213.rdf.gz	2019-03-08 10:54	72M	806715	I
authorities-kongress_lds_20190213.ttl.gz	2019-03-08 10:41	72M	806715	I
authorities-name_lds_20190213.jsonld.gz	2019-03-08 10:44	422M	7042949	
authorities-name_lds_20190213.rdf.gz	2019-03-08 11:06	359M	7042949	
authorities-name_lds_20190213.ttl.gz	2019-03-08 10:41	404M	7042949	
authorities-person_lds_20190213.jsonld.gz	2019-03-08 10:45	793M	5017126	
authorities-person_lds_20190213.rdf.gz	2019-03-08 10:54	745M	5017126	
authorities-person_lds_20190213.ttl.gz	2019-03-08 10:42	762M	5017126	
authorities-sachbegriff_lds_20190213.jsonld.gz	2019-03-08 10:45	18M	211605	I
authorities-sachbegriff_lds_20190213.rdf.gz	2019-03-08 10:54	17M	211605	I
authorities-sachbegriff_lds_20190213.ttl.gz	2019-03-08 10:42	17M	211605	I
authorities-werk_lds_20190213.jsonld.gz	2019-03-08 10:45	35M	372053	I
authorities-werk_lds_20190213.rdf.gz	2019-03-08 10:54	34M	372053	I
authorities-werk_lds_20190213.ttl.gz	2019-03-08 10:42	33M	372053	I
authorities_entityfacts_20190305.jsonld.gz	2019-03-12 13:47	728M	6741310	

https://data.dnb.de/opendata/


```
{
  "gndIdentifier" : "16187379-0",
  "id" : "http://d-nb.info/gnd/16187379-0",
  "preferredName" : "Schloss (Mannheim)",
  "wikipedia" : [
  {
    "id" : "https://de.wikipedia.org/wiki/Schloss_Mannheim",
    "label" : https://de.wikipedia.org/wiki/Schloss_Mannheim
  } ],
  "type" : [ "CorporateBody", "AuthorityResource" ],
  "geographicAreaCode" : [
  {
    "id" : "http://d-nb.info/standards/vocab/gnd/geographic-a
    "label" : "Deutschland„
  },
  {
    "id" : "http://d-nb.info/standards/vocab/gnd/geographic-a
    "label" : "Deutschland, Deutsches Reich„
  } ],
  "placeOfBusiness" : [
  {
    "id" : "http://d-nb.info/gnd/4037372-1",
    "label" : "Mannheim„
```

DEUTSCHE NATIONAL BIBLIOTHEK

Die Forschungsdatencommunities wollen Provenienz- und Zeit-/Raum-Information

- **Woher** kommen die Daten, wie sind sie entstanden?
- **Welchen geografischen Raum** decken die Daten ab?
- **Welchen Zeitraum** decken die Daten ab?

*Ich will Daten zur **Luftqualität** in **Mannheim** für **2018**, die die **Deutsche Umwelthilfe** überprüft haben*

Quelle: LUBW
https://www.lubw.baden-wuerttemberg.de/documents/10184/221750/DEBW098.png

DEUTSCHE NATIONAL BIBLIOTHEK

Und alle wollen Best Practices für Nutzung und Dokumentation von Anwendungsprofilen

Foto von Bran (PD): https://commons.wikimedia.org/wiki/File:Steckdose.jpg

Foto von secretlondon (PD): https://commons.wikimedia.org/wiki/File:BritishPlugforWikipedia.jpg

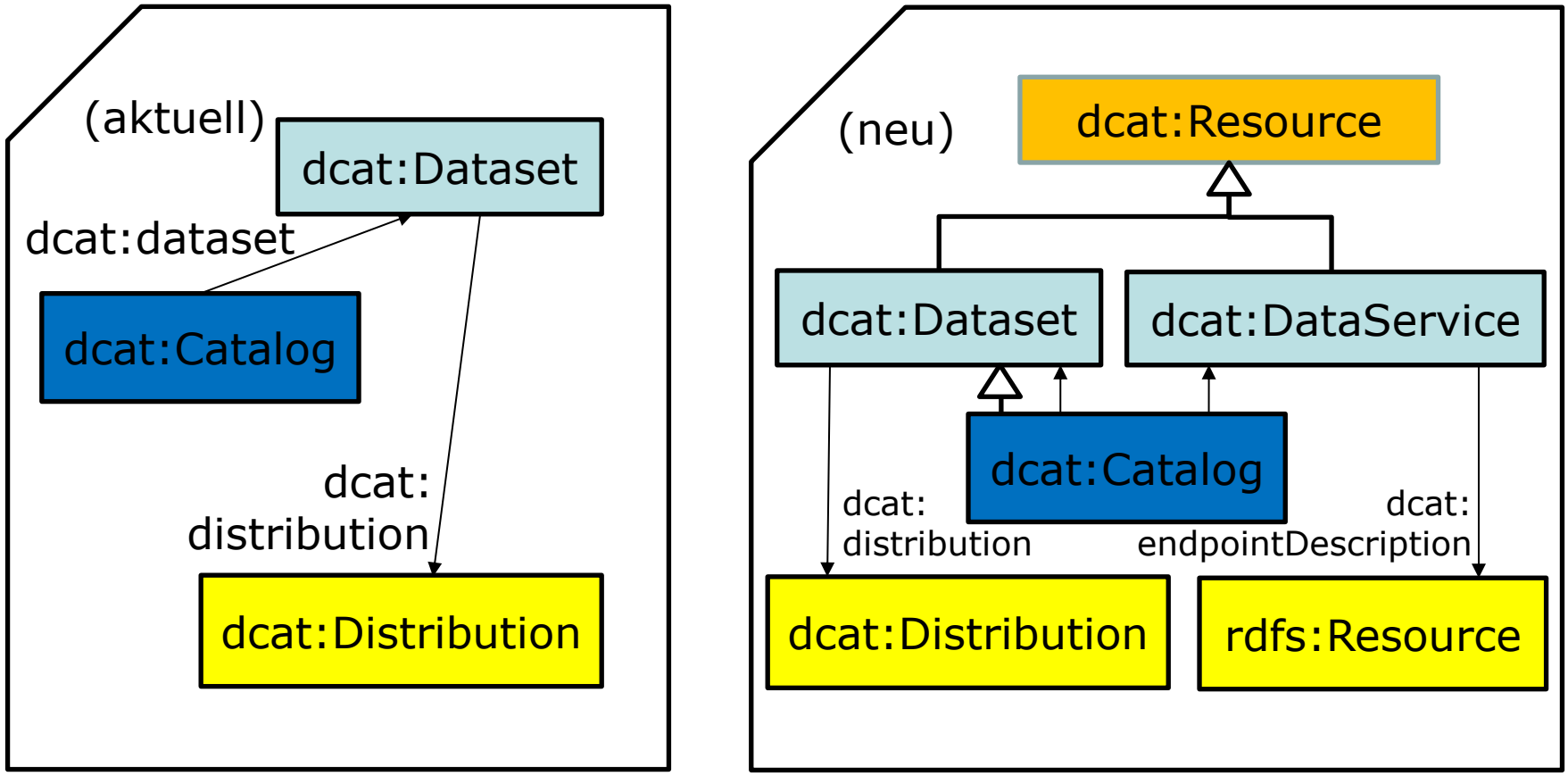
Foto von Mattes (PD): https://commons.wikimedia.org/wiki/File:Reisestecker.jpg

230 V
50 Hz

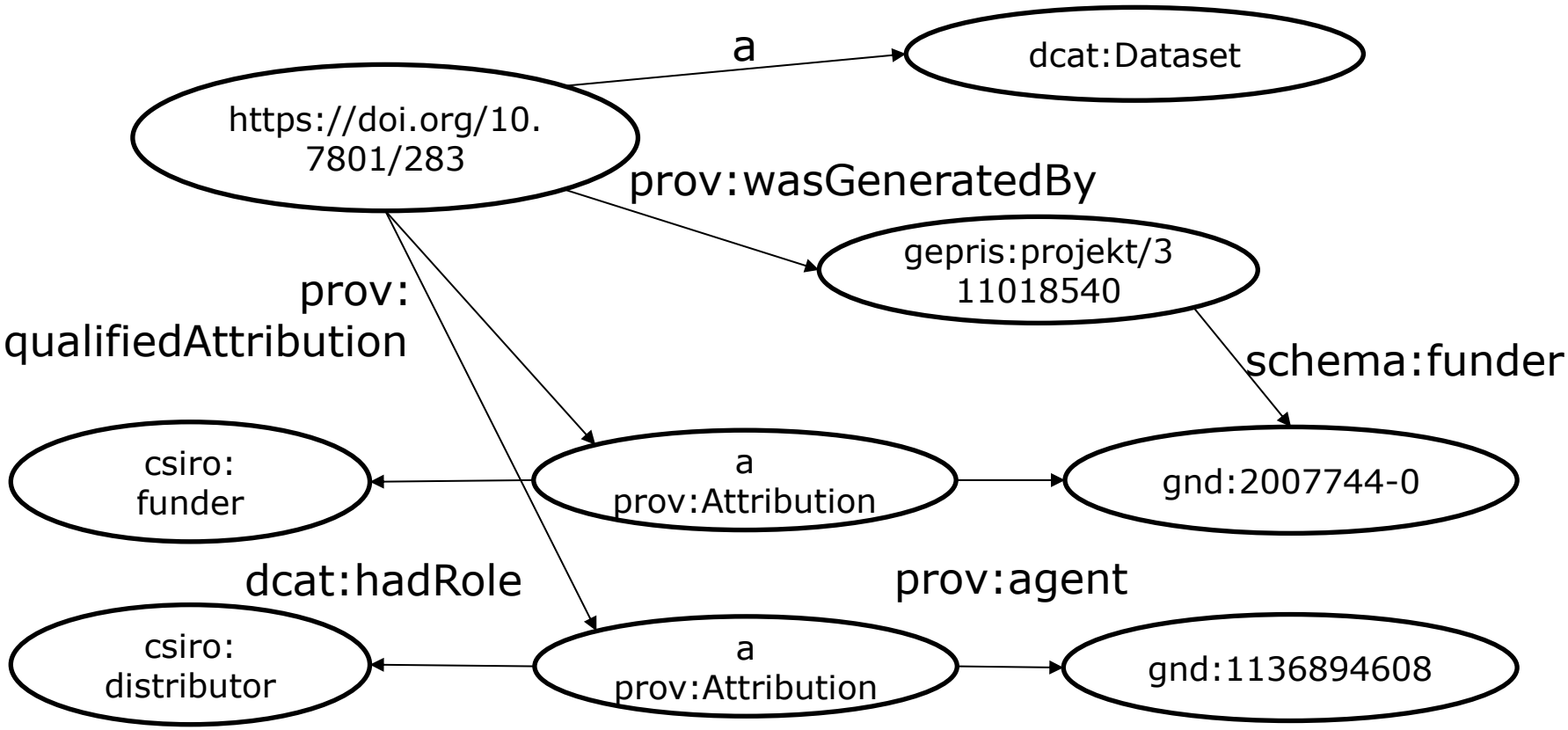
110 V
60 Hz

Foto von Chameleon (PD): https://commons.wikimedia.org/wiki/File:B_plug.jpg

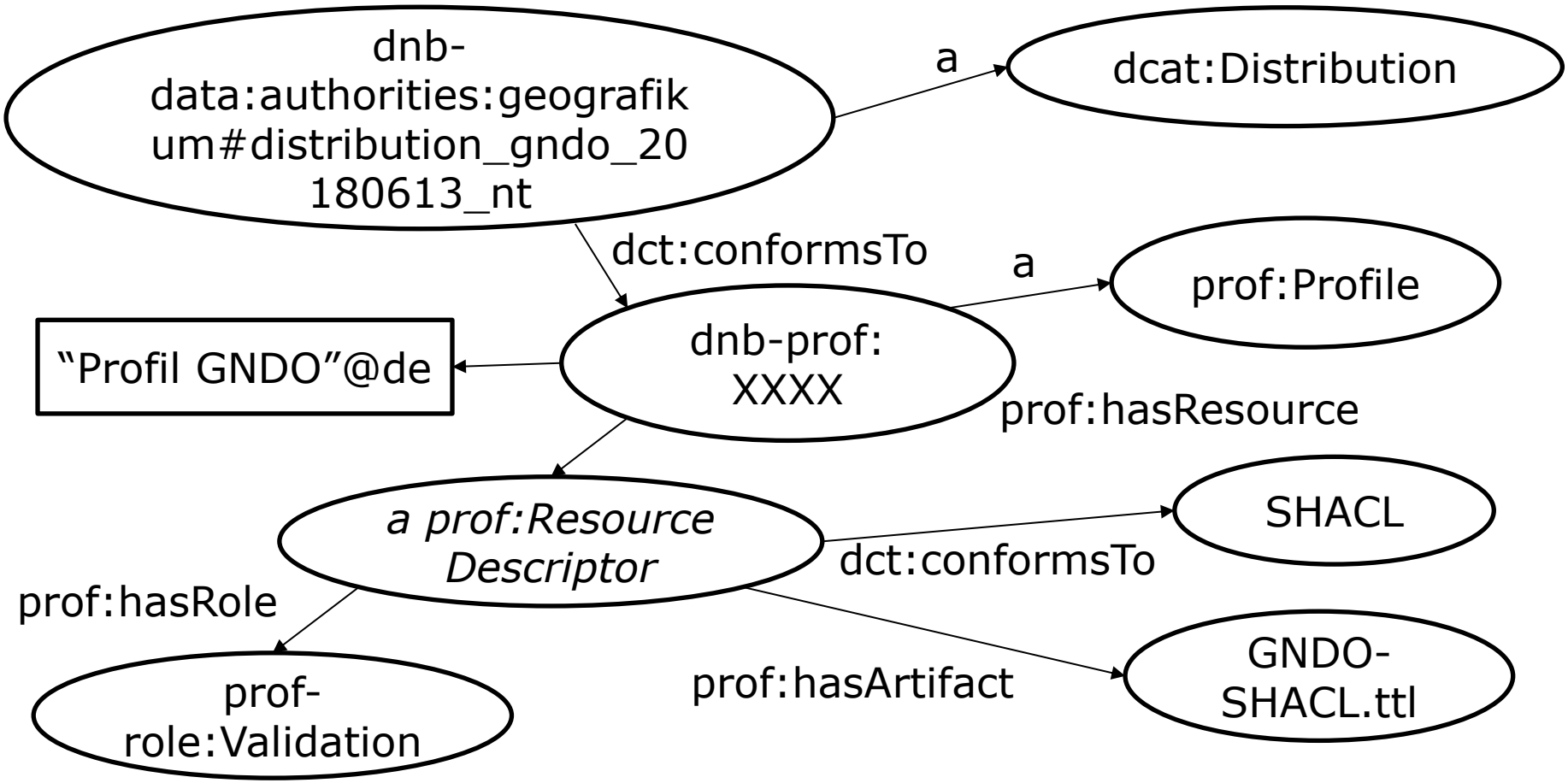
DCAT 2 führt neue Strukturen ein, bleibt aber abwärtskompatibel



Durch die Verwendung von prov-o kann man Datensets an bestimmte Projekte binden



Konformität mit Anwendungsprofilen kann man mit dct:conformsTo ausdrücken



DEUTSCHE NATIONAL BIBLIOTHEK

1. **Motivation**
2. Ziel des Projektes
3. **Umsetzung**
4. Statistik

DEUTSCHE NATIONAL BIBLIOTHEK

Motivation I (Gesamtziele allgemein)

- Suchmaschinenoptimierte Homepage für Datendienste zwecks Auffindbarkeit und Zugriff auf veröffentlichte Daten (Hinweis aus W3C-Workshop Open Data on the Web 2013

  [www.w3.org/2013/04/odw]
- Optimierung für Suchmaschinen und Crawler durch maschinenlesbare Informationen
- Präsenz in Suchergebnissen verbessern
- Informellen Gesamteinstieg bieten indem wir strukturierten Metadaten über unsere Daten anbieten

DEUTSCHE NATIONAL BIBLIOTHEK

Motivation II (Projektziele)

Vor Projekt: Datendienstwebseite der DNB stellt keine Datensetbeschreibungen und maschinenlesbare Informationen bereit

1. Vereinheitlichung (Single-Entry-Point) der Datensetbeschreibung zu den Datensets und deren Bereitstellung über d-nb.info (PermalinkURL)
2. Sichtbarkeit, Auffindbarkeit, direkter Verweis auf unsere Datensets
3. Datensetbeschreibung soll Gesamtüberblick über unsere Datensets verschaffen und den Zugriff vereinfachen
4. Zugriffsstatistik soll Mehrwert aufzeigen

DEUTSCHE NATIONAL BIBLIOTHEK

Ziele

- Infos zu welche Objekte haben wir
- Infos zu welche Formate bieten wir an
- autom. (Roboter) Zugriffe (Crawler) ermitteln
- Wie oft werden die Daten allgemein angefragt?
- In welchen Formaten werden sie angefragt?

Eine AP-konforme Datensetbeschreibung muss Instanzen der **Klassen dcat:Catalog, dcat:Dataset** und **dcat:Distribution** beinhalten. Für die Beschreibung der Instanzen der jeweiligen Klassen gibt es wiederum Vorgaben über vorgeschriebene, empfohlene und optionale Properties.

DEUTSCHE NATIONAL BIBLIOTHEK

dcat:Catalog / schema:DataCatalog

Mandatory	Recommended	Optional	Schema.org-Equivalent	VOID	Wert kommt woher?
dcat:dataset (umgesetzt)			schema:dataset (umgesetzt)		Fest eingetragen (http://d-nb.info/authorities bzw. http://d-nb.info/titles)
dct:description (umgesetzt)			schema:description (umgesetzt)		Fest eingetragen, Text kommt von 2D
dct:publisher (umgesetzt)			schema:publisher (umgesetzt)		http://ld.zdb-services.de/resource/organisations/DE-588
dct:title (umgesetzt)			schema:name (umgesetzt)		Fest eingetragen („Daten der DNB" oder so etwas ähnlich. Text von 2D)
	foaf:homepage				Setzen wir noch nicht
	dct:language		schema:inLanguage(?)		Setzen wir noch nicht
	dct:license (umgesetzt)		schema:license (umgesetzt)		Fest eingetragen: https://creativecommons.org/publicdomain/zero/1.0/
	dct:issued (umgesetzt)		schema:datePublished (umgesetzt)		Fest eingetragen: Datum der Erstveröffentlichung in ISO 8601
	dct:modified (umgesetzt)		schema:dateModified (umgesetzt)		Wird eingetragen wenn wir etwas an der Seite ändern (wohl recht selten), in ISO 8601
	dcat:themeTaxonomy				Setzen wir noch nicht
		dcat:hasPart			Setzen wir noch nicht
		dcat:isPartOf			Setzen wir noch nicht
		dcat:record			Setzen wir noch nicht
		dct:rights			Setzen wir noch nicht
		dct:spatial			Setzen wir noch nicht

<https://d-nb.info/datasets/#catalogue> a dcat:Catalog ;
dct:description "The German National Library's Catalogue of Datasets is the central point of access to the regularly produced data dumps."@en , "Der Katalog der Datensets der Deutschen Nationalbibliothek ist der zentrale Zuganspunkt zu den regelmäßig zur Verfügung gestellten Datenabzügen."@de ;
dct:license <http://dcat-ap.de/def/licenses/cc-zero> ;
dct:modified "2018-09-21"^^xsd:date ;
dct:publisher isil:DE-101 ;
dct:title "Katalog der Datensets / Deutsche Nationalbibliothek"@de , "Catalogue of Datasets / German National Library"@en ;
dcat:dataset <https://d-nb.info/datasets/authorities:sachbegriff#dataset> ,
<https://d-nb.info/datasets/authorities:name#dataset> ,
<https://d-nb.info/datasets/authorities:person#dataset> ,
<https://d-nb.info/datasets/authorities:werk#dataset> ,
<https://d-nb.info/datasets/authorities#dataset> ,
<https://d-nb.info/datasets/bib#dataset> ,
<https://d-nb.info/datasets/authorities:geografikum#dataset> ,
<https://d-nb.info/datasets/dnb-all#dataset> ,
<https://d-nb.info/datasets/zdb#dataset> ,
<https://d-nb.info/datasets/authorities:kongress#dataset> ,
<https://d-nb.info/datasets/authorities:koerperschaft#dataset> .

DEUTSCHE NATIONAL BIBLIOTHEK

dcat:Dataset / schema:Dataset

Mandatory	Recommended	Optional	Schema.org-Equivalent	VOID	Wert kommt woher?
dct:description (umgesetzt)			schema:description (umgesetzt)		Fest eingetragen, Text kommt von 2D (Beschreibung des Titeldatensets und der GND)
dct:title (umgesetzt)			schema:name (umgesetzt)		Fest eingetragen („Gemeinsame Normdatei" bzw. „Alle Titeldaten der DNB" oder so etwas ähnlich. Text von 2D)
	dcat:contactPoint				
	dcat:distribution (umgesetzt)		schema:distribution (umgesetzt)		Änderung jedes Mal, wenn wir eine neue Distribution (neues Veröffentlichungsdatum) erzeugen.
	dcat:keyword		schema:keywords		
	dct:publisher (umgesetzt)		schema:publisher (umgesetzt)		http://ld.zdb-services.de/resource/organisations/DE-588
	dcat:theme (dct:subject)		schema:about		
		dct:accessRights			
		dct:conformsTo			
		foaf:page			http://dnb.de/gnd; http://d-nb.de/nationalbibliografie bzw. deren englishen Äquivalenten
		dct:accrualPeriodicity			
		dct:hasVersion			
		dct:identifier			
		dct:isVersionOf			
		dcat:landingPage			
		dct:language			
		adms:identifier			
		dct:provenance			

<**https://d-nb.info/datasets/authorities#dataset**> a dcat:Dataset ;
dct:description "The Integrated Authority File (GND) is an […]other institutions."@en , "Die Gemeinsame Normdatei (GND) ist eine Normdatei für Personen, Körperschaften, Konferenzen, Geografika, Sachschlagwörter und Werktitel, die vor allem zur Katalogisierung von Literatur in Bibliotheken dient, zunehmend aber auch von Archiven, Museen, Projekten und in Webanwendungen genutzt wird. Sie wird von der Deutschen Nationalbibliothek, allen deutschsprachigen Bibliotheksverbünden mit den angeschlossenen Bibliotheken, der Zeitschriftendatenbank (ZDB) und zahlreichen weiteren Einrichtungen gemeinschaftlich geführt."@de ;
dct:publisher isil:DE-588 ;
dct:title "Gemeinsame Normdatei (GND)"@de , "Integrated Authority File (GND)"@en ;
dcat:contactPoint [a
vcard:Organization ;
vcard:fn "Data Services" ;
vcard:hasEmail <mailto:datendienste@dnb.de> ;
vcard:hasTelephone "+49 69 1525 1630"
] ;
dcat:distribution
<https://d-nb.info/datasets/authorities#distribution_gndo_20180613_hdt> ,
<https://d-nb.info/datasets/authorities#distribution_ef_20180613_jsonld> .

DEUTSCHE NATIONAL BIBLIOTHEK

dcat:Distribution / schema:DataDownload

Mandatory	Recommended	Optional	Schema.org-Equivalent	VOID	Wert kommt woher?
dcat:accessURL (umgesetzt)			schema:contentUrl (umgesetzt)		Wahrsch. fest eingetragen (wenn wir immer die gleichen URLs wiederverwenden und keine versionsabhängigen haben
	dct:description (umgesetzt)		schema:description (umgesetzt)		Wohl hauptsächlich fest eingetragen („Die DNB als RDF in Turtle"). Text kommt wohl von 2D.
	dct:format (umgesetzt)		schema:encodingFormat (umgesetzt)		Fest eingetragen („zip", „gzip", ...)
	dct:license (umgesetzt)		schema:license (umgesetzt)		Fest eingetragen: https://creativecommons.org/publicdomain/zero/1.0/
		dcat:byteSize (umgesetzt)	schema:contentSize (umgesetzt)		Tatsächliche Dateigröße (immer veränderlich)
		spdx:checksum			
		foaf:page			Setzen wir noch nicht (wir werden wohl keine Einzelseiten pro Distribution haben)
		dcat:downloadURL			Setzen wir wahrscheinlich nicht (was ist Unterschied zu accessURL?)
		dct:language			
		dct:conformsTo			
		dcat:mediaType (umgesetzt)	schema:encodingFormat (umgesetzt)		Fest für jede Distribution (text/turtle, application/marcxml+xml), ...
		dct:issued	schema:datePublished		Jedes Mal neu eintragen (Veröffentlichungsdatum)
		dct:rights			Setzen wir wahrsch. nicht (unnötig, da wir CC0 verwenden)
		adms:status			

<**https://d-nb.info/datasets/authorities:name#distribution_gndo_20180613_ttl**> a dcat:Distribution ;
dct:format <http://publications.europa.eu/mdr/resource/authority/file-type/GZIP> ;
dct:license <http://dcat-ap.de/def/licenses/cc-zero> ;
dct:title "Gemeinsame Normdatei, GND-Entität Name, Profil GNDO, Format RDF (Turtle)"@de , "Integrated Authority File, GND-entity name, Profile GNDO, Format RDF (Turtle)"@en ;
dcat:accessURL <https://data.dnb.de/opendata/authorities-name_gndo_20180613.ttl.gz> ;
dcat:byteSize "309454928"^^xsd:decimal .

<**https://d-nb.info/datasets/authorities:koerperschaft#distribution_gndo_20180613_jsonld**> a
dcat:Distribution ;
dct:format <http://publications.europa.eu/mdr/resource/authority/file-type/GZIP> ;
dct:license <http://dcat-ap.de/def/licenses/cc-zero> ;
dct:title "Integrated Authority File, GND-entity corporate body, Profile GNDO, Format RDF (JSON-LD)"@en , "Gemeinsame Normdatei, GND-Entität Körperschaft, Profil GNDO, Format RDF (JSON-LD)"@de ;
dcat:accessURL <https://data.dnb.de/opendata/authorities-koerperschaft_gndo_20180613.jsonld.gz> ;
dcat:byteSize "135848946"^^xsd:decimal .

<**https://d-nb.info/datasets/zdb#distribution_dini-kim_20180206_hdt**> a dcat:Distribution ;
dct:format <http://publications.europa.eu/mdr/resource/authority/file-type/GZIP> ;
dct:license <http://dcat-ap.de/def/licenses/cc-zero> ;
dct:title "German Union Catalogue of Serials (ZDB), Profile DINI-KIM, Format RDF (HDT)"@en , "Katalog der Zeitschriftendatenbank, Profil DINI-KIM, Format RDF (HDT)"@de ;
dcat:accessURL <https://data.dnb.de/opendata/zdb_dini-kim_20180206.hdt.gz> ;
dcat:byteSize "163465851"^^xsd:decimal .

DEUTSCHE NATIONAL BIBLIOTHEK

Statistik Dump Downloads

Monat	GND .ttl	GND .rdf	GND .hdt	GND. jsonld	DNBTitel .ttl	DNBTitel .rdf	DNBTitel .hdt	DNBTitel .jsonld	ZDBTitel .ttl	ZDBTitel .rdf	ZDBTitel .hdt	ZDBTitel .jsonld	googlebot
März 2018	69	104	13	31	22	33	4	25	28	21	19	19	56
April 2018	34	57	10	37	25	41	6	16	17	22	21	17	6
Mai 2018	60	92	12	57	45	55	10	38	26	29	23	17	41
Juni 2018	58	58	20	41	43	49	12	36	24	28	27	29	38
Juli 2018	18	15	29	14	27	30	12	22	50	34	22	23	63
August 2018	36	31	41	42	66	108	33	62	41	54	47	46	125
September 2018	21	25	22	24	38	36	16	33	21	19	18	17	151
Oktober 2018	21	20	27	13	35	44	7	29	24	30	27	19	74
November 2018	17	30	16	10	24	50	11	22	27	26	28	20	100
Dezember 2018	21	21	25	22	20	30	14	34	24	23	27	22	111
Januar 2019	20	22	13	10	25	39	14	39	16	22	15	21	74
Februar 2019	9	8	21	7	11	6609	8	14	21	16	22	19	54

DEUTSCHE NATIONAL BIBLIOTHEK