Fraunhofer FOKUS

Fraunhofer-Institut für Offene Kommunikationssysteme

Fit for Use! Datenqualität mit Kontext

Juan Carlos Carvajal B.

Rückblick

Qua·li·tät

Aus dem lateinischen „qualitas“. Beschaffenheit, Eigenschaft.

FAIR Principles

Ausgangslage

Heterogene und schlechte Datenbasis

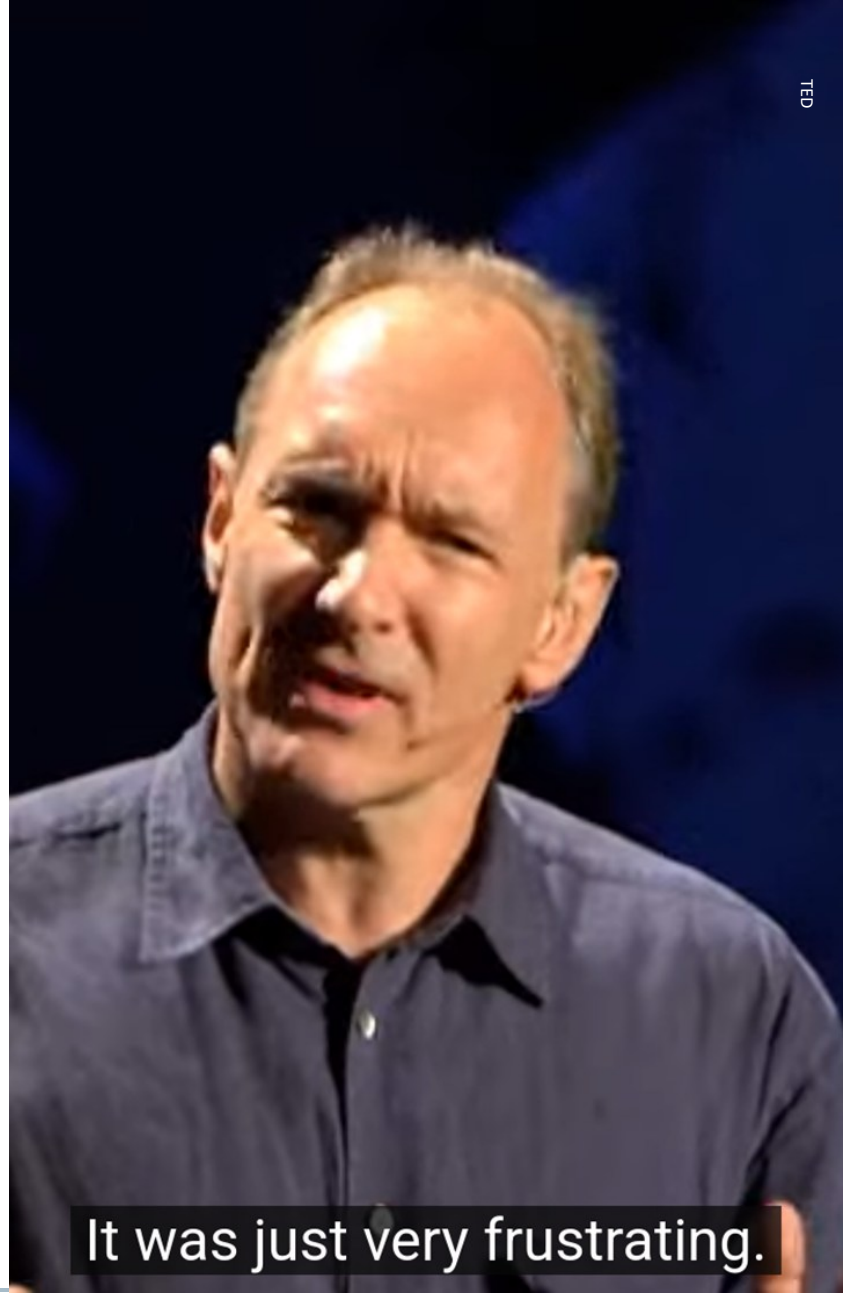
- Daten stammen aus unterschiedlichen Quellen und Formaten, was die Integration erschwert hat.

Keine maschinenlesbaren Formate

- Daten waren meist in unstrukturierten Formaten (z. B. PDFs, Textdokumente) vorhanden, die keine automatisierte Verarbeitung ermöglichten.

Vertraulichkeit „Database Hugging“ und Datensilos

- Unternehmen und Institutionen hielten ihre Daten oft geheim, was den Austausch und die Zusammenarbeit einschränkte.
- Fehlende Standards führten dazu, dass Daten in isolierten Systemen gespeichert wurden.



Erster Ansatz

Öffentlicher Zugang zu Daten

- Jedes Datenobjekt ist über eine eindeutige Adresse (http) zugänglich.

Einheitliche Formate

- Bei der Abfrage eines HTTP-Namens erhält man nützliche Informationen in einem standardisierten Format.

Datenverknüpfung

- Daten enthalten Verbindungen zwischen verschiedenen Entitäten, die ebenfalls durch HTTP-Namen referenziert werden.
- Durch so einen vernetzten Datenraum, wird eine redundante Datenhaltung vermieden und die Effizienz gesteigert.

Hintergrund – Offene Daten

FAIR steht für Findable, Accessible, Interoperable, Reusable

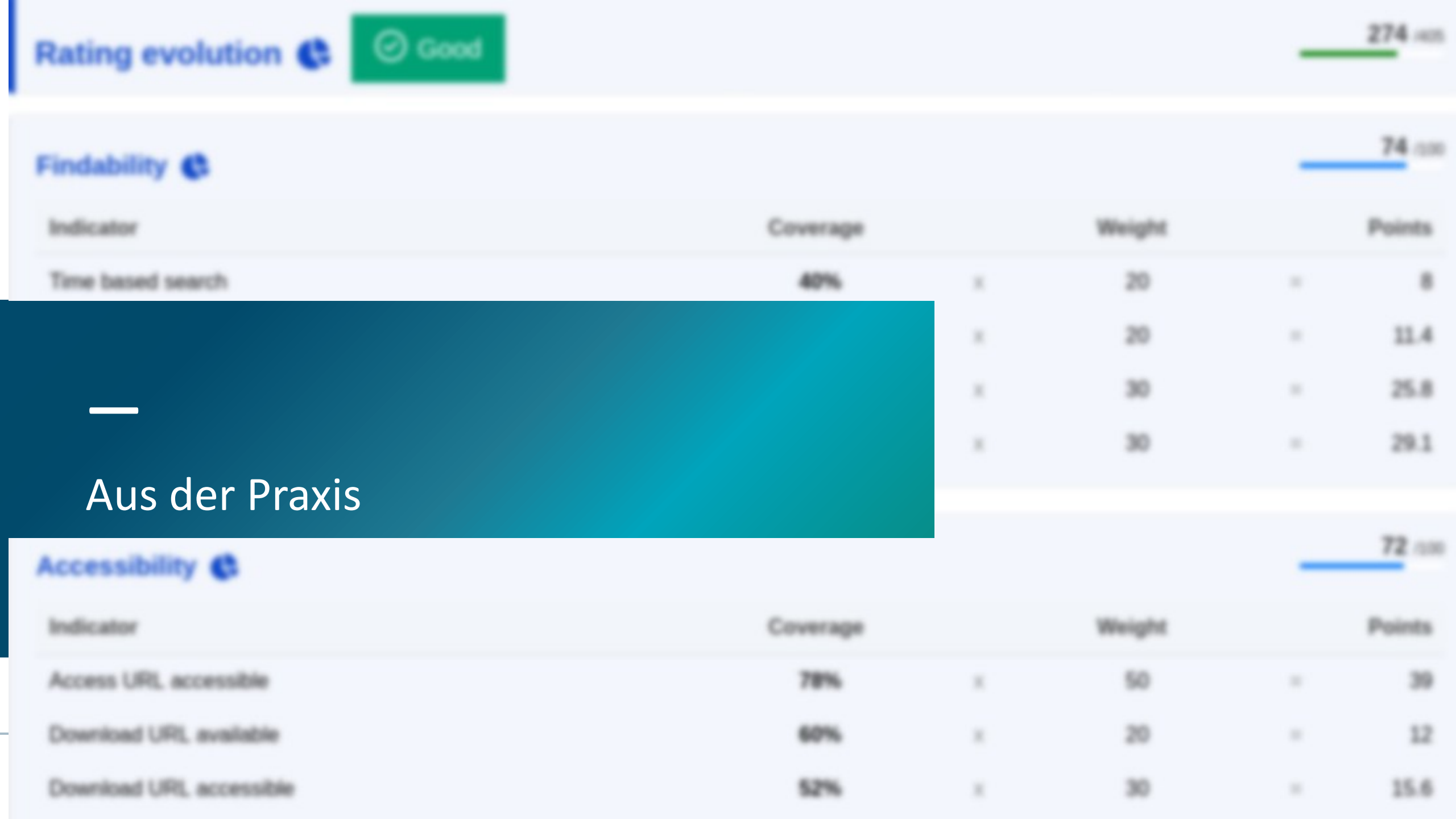
- Ziel: Daten so gestalten, dass sie leicht auffindbar, zugänglich, interoperabel und wiederverwendbar sind.

- Umsetzungsschwerpunkt: eindeutige Identifikatoren, umfangreiche Metadaten, Standardformate, klare Nutzungsbedingungen

Hintergrund – Offene Daten

Der Kontext vom FAIR war geprägt von geschlossene Datenquellen.

Das Ziel ist stark an die Veröffentlichung von Daten eingerichtet.


Rating evolution
Good
274 /405
Findability
74 /100
Indicator
Coverage
Weight
Points
Time based search
40%
x
20
=
8
x
20
=
11.4
x
30
=
25.8
x
30
=
29.1
Aus der Praxis
Accessibility
72 /100
Indicator
Coverage
Weight
Points
Access URL accessible
78%
x
50
=
39
Download URL available
60%
x
20
=
12
Download URL accessible
52%
x
30
=
15.6

DCAT-AP - Allgemeine Qualität

Allgemeine Qualität nach dem FAI Prinzipien

- Findability
- Accessibility
- Interoperability
- Reusability
- Contextuality

Rating evolution Good — 274 /405

Findability — 74 /100

Indicator	Coverage		Weight		Points
Time based search	40%	x	20	=	8
Geo search	57%	x	20	=	11.4
Keyword usage	86%	x	30	=	25.8
Categories	97%	x	30	=	29.1

Accessibility — 72 /100

Indicator	Coverage		Weight		Points
Access URL accessible	78%	x	50	=	39
Download URL available	60%	x	20	=	12
Download URL accessible	52%	x	30	=	15.6

Interoperability — 63 /110

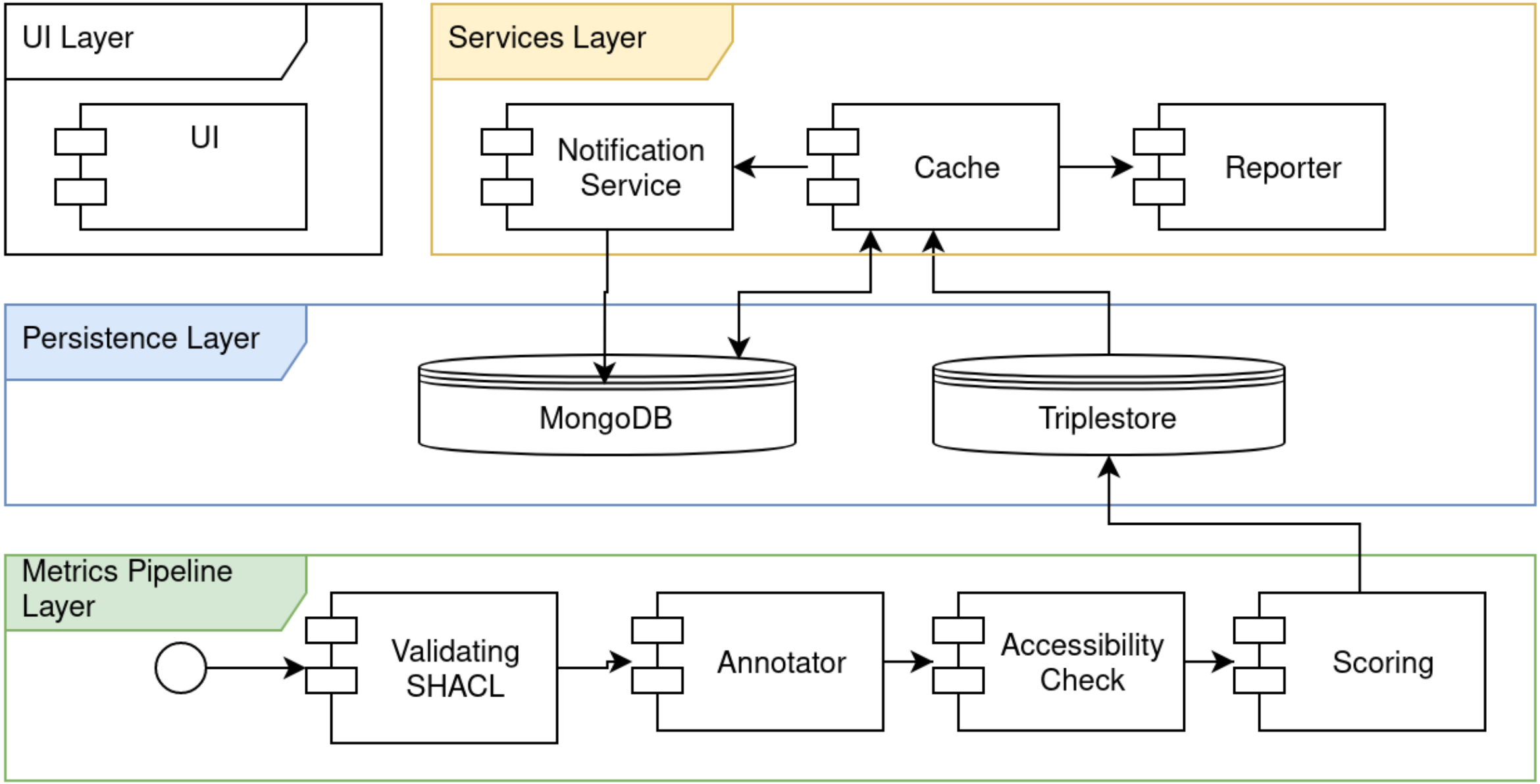
Indicator	Coverage		Weight		Points
Non-proprietary	59%	x	20	=	11.8
Format / Media type from vocabulary	79%	x	10	=	7.9
Machine readable	46%	x	20	=	9.2
DCAT-AP compliance	26%	x	30	=	7.8
Media type	56%	x	10	=	5.6
Format	79%	x	20	=	15.8

Reusability — 50 /75

Indicator	Coverage		Weight		Points
Contact point	100%	x	20	=	20
License information	100%	x	20	=	20
License vocabulary	0%	x	10	=	0
Access restrictions	0%	x	10	=	0
Publisher	100%	x	10	=	10
Access restrictions vocabulary	0%	x	5	=	0

Contextuality — 15 /20

Indicator	Coverage		Weight		Points
Date of issue	91%	x	5	=	4.55
Modification date	76%	x	5	=	3.8
File size	14%	x	5	=	0.7
Rights	91%	x	5	=	4.55

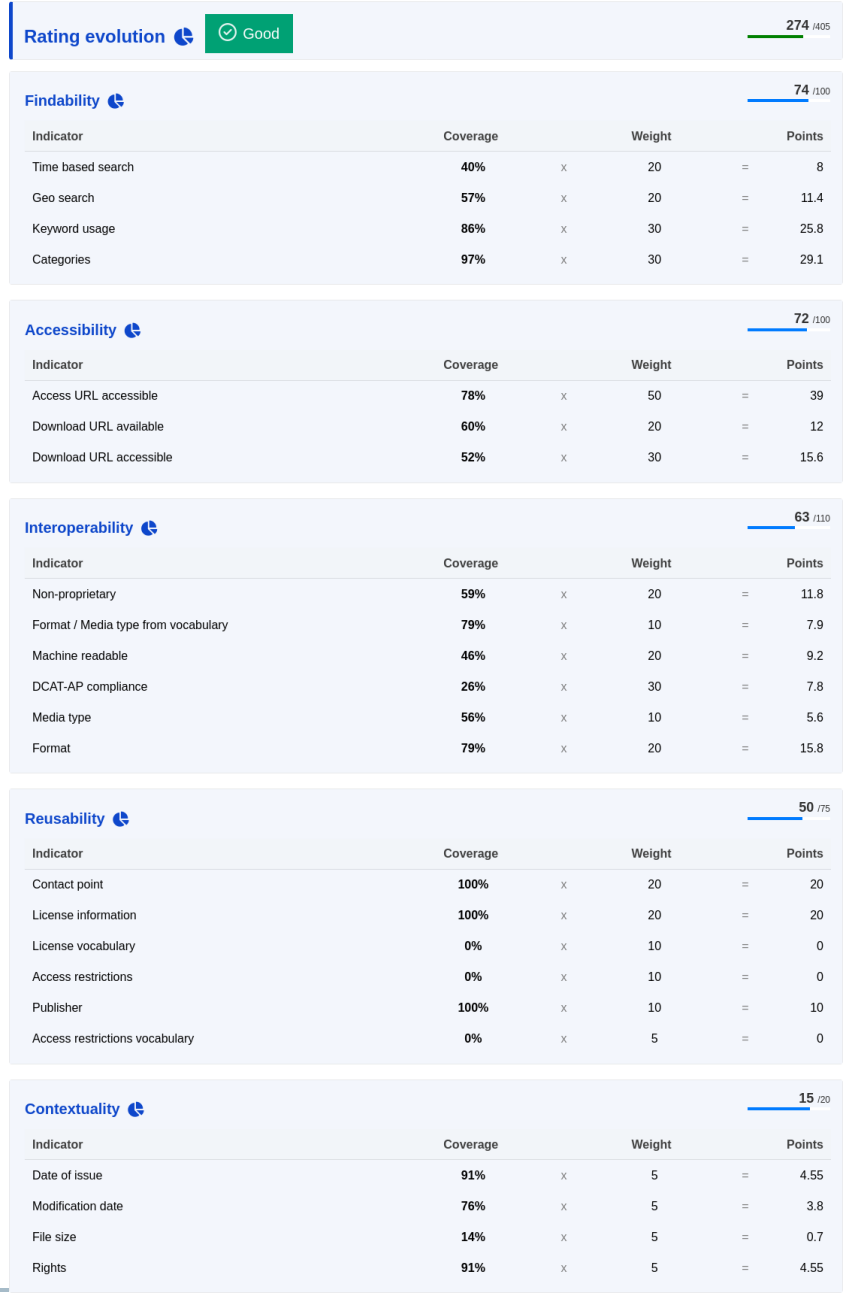


DCAT-AP - Allgemeine Qualität

URL Check, eine Gute Idee?

- Accessibility
- Technisch leider eine Herausforderung
- HEAD Request
- Rate Limiting

Brauchen wir eigentlich einen Crawler?



Rating evolution Good — 274 /405

Findability — 74 /100

Indicator	Coverage		Weight		Points
Time based search	40%	x	20	=	8
Geo search	57%	x	20	=	11.4
Keyword usage	86%	x	30	=	25.8
Categories	97%	x	30	=	29.1

Accessibility — 72 /100

Indicator	Coverage		Weight		Points
Access URL accessible	78%	x	50	=	39
Download URL available	60%	x	20	=	12
Download URL accessible	52%	x	30	=	15.6

Interoperability — 63 /110

Indicator	Coverage		Weight		Points
Non-proprietary	59%	x	20	=	11.8
Format / Media type from vocabulary	79%	x	10	=	7.9
Machine readable	46%	x	20	=	9.2
DCAT-AP compliance	26%	x	30	=	7.8
Media type	56%	x	10	=	5.6
Format	79%	x	20	=	15.8

Reusability — 50 /75

Indicator	Coverage		Weight		Points
Contact point	100%	x	20	=	20
License information	100%	x	20	=	20
License vocabulary	0%	x	10	=	0
Access restrictions	0%	x	10	=	0
Publisher	100%	x	10	=	10
Access restrictions vocabulary	0%	x	5	=	0

Contextuality — 15 /20

Indicator	Coverage		Weight		Points
Date of issue	91%	x	5	=	4.55
Modification date	76%	x	5	=	3.8
File size	14%	x	5	=	0.7
Rights	91%	x	5	=	4.55

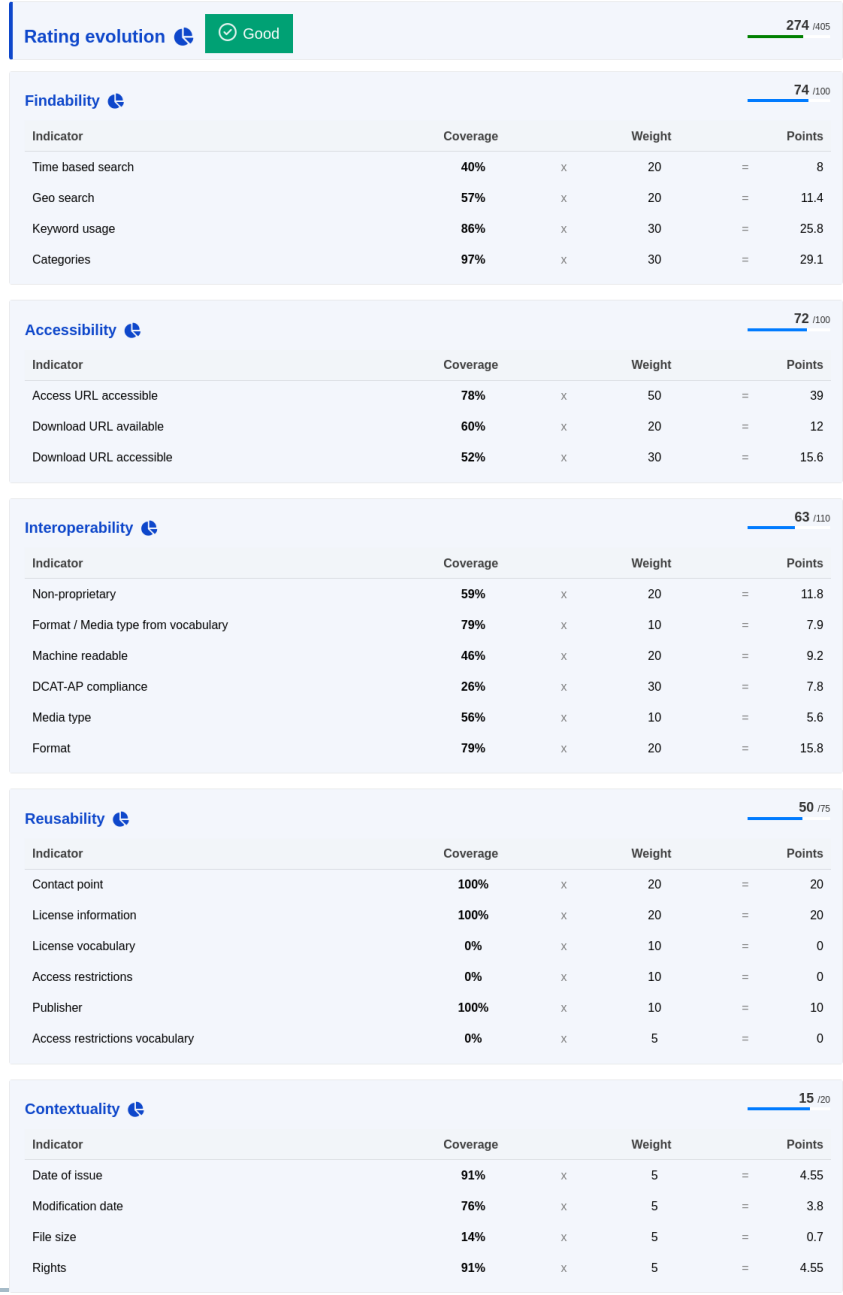
DCAT-AP - Allgemeine Qualität

DCAT-AP Schema Komformität:

- Kann nur auf Kataloguen angewendet werden, die Native RDF liefern.

Fragliche Annahmen:

- Alle Distributionen sind nur unterschiedliche Abbildungen eines Datensätzes
- Daher wird nur die beste Distribution pro Datensatz bewertet



Rating evolution Good — 274 /405

Findability — 74 /100

Indicator	Coverage		Weight		Points
Time based search	40%	x	20	=	8
Geo search	57%	x	20	=	11.4
Keyword usage	86%	x	30	=	25.8
Categories	97%	x	30	=	29.1

Accessibility — 72 /100

Indicator	Coverage		Weight		Points
Access URL accessible	78%	x	50	=	39
Download URL available	60%	x	20	=	12
Download URL accessible	52%	x	30	=	15.6

Interoperability — 63 /110

Indicator	Coverage		Weight		Points
Non-proprietary	59%	x	20	=	11.8
Format / Media type from vocabulary	79%	x	10	=	7.9
Machine readable	46%	x	20	=	9.2
DCAT-AP compliance	26%	x	30	=	7.8
Media type	56%	x	10	=	5.6
Format	79%	x	20	=	15.8

Reusability — 50 /75

Indicator	Coverage		Weight		Points
Contact point	100%	x	20	=	20
License information	100%	x	20	=	20
License vocabulary	0%	x	10	=	0
Access restrictions	0%	x	10	=	0
Publisher	100%	x	10	=	10
Access restrictions vocabulary	0%	x	5	=	0

Contextuality — 15 /20

Indicator	Coverage		Weight		Points
Date of issue	91%	x	5	=	4.55
Modification date	76%	x	5	=	3.8
File size	14%	x	5	=	0.7
Rights	91%	x	5	=	4.55

DCAT-AP - Allgemeine Qualität

Bestehende MQA-Methode wendet einheitliche Bewertungskriterien auf alle Datensätze an

- Führt zu systematischen Verzerrungen.

Unklare Hierarchie und Regelwerk

- Aggregation von Eigenschaften über Klassen hinweg (Dataset vs. Distribution) ohne klare Regeln.
- Bewertet nur eine Distribution pro Datensatz.

Rating evolution Good — 274 /405

Findability — 74 /100

Indicator	Coverage		Weight		Points
Time based search	40%	x	20	=	8
Geo search	57%	x	20	=	11.4
Keyword usage	86%	x	30	=	25.8
Categories	97%	x	30	=	29.1

Accessibility — 72 /100

Indicator	Coverage		Weight		Points
Access URL accessible	78%	x	50	=	39
Download URL available	60%	x	20	=	12
Download URL accessible	52%	x	30	=	15.6

Interoperability — 63 /110

Indicator	Coverage		Weight		Points
Non-proprietary	59%	x	20	=	11.8
Format / Media type from vocabulary	79%	x	10	=	7.9
Machine readable	46%	x	20	=	9.2
DCAT-AP compliance	26%	x	30	=	7.8
Media type	56%	x	10	=	5.6
Format	79%	x	20	=	15.8

Reusability — 50 /75

Indicator	Coverage		Weight		Points
Contact point	100%	x	20	=	20
License information	100%	x	20	=	20
License vocabulary	0%	x	10	=	0
Access restrictions	0%	x	10	=	0
Publisher	100%	x	10	=	10
Access restrictions vocabulary	0%	x	5	=	0

Contextuality — 15 /20

Indicator	Coverage		Weight		Points
Date of issue	91%	x	5	=	4.55
Modification date	76%	x	5	=	3.8
File size	14%	x	5	=	0.7
Rights	91%	x	5	=	4.55

MQA - Diagnose

- Datenbestand ist heterogen; domänenspezifische Katalogen (Geospatial, Temporal, Service) beeinflussen Bewertungen
- Ohne Klassenhierarchie fehlen klare Propagationsregeln von Distribution zu Dataset zu Catalogue
- Score spiegeln oft nur Übereinstimmung mit impliziten Annahmen wider, nicht Nutzen für konkrete Anwendungen

Coverage		Weight		Points
15%	x	20	=	3
74%	x	20	=	14.8
84%	x	30	=	25.2
81%	x	30	=	24.3

Rating evolution Good 274 /405

Findability 74 /100

Indicator	Coverage		Weight		Points
Time based search	**40%**	x	20	=	8
Geo search	**57%**	x	20	=	11.4
Keyword usage	**86%**	x	30	=	25.8
Categories	**97%**	x	30	=	29.1

Accessibility 72 /100

Indicator	Coverage		Weight		Points
Access URL accessible	**78%**	x	50	=	39
Download URL available	**60%**	x	20	=	12
Download URL accessible	**52%**	x	30	=	15.6

Interoperability 63 /110

Indicator	Coverage		Weight		Points
Non-proprietary	**59%**	x	20	=	11.8
Format / Media type from vocabulary	**79%**	x	10	=	7.9
Machine readable	**46%**	x	20	=	9.2
DCAT-AP compliance	**26%**	x	30	=	7.8
Media type	**56%**	x	10	=	5.6
Format	**79%**	x	20	=	15.8

Reusability 50 /75

Indicator	Coverage		Weight		Points
Contact point	**100%**	x	20	=	20
License information	**100%**	x	20	=	20
License vocabulary	**0%**	x	10	=	0
Access restrictions	**0%**	x	10	=	0
Publisher	**100%**	x	10	=	10
Access restrictions vocabulary	**0%**	x	5	=	0

Contextuality 15 /20

Indicator	Coverage		Weight		Points
Date of issue	**91%**	x	5	=	4.55
Modification date	**76%**	x	5	=	3.8
File size	**14%**	x	5	=	0.7
Rights	**91%**	x	5	=	4.55

Qualität Schemas

Qualitäts-Schema-Ansatz

Wir wollen einen Zweckspezifischen Ansatz einführen.

Qualität nach Use case.

- "Sind die Daten gut?" » "Wofür sind die Daten geeignet?"

Qualitäts-Schema

Schemas sind Property-Sets für spezifische Use Cases

Jedes Schema:

- Legt erforderliche/empfohlene Eigenschaften fest
- Verknüpft Schema mit konkreten Anwendungsszenarien
- Liefert fit-for-purpose-Indikatoren statt eines Gesamt-Scores

Qualitäts-Schema

- Beseitigung falscher Vergleiche zwischen inkompatiblen Datentypen
- Mehrere Schemata pro Datensatz möglich (Datensatz dient mehreren Zwecken)
- Bietet praktische, handlungsorientierte Vorteile für Herausgeber
- Benutzer können Datensätze anhand ihrer Verwendbarkeit für bestimmte Anwendungen identifizieren

Qualitäts-Schema

Real-Time Anwendungen

- dcat:accessUrl (mandatory)
- dcat:endpointURL (mandatory)
- dcat:dataService (recommended)
- dct:conformsTo (recommended)
- dcat:endpointDescription (recommended)
- dcat:accessRights (recommended)
- dct:format (mandatory)
- dcat:temporalResolution (mandatory)
- dct:modificationDate (mandatory)
- dcat:endpointDescription (recommended)

Qualitäts-Schema

Geographische Anwendungen

- dct:spatial (mandatory)
- dcat:spatialResolutionInMeters (recommended)
- geodcatap:referenceSystem (recommended)
- geodcatap:spatialResolutionAsText (recommended)
- dcat:bbox (recommended)
- geodcatap:referenceSystem (recommended)

Schema-basiertes Qualitätsystem

Output 1: Deskriptive Statistik pro Qualitätsschema.

- Deskriptive Statistik: Validitätsraten pro Property (Format, Typ, URL, Bereich).

Output 2: Aggregation der Statistiken zu einem Score mit hierarchischer Struktur.

- Score: Hierarchische Aggregation entlang DCAT-Hierarchie (Catalog → Dataset → Distribution).

Deskriptive Statistik

Ziel der Statistik: **Neutraler Überblick über den statistischen Universum** des Katalogs; Hilft Providern, Verbesserungsbereiche zu identifizieren.

Validität: Property ist vorhanden und Wert entspricht Format/Datentyp (z. B. gut formatiertes URL, korrekter Typ, zulässiger Bereich).

- 56% der Distributionen haben eine gültige räumliche Property.
- 45% der Datasets haben eine gültige räumliche Property.
- 75% der Distributionen haben eine gültige Download-URL (Rein formal).

Aggregation der Statistik (Score)

Ziel: Einen einzelnen, vergleichbaren Score pro Katalog ableiten.

Grundprinzip: **Hierarchische Aggregation** der Deskriptivstatistiken gemäß den Properties eines Qualitätsschemas.

Aggregation der Statistik (Score)

1) Distribution-Scores: Anteil gültiger Distribution-Properties pro Distribution im Scheme Q.

2) Dataset-Scores: Anteil gültiger Dataset-Properties im Scheme Q.

3) Dataset-Overall-Score: (Summe aller Distribution-Scores + Dataset-Score) geteilt durch (Anzahl Distributionen + 1). Falls keine Distributionen, Dataset-Overall-Score = Dataset-Score.

4) Catalogue-Score: Mittelwert aller Dataset-Overall-Scores über alle Datasets.

Beispielformel: Dataset-Overall-Score = (60 + 60 + 75) / 3 = 65%.

Aggregation der Statistik (Score)

Vorgehen (Schritte):

1) Distribution-Scores: Anteil gültiger Distribution-Properties pro Distribution im Scheme Q.

2) Dataset-Scores: Anteil gültiger Dataset-Properties im Scheme Q.

3) Dataset-Overall-Score: (Summe aller Distribution-Scores + Dataset-Score) geteilt durch (Anzahl Distributionen + 1). Falls keine Distributionen, Dataset-Overall-Score = Dataset-Score.

4) Catalogue-Score: Mittelwert aller Dataset-Overall-Scores über alle Datasets.

Beispielformel: Dataset-Overall-Score = (60 + 60 + 75) / 3 = 65%.

Offene Fragen

Aggregationsoptionen: Durchschnitt, Median, gewichteter Durchschnitt.

Feedback zu den vorgeschlagenen Änderungen: Macht es Sinn für die Datenanbieter?

Priorisierte Anwendungsfälle: Welche Szenarien sind für Städte, Kantonen relevant?

Welche Properties sind für jedes Qualitätschema relevant?

Umgang mit Datensätzen ohne Distributionen: Führt das zu einer impliziten Strafe?

Weitere Schritte

- Detaillierte Spezifikation der Property-Listen pro Schema Q.
- Validierungsregeln und Testdaten/Beispiele erstellen.
- Prototyping des Score-Computations-Workflows und Metrik-Exportformate definieren.

Take aways

Was ist Qualität?

- Ohne Kontext ist Qualität ein abstraktes Konzept
- Objekte und Daten können unterschiedliche Qualitäten haben, die nicht aussagekräftig sind.
- Es kommt immer auf der Anwendungsfall, ob die Qualität gut oder schlecht ist.

Kann es zu viel werden?

- Übermäßige und unklare Standards können tatsächlich zu einem Hindernis werden.
- Ohne konkrete Anwendung, kann man eigentlich keine Standards setzen.
- Für DCAT-AP wird ein klares konzept für die hierarchische Agreggation gebraucht.

Kontakt

Dr. Juan Carlos Carvajal B.

Geschäftsbereich Digital Public Services

juan.carvajal@fokus.fraunhofer.de

Fraunhofer FOKUS

Kaiserin-Augusta-Allee 31

10589 Berlin

www.fokus.fraunhofer.de