
Adrian Wilke
Michael Röder, Kevin Dreßler, Daniel Vollmers
Prof. Dr. Axel Ngonga
dice-research.org
Universität Paderborn

15.12.2020

# OPAL
# Open Data Portal Germany

# Abschlusspräsentation

# Agenda

- Projekt: **9 Arbeitspakete** (AP) mit **39 Deliverables**
- Arbeitsplan, Arbeitsziele, Soll-IST-Zustand
  - **SOLL**: Kurze Beschreibung der AP
  - **IST**: Vorstellung der Projektergebnisse
- **Fragen**: jederzeit, bestenfalls nach Deliverables
  Markierung dazu in dieser Präsentation: ✓

## Zeitplan (flexibel)

- OPAL Portal Demo
- 09:15 – 10:30 Uhr
- 10:45 – 12:00 Uhr

## Gesamtziel des Vorhabens

"Das OPAL-Projekt konzipiert und entwickelt ein **Linked-Open-Data-Portal** mit Fokus auf **öffentlichen Datenquellen aus Deutschland** [...]."

## Arbeitspakete

1. Anforderungsanalyse und Architektur
2. Datenakquisition
3. Datenanalyse
4. Datenkonvertierung
5. Datenintegration
6. Datenselektion
7. Anwendungsfälle
8. Portalentwicklung
9. Projektmanagement

# Arbeitspaket 1:
# Anforderungsanalyse und Architektur

# Arbeitspaket 1:
# Anforderungsanalyse und Architektur

Ziel: "Erfassung der Anforderungen,
die Untersuchung & initiale Analyse der Datenbestände
der Entwurf einer Gesamtarchitektur für OPAL."

# Arbeitspaket 1:
# Anforderungsanalyse und Architektur

- D1.1 Anforderungsanalyse
- D1.2 Datenanalyse
- D1.3 Architektur

## D1.1 Anforderungsanalyse

- Analyse
  - Ergebnisse systematischer Literaturrecherche
  - Ergebnisse von Fragebögen
  - Ergebnisse mFUND Workshop
  - Erkenntnisse aus Fachkonzeption und Usability-Untersuchung des mCLOUD-Portals
- Deliverable als PDF-Datei
- Ergebnis: **21 konsolidierte Anforderungen**

## D1.1 Konsolidierte Anforderungen

1. Semantische Suche
2. Räumliche Suche
3. Zeitliche Suche
4. Zeitverlauf und Aktualisierungen
5. Komponentenbasiertes Metadatenportal
6. Programmatischer Zugriff
7. Konvertierung in Linked Data
8. (Meta-) Datenqualität
9. Automatisierte Verknüpfung von Datensätzen
10. Lizenzinformationen
11. Fokussierter Crawler

## D1.1 Konsolidierte Anforderungen

12. Automatisierte Extraktion von Metadaten
13. Selektion von Teilmengen eines Datensatzes
14. Mobile Anwendung / lokal relevante Datensätze
15. Question-Answering Assistent für soziale Netzwerke
16. Untersuchung von Datensätzen
17. Persistente versionierte Speicherung von Metadaten
18. Anzeige von existierender und neuer Daten
19. Empfehlungen von relevanten Datensätzen
20. Kommentierung bzgl. Qualität- und Inhalt
21. Bewertung der Datensätze

    *Deliverable abgeschlossen* ✓

# D1.2 Datenanalyse

- Technische und statistische Analyse
- Detaillierte Analyse zufällig ausgewählter Datensätze

| Datenquelle | Anzahl Datensätze |
|---|---|
| mCLOUD | 652 |
| MDM | 119 |
| GovData | 19.754 |
| OffeneDaten.de | 28.542 |
| European Data Portal | 817.755 (206.068 aus Deutschland) |

  - Deliverable als PDF-Datei
  - Heute: Informationen obsolet (folgt gleich)

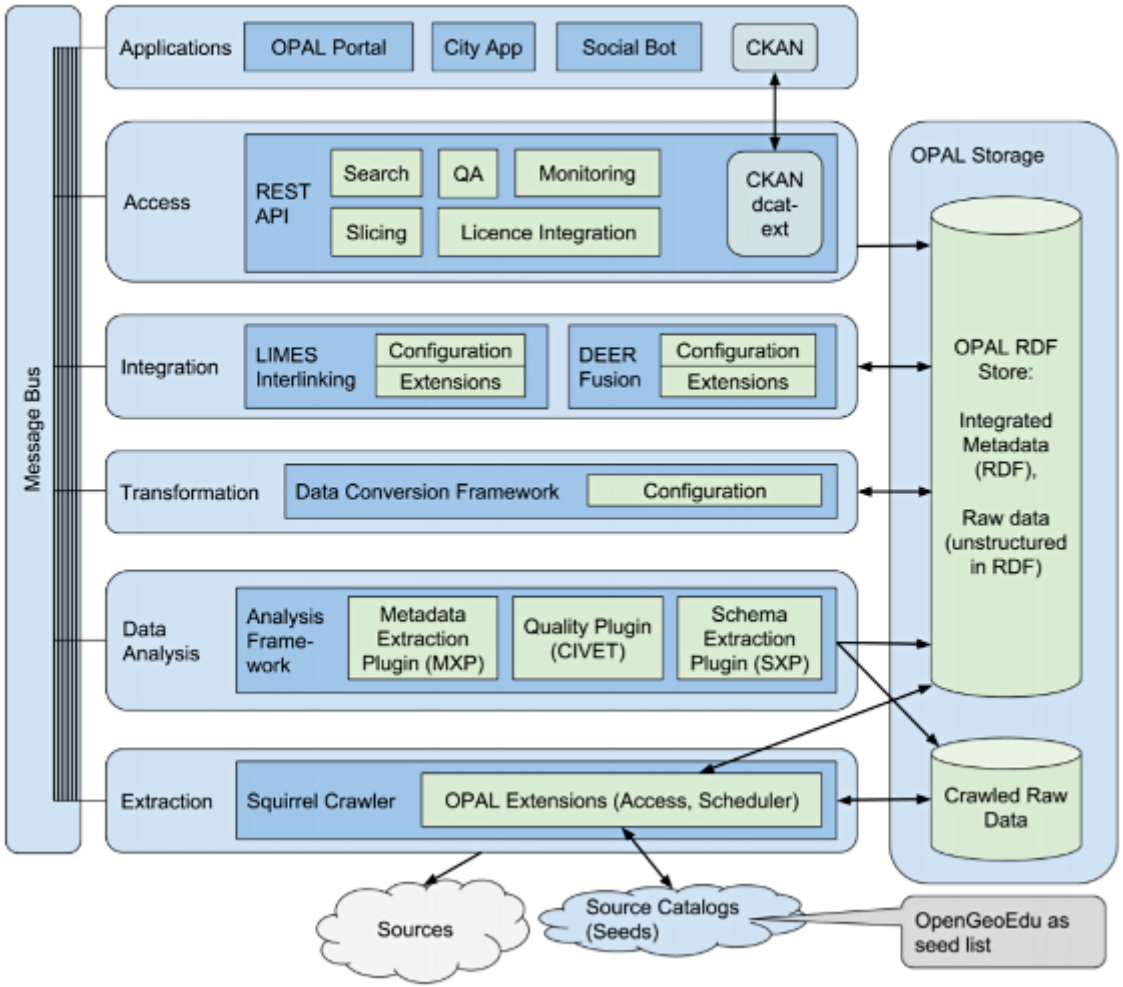
## D1.2 Datenanalyse

- **Datenvolumen** (2,5 Jahre: Mai 2018 – Dezember 2020)
  - mCLOUD: 652 → 3.276
  - EDP: 817.000 → 1.184.000
- Neuerungen / Obsoleszenz:
  - **jQuery** obsolet (MDM Relaunch) → HTML
  - **DCAT**-AP.de XML/RDF (mCLOUD 1.5.0, 11.04.19)
  - **Datenfluss**: mCLOUD → Govdata → EDP (mCLOUD 1.6.0, 16.07.19)

## D1.2 Datenanalyse

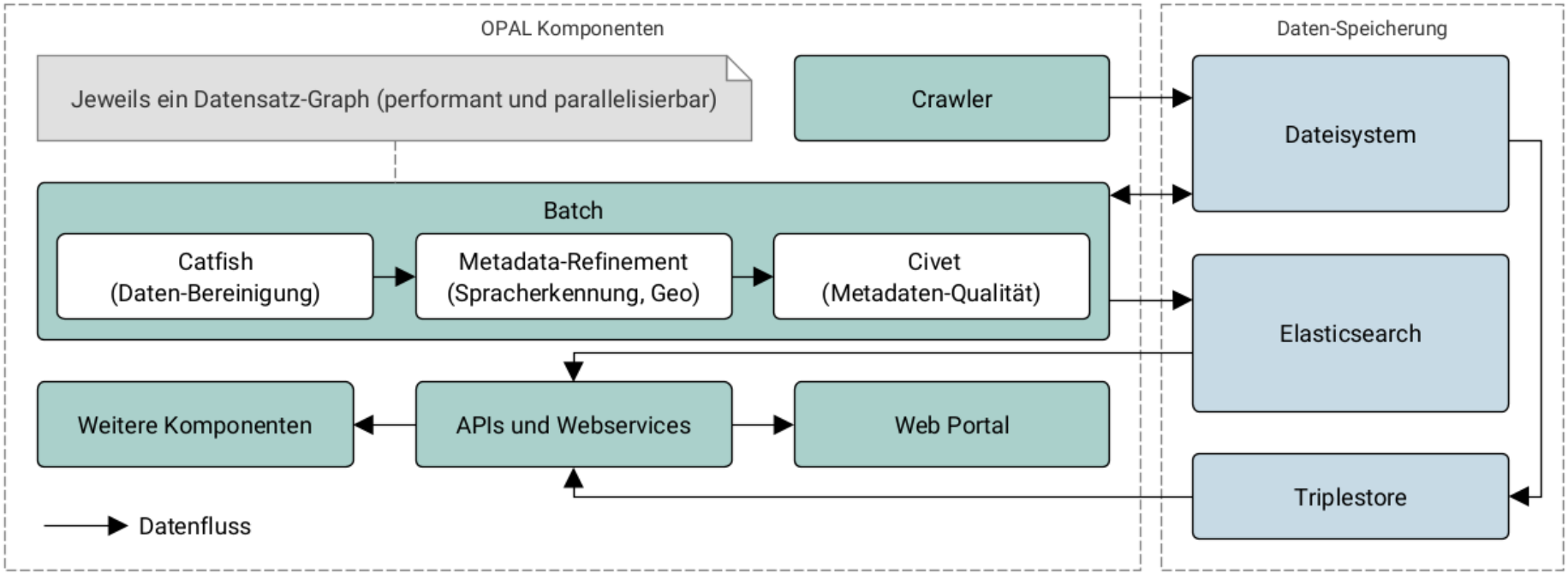
- OPAL Graph (Oktober/November 2020)
  - MDM: 203
  - mCLOUD: 2.853
  - GovData: 37.932
  - European Data Portal (EDP): 795.387
    (191.374 mit deutschen und englischen Titeln)
- OPAL Daten auf mCLOUD
  *Deliverable abgeschlossen ✓*

# D1.3 Architektur (2017/2018)




Deliverable als PDF-Datei

# D1.3 Architektur (Auszug 2020)




- Neu: Sequentielle Behandlung von Datensatz-Graphen
- Neu: Elasticsearch
- Code: OPAL Batch ✓

# Arbeitspaket 2:
# Datenakquisition

# Arbeitspaket 2:
# Datenakquisition

**Ziel**: "In Arbeitspaket 2 wird ein **fokussierter Crawler** entwickelt, der **Informationen** zu offenen Datensätzen aus Webseiten **extrahiert**."

## Arbeitspaket 2:
## Datenakquisition

- D2.1 Spezifikation der Crawler-Komponente
- D2.2 Erste Version der Crawler-Komponente
- D2.3 Benchmark-Spezifikation und Ergebnisse des ersten Crawlers
- D2.4 Metadatenbasierte Crawlingstrategien
- D2.5 Finale Crawler-Komponente
- D2.6 Finale Crawler-Benchmark-Ergebnisse

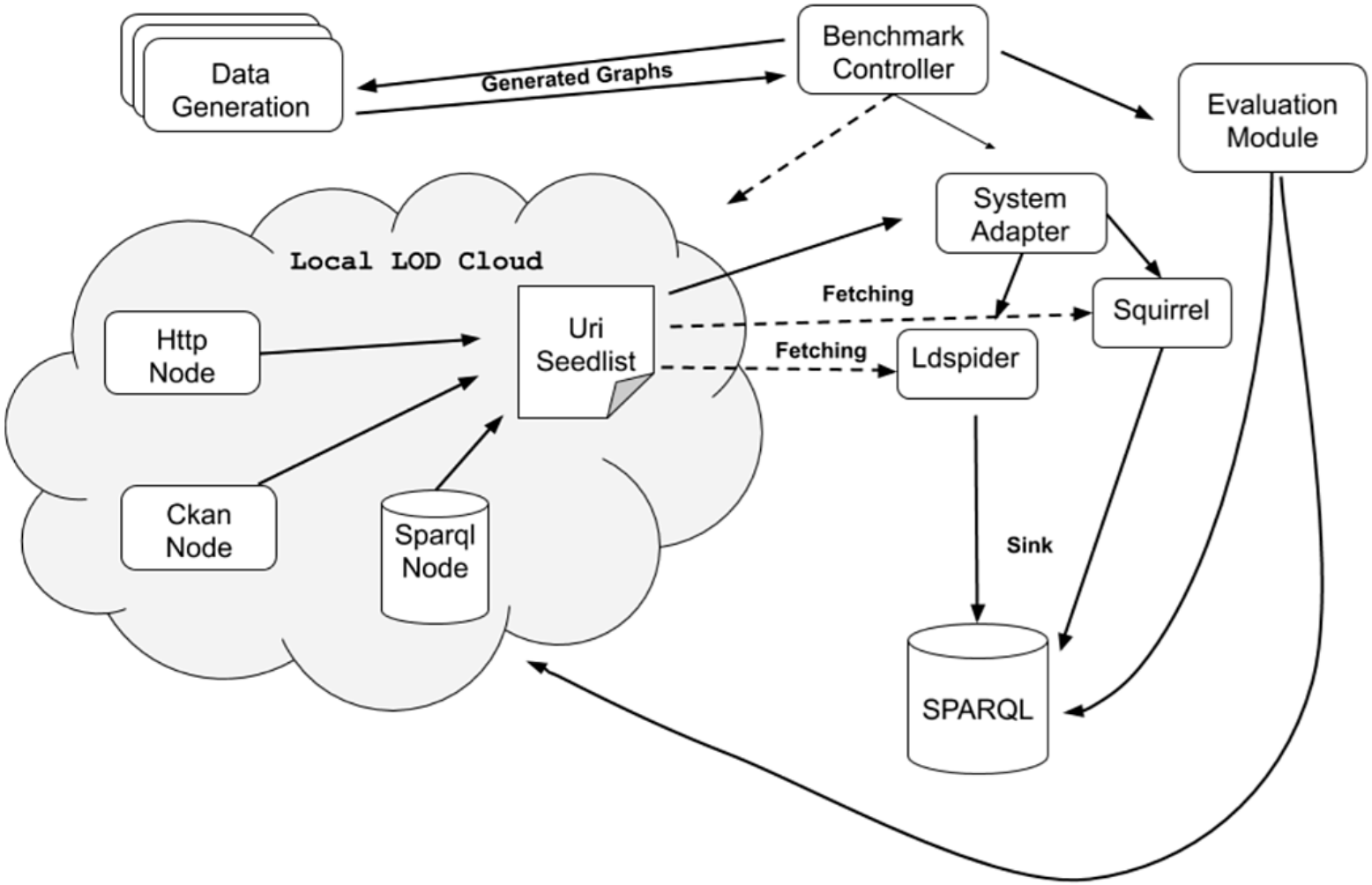
## D2.1 Spezifikation der Crawler-Komponente

- 9 funktionale Anforderungen
- 7 nicht-funktionale Anforderungen
- Schnittstellen, Operationen, Datenformate
- HTML, RDF, verschiedene Protokolle
- Vergleich von 10 Alternativen
- Deliverable als PDF-Datei ✓

# D2.2 Erste Version der Crawler-Komponente

Deliverable (Code): Squirrel, Version vom 13. Juni 2019 ✓

# D2.3 Benchmark-Spezifikation und Ergebnisse des ersten Crawlers



# D2.3 Benchmark-Spezifikation und Ergebnisse des ersten Crawlers

| | Squirrel | LdSpider |
|---|---|---|
| Average node graph degree | 4 | 4 |
| Average rdf graph degree | 4 | 4 |
| Number of Nodes | 5 | 5 |
| Triples evaluated | 5009 | 5009 |
| Triples per Node | 1000 | 1000 |
| HttpNode Weight | 1.0 | 1.0 |
| CkanNode Weight | 0 | 0 |
| SparqlNode Weight | 0 | 0 |
| **Recall** | **0.998** | **0.893** |

**Table 1. Http node only Scenario.**

| | Squirrel | LdSpider |
|---|---|---|
| Average node graph degree | 2 | 2 |
| Average rdf graph degree | 5 | 5 |
| Number of Nodes | 3 | 3 |
| Triples evaluated | 3003 | 3003 |
| Triples per Node | 1000 | 1000 |
| HttpNode Weight | 1.0 | 1.0 |
| CkanNode Weight | 1.0 | 1.0 |
| SparqlNode Weight | 1.0 | 1.0 |
| **Recall** | **1.0** | **0.5** |

**Table 2. Multi nodes run scenario.**

Deliverable als PDF-Datei ✓

# D2.4 Metadatenbasierte Crawlingstrategien

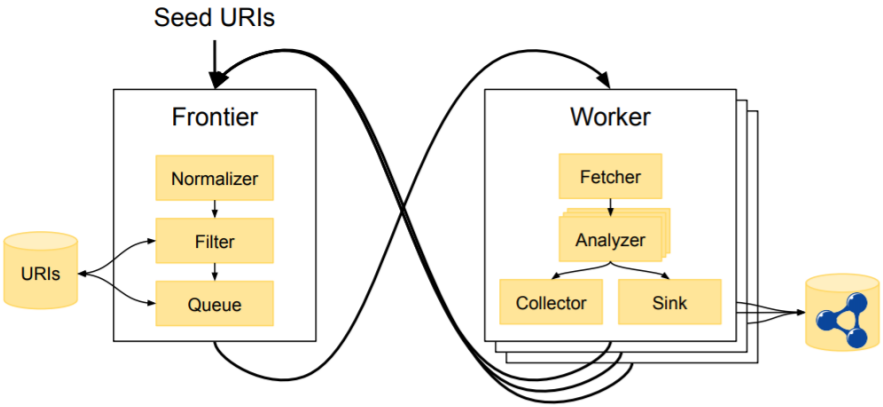



**Fig. 1.** Squirrel Core Achitecture

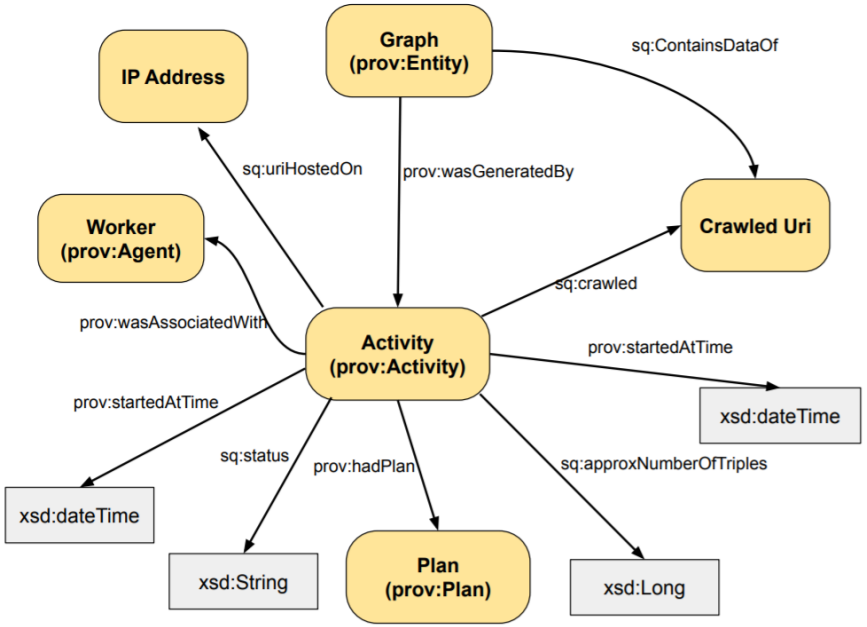



**Fig. 2.** Squirrel Activity, extending the PROV ontology

Deliverable als PDF-Datei ✓

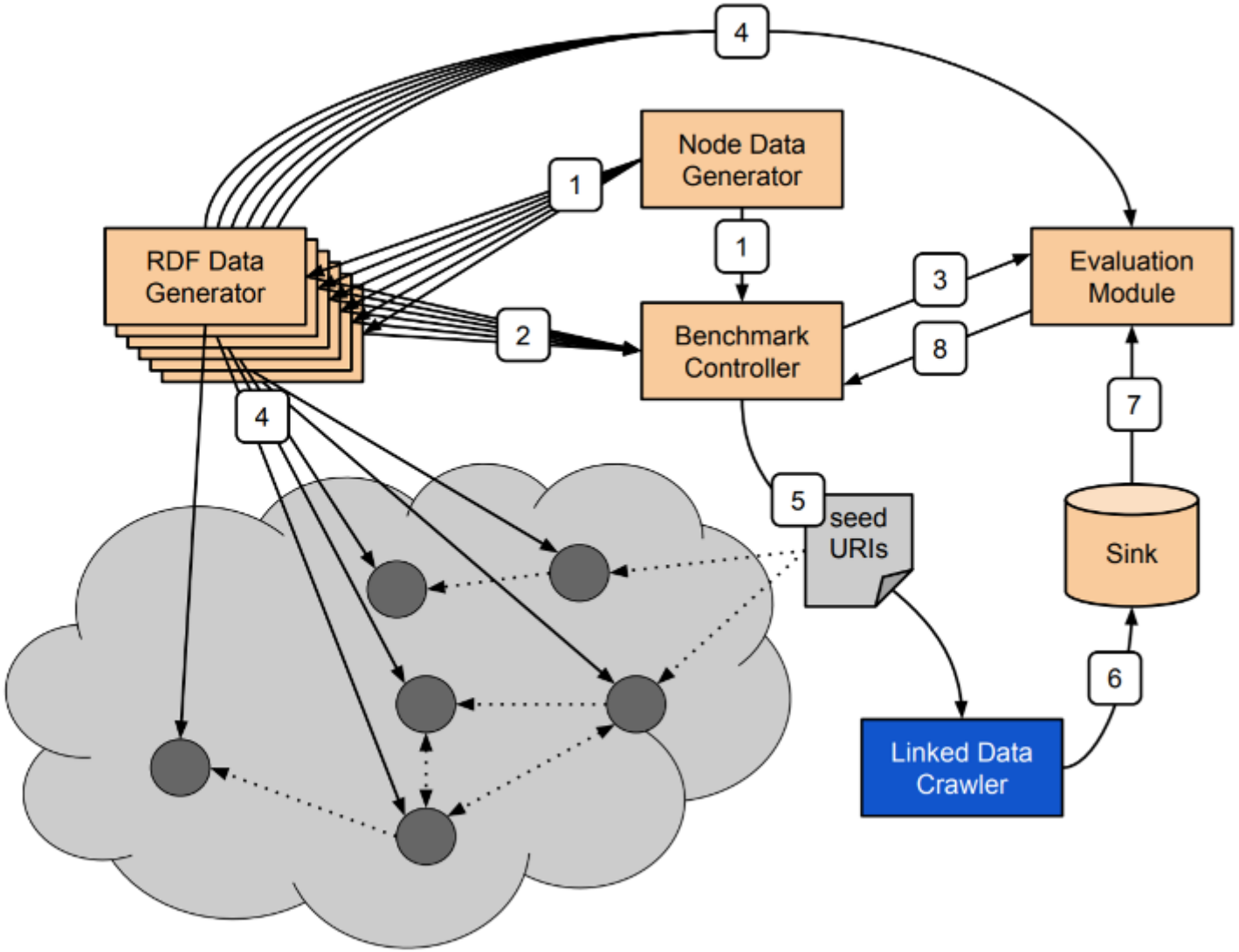
# D2.5 Finale Crawler-Komponente

Deliverable:

- Konfiguration: squirrel-portals-config
- Code: Squirrel, Version 0.3 ✓

# D2.6 Finale Crawler-Benchmark-Ergebnisse

## Evaluierung: Synthetischer Benchmark



# D2.6 Finale Crawler-Benchmark-Ergebnisse

| | RDF Serialisations | | | | | | | | | | Comp. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RDF/XML | RDF/JSON | Turtle | N-Triples | N-Quads | Notation 3 | JSON-LD | TriG | TriX | HDT | ZIP | Gzip | bzip2 |
| ORCA | ✓ | (✓) | ✓ | ✓ | (✓) | ✓ | (✓) | (✓) | (✓) | – | ✓ | ✓ | ✓ |
| LDSpider | ✓ | – | ✓ | X | ✓ | ✓ | ✓ | – | – | – | – | – | – |
| Squirrel | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

# D2.6 Finale Crawler-Benchmark-Ergebnisse

| Crawler | Data Web | | Efficiency | | | |
|---|---|---|---|---|---|---|
| | Micro Recall | Runtime (in s) | Micro Recall | Runtime (in s) | CPU (in s) | RAM (in GB) |
| LDSpider (T8) | 0.00 | 67 | – | – | – | – |
| LDSpider (T16) | 0.00 | 73 | – | – | – | – |
| LDSpider (T32) | 0.00 | 74 | – | – | – | – |
| LDSpider (T1,FS) | 0.31 | 1 798 | 1.00 | 2 031 | 320.0 | 1.2 |
| LDSpider (T8,FS) | 0.30 | 1 792 | 1.00 | 2 295 | 365.9 | 2.8 |
| LDSpider (T16,FS) | 0.31 | 1 858 | 1.00 | 1 945 | 345.4 | 1.6 |
| LDSpider (T32,FS) | 0.31 | 1 847 | 1.00 | 2 635 | 588.7 | 2.6 |
| LDSpider (T32,FS,LBS) | 0.03 | 66 | 0.54 | 765 | 182.1 | 7.5 |
| Squirrel (W1) | 0.98 | 6 663 | 1.00 | 11 821 | 991.3 | 3.9 |
| Squirrel (W3) | 0.98 | 2 686 | 1.00 | 4 100 | 681.4 | 8.6 |
| Squirrel (W9) | 0.98 | 1 412 | 1.00 | 1 591 | 464.8 | 18.1 |
| Squirrel (W18) | 0.97 | 1 551 | 1.00 | 1 091 | 279.8 | 22.1 |

Deliverable als PDF-Datei, Update 2020 ✓

# Arbeitspaket 3: Datenanalyse

## Arbeitspaket 3:
## Datenanalyse

**Ziel**: "Das dritte Arbeitspaket entwickelt Komponenten zur **Untersuchung und Gewinnung von Metadaten** der in AP2 gefundenen Information."

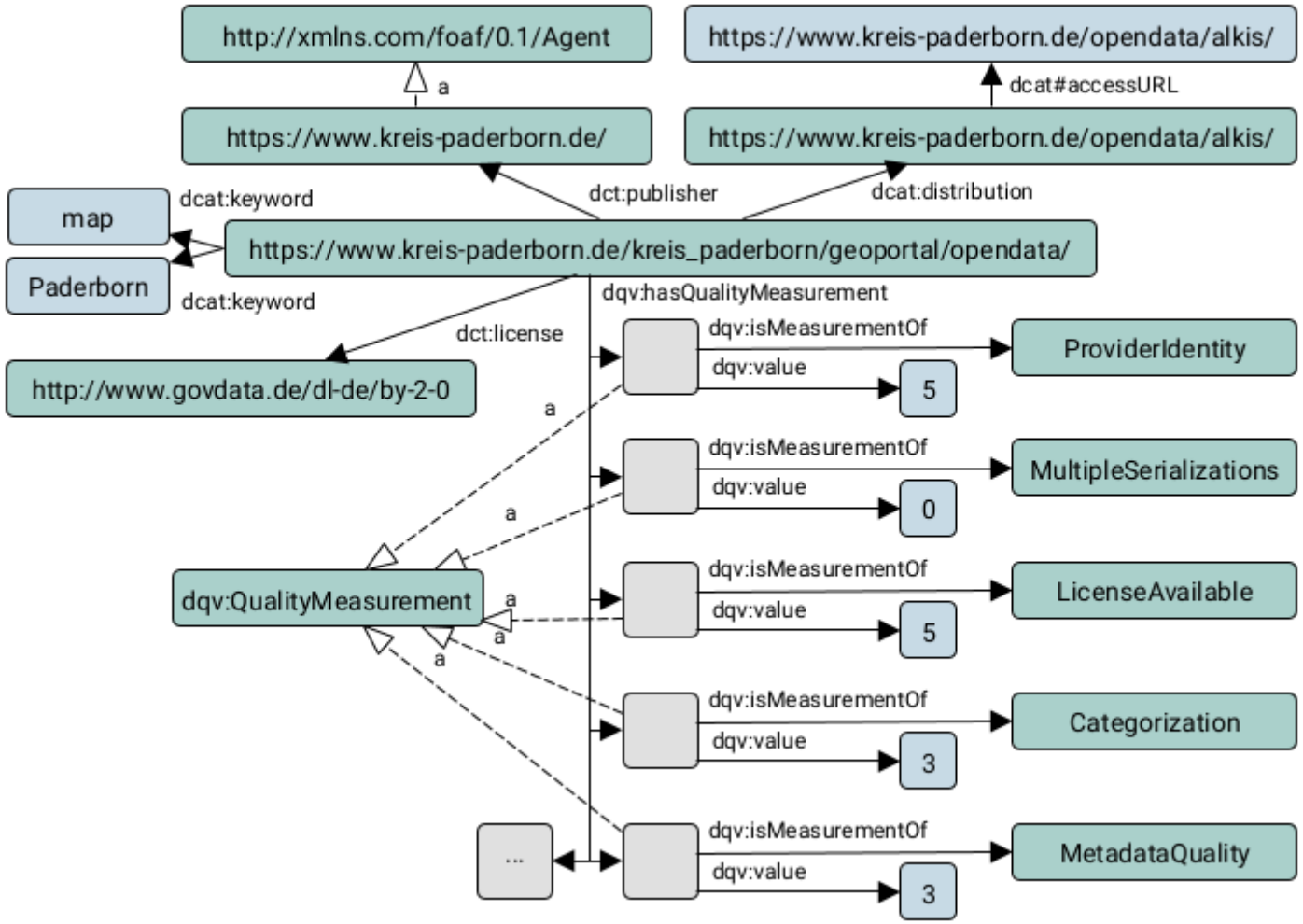
## Arbeitspaket 3:
## Datenanalyse

- D3.1 Spezifikation von Qualitätskriterien
- D3.2 Qualitätsanalyse-Komponente
- D3.3 Erste Metadatenextraktionskomponente
- D3.4 Topic-Extraktionskomponente
- D3.5 Finale Datenanalysekomponenten

## D3.1 Spezifikation von Qualitätskriterien

- Sichtung von wissenschaftlichen Artikeln
- Aggregation und Anpassung von Kriterien
- Ergebnis: Katalog, bestehend aus
  - 13 Qualitätsdimensionen und
  - 48 zugehörigen Qualitätskriterien und -metriken
- Deliverable als PDF-Datei ✓

# D3.2 Qualitätsanalyse-Komponente




- Vokabular in D4.1 erläutert
- Deliverable als PDF-Datei und Code (Civet) ✓

## D3.3 Erste Metadatenextraktionskomponente

- **Named Entity Recognition** mit FOX
  nutzt **Disambiguierung**: AGDISTIS / MAG (D4.4)
- **Spracherkennung** mit Apache OpenNLP
- Deliverables:
  - Code: Generation of RDF data statistics
  - Code: metadata-refinement (alpha)
  - Deliverable als PDF-Datei ✓

# D3.4 Topic-Extraktionskomponente

## Ansatz 1: Topic-Extraction

- Extrahiert Entitäten aus **englischsprachigen Texten** (z.B. **Topics**: **Ort**, **Datum**)
- Ergebnis:
  - date precision: 1.0
  - date recall: 0.087
  - date f1-score: 0.16
  - place precision: 1.0
  - place recall: 0.324
  - place f1-score: 0.489
- Problem: Viele Herausgeber → **heterogene Texte**
- Deliverable (Code): Topic-Extraction

# D3.4 Ansatz 2: Klassifizierung Kategorien

- Klassifizierung von **DCAT Kategorien** (themes)
- Nutzung von **Entscheidungsbäumen** und **TF-IDF**

The following accuracy was obtained for the cross-validation method with 4 folds:

| Classifier | 1-gram | 2-gram | 3-gram | 4-gram |
|---|---|---|---|---|
| J48 | 75,625% | 59,375% | 59,375% | 59,375% |
| NaiveBayes | 47,5% | 31,875% | 36,875% | 35% |

The following accuracy was obtained for the evaluation of the test data.

| Classifier | 1-gram | 2-gram | 3-gram | 4-gram |
|---|---|---|---|---|
| J48 | 62,07% | 50% | 59,09% | 55,32% |
| NaiveBayes | 28,09% | 29,35% | 27,59% | 28,05% |

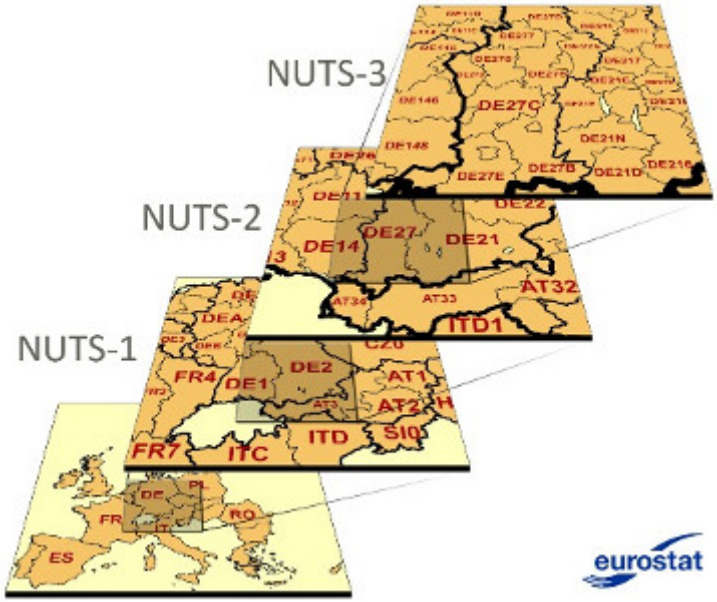
Deliverable (Code): Classification ✓

## D3.5 Finale Datenanalysekomponenten

- Qualitätsanalyse-Komponente Civet (D3.1, D3.2)
- Metadatenextraktions-Komponente (D3.3)
  - Erweiterung um **Geo-Daten** (D3.5, folgt gleich)
- Topic-Extraction und Klassifizierung (D3.4)

# D3.5 Finale Datenanalysekomponenten

## Metadatenextraktions-Komponente: LauNuts




- Neu: Metadatenextraktions-Komponente mit
  - Nomenclature of Territorial Units for Statistics (NUTS)
  - Local Administrative Units (LAU)
- **11.953 Orte** (DE/AU) mit Namen und Koordinaten
- Code: metadata-refinement ✓

# Arbeitspaket 4:
# Datenkonvertierung

## Arbeitspaket 4:
## Datenkonvertierung

**Ziel**: "Die extrahierten Metadaten werden in für Linked-Data-Anwendungen geeignete Formate **konvertiert**. Für die effiziente Durchsuchbarkeit werden geeignete **Indexstrukturen** entwickelt."

## Arbeitspaket 4:
## Datenkonvertierung

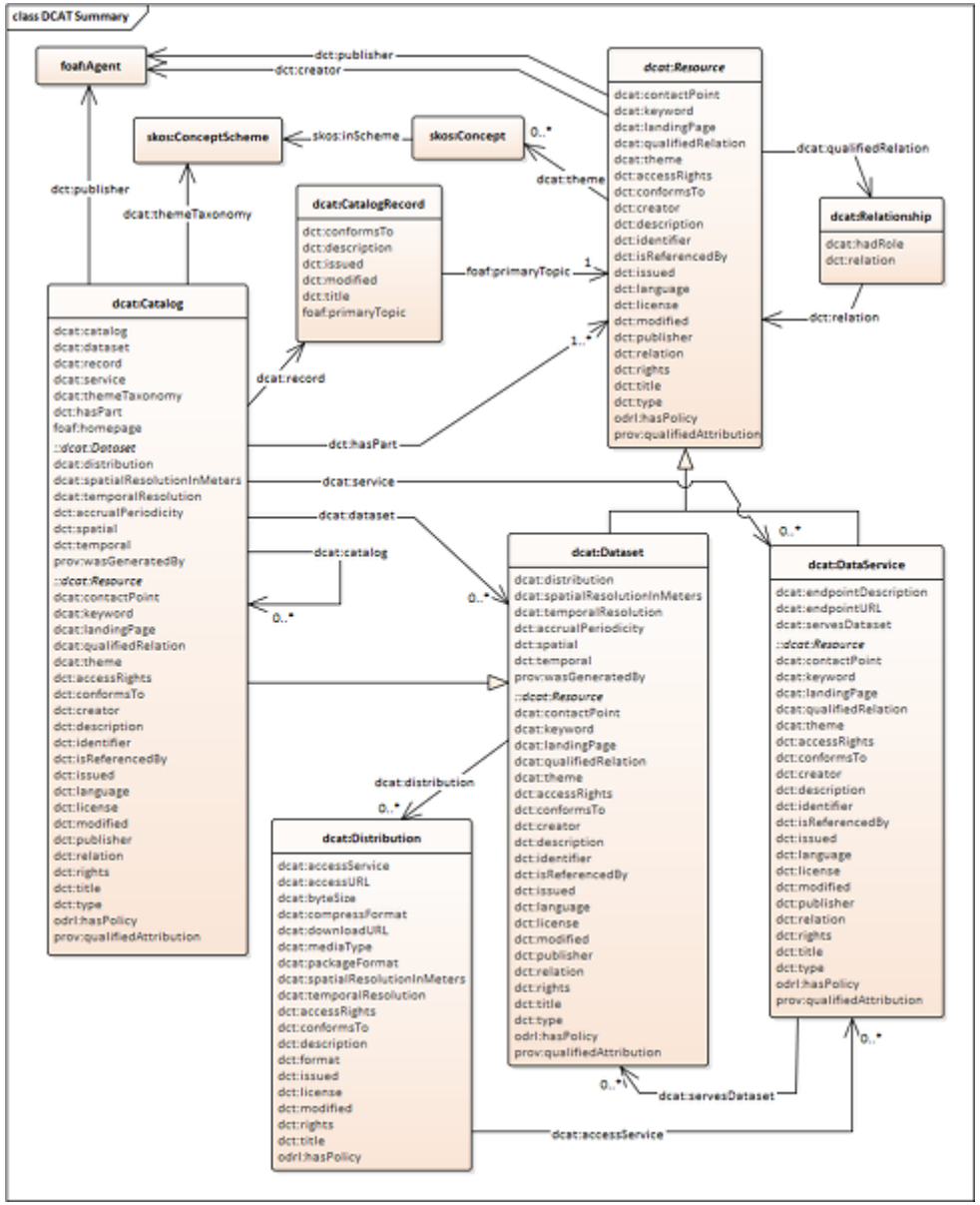
- D4.1 Vokabularspezifikation
- D4.2 Konvertierungskomponente
- D4.3 Prototyp Indexstrukturen und Entitätserkennung
- D4.4 Indizierungskomponente

## D4.1 Vokabularspezifikation

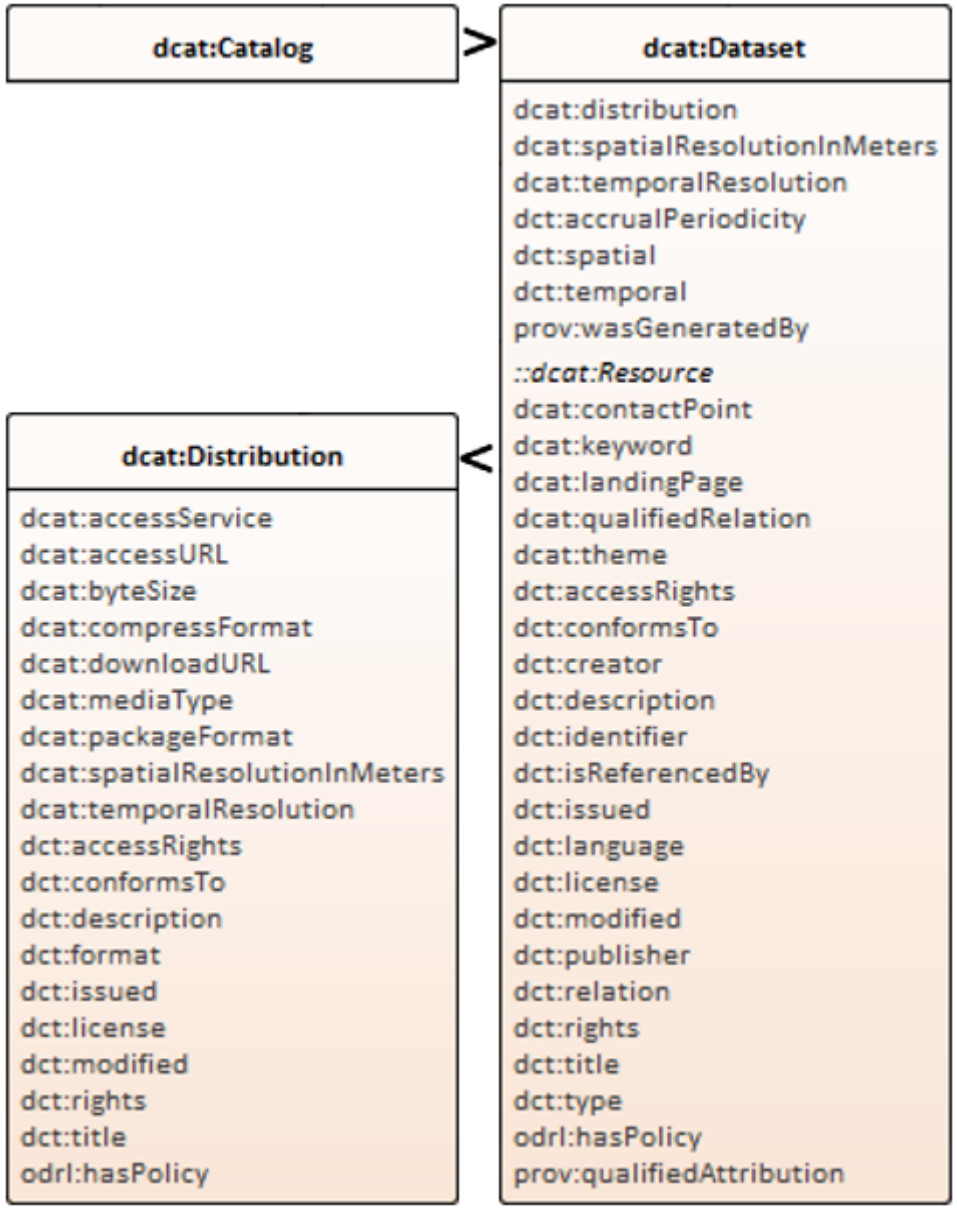
- Data Catalog Vocabulary (DCAT)
  Version 2, W3C Recommendation 04 February 2020
  www.w3.org/TR/vocab-dcat-2

- Data Quality Vocabulary (DQV)
  15 December 2016
  www.w3.org/TR/vocab-dqv

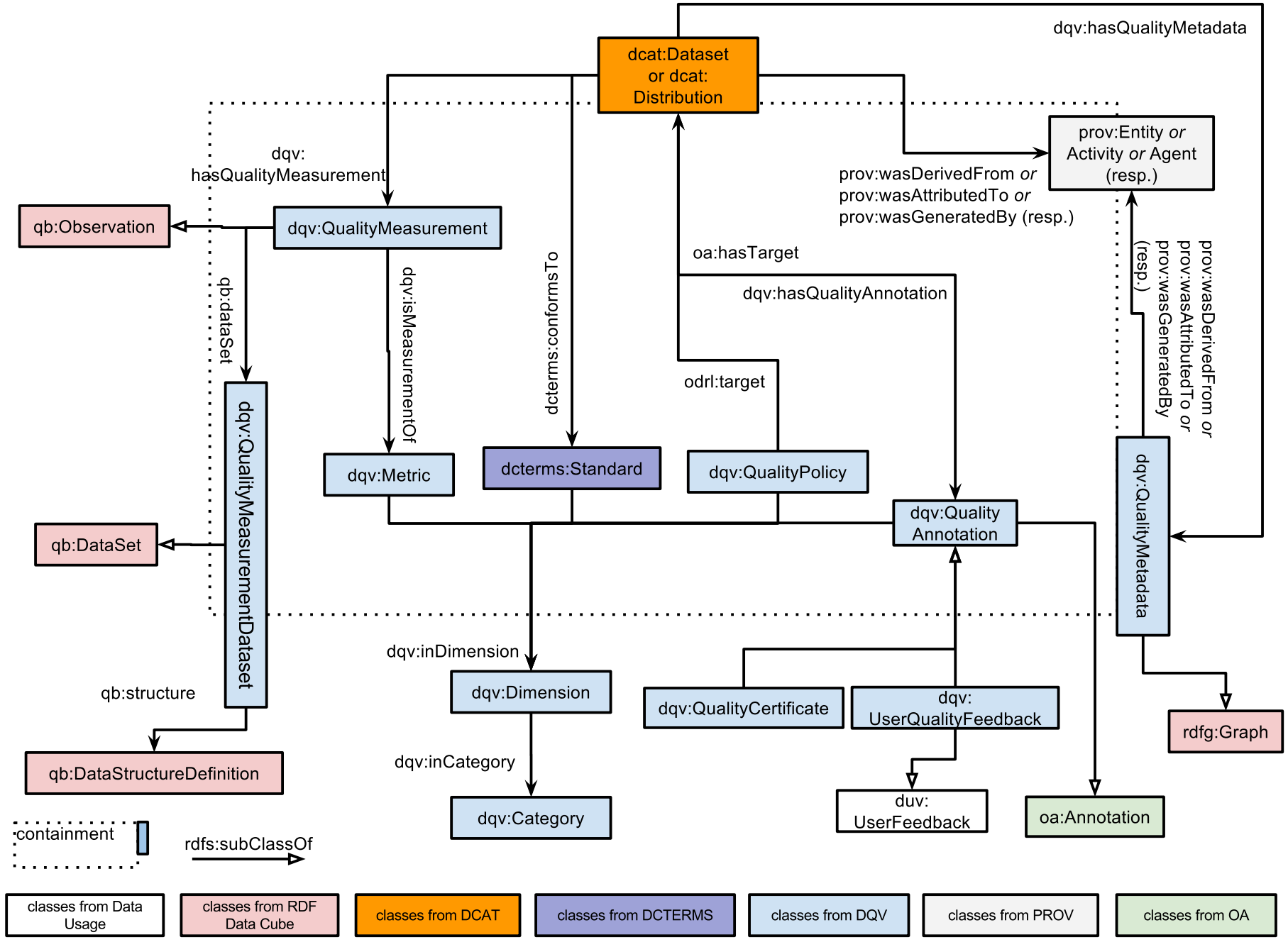
# D4.1 Vokabularspezifikation: DCAT



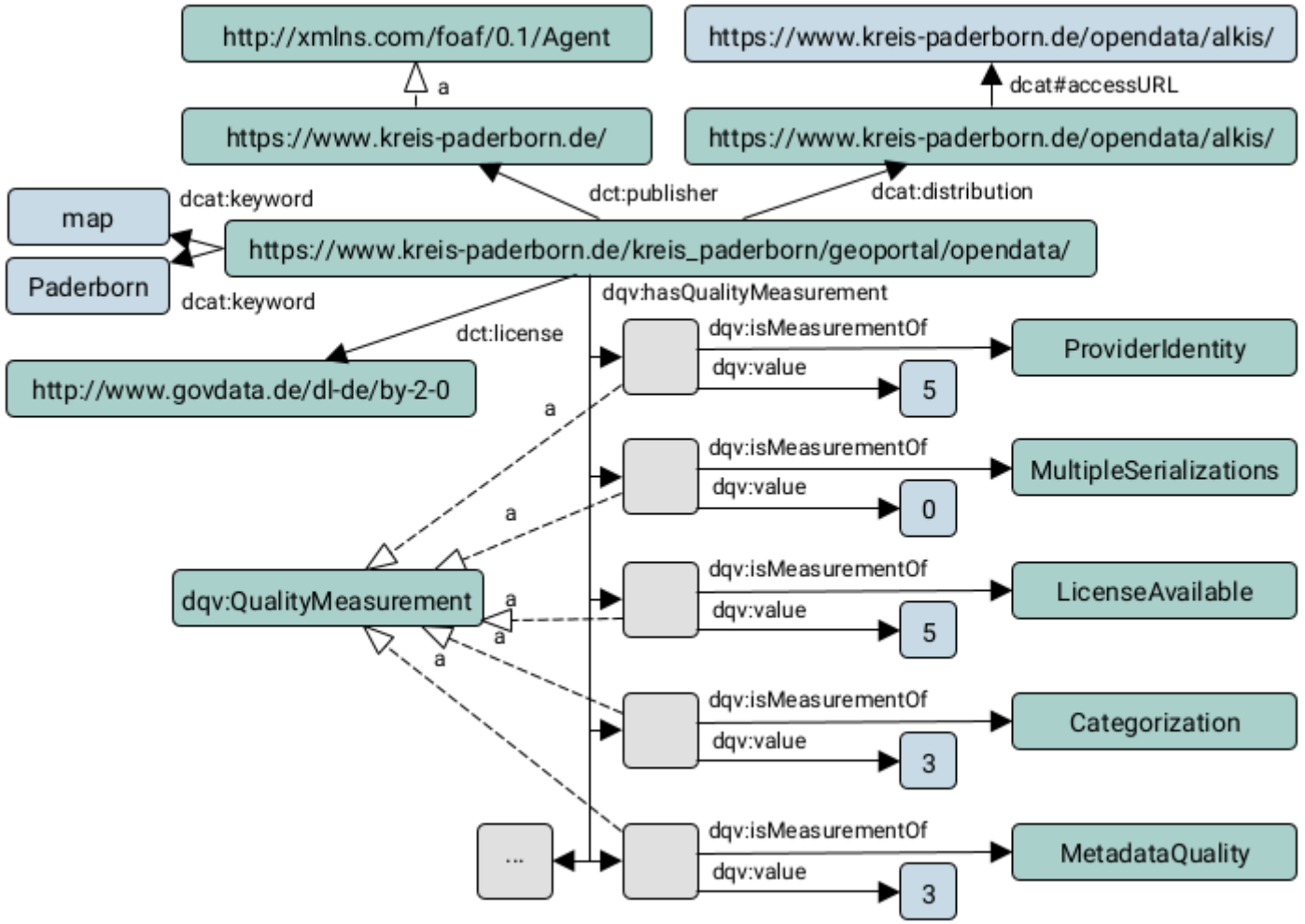
# D4.1 Vokabularspezifikation: DCAT



# D4.1 Vokabularspezifikation: DQV

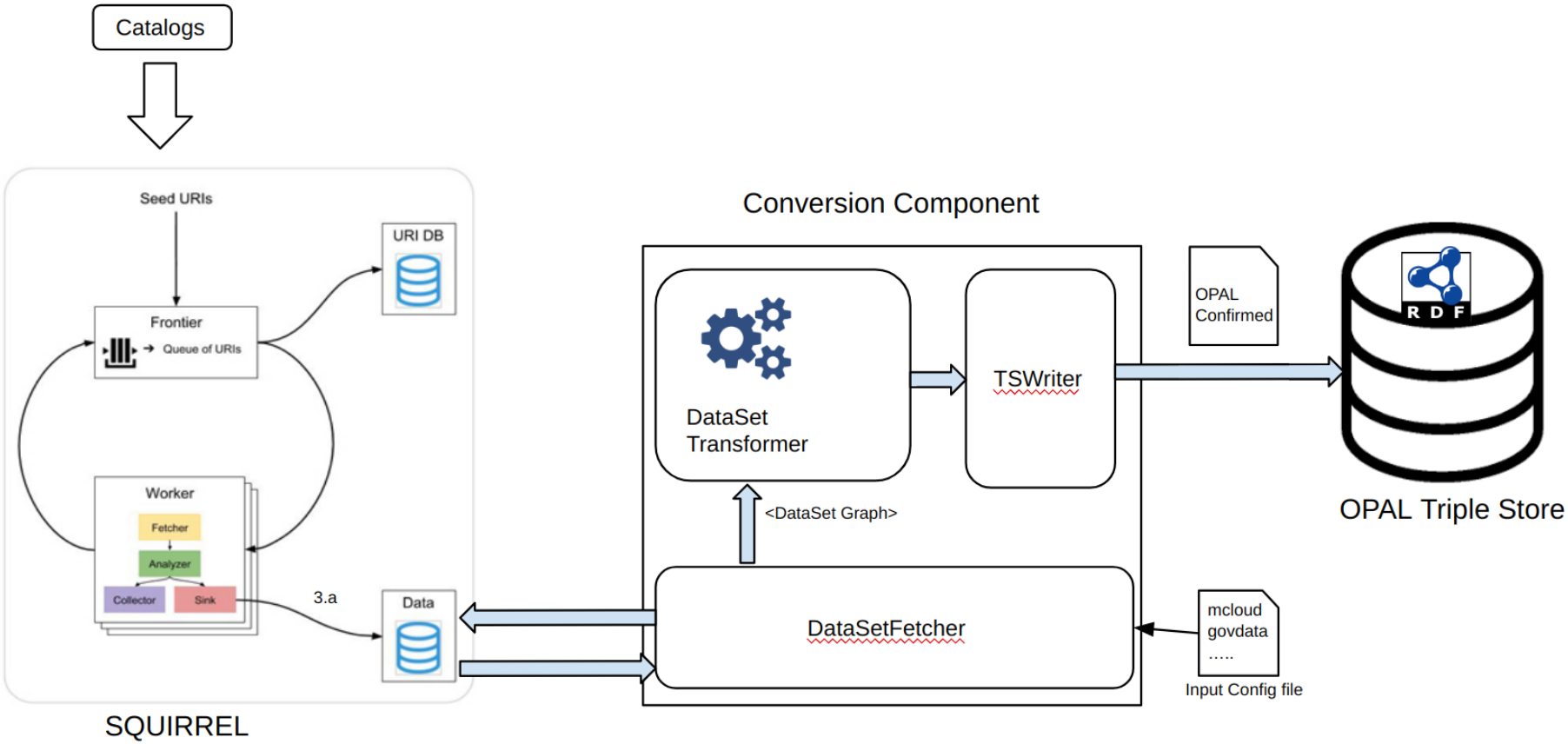

dqv:hasQualityMetadata
dcat:Dataset or dcat: Distribution
prov:Entity or Activity or Agent (resp.)
dqv: hasQualityMeasurement
prov:wasDerivedFrom or prov:wasAttributedTo or prov:wasGeneratedBy (resp.)
qb:Observation
dqv:QualityMeasurement
oa:hasTarget
dqv:hasQualityAnnotation
prov:wasDerivedFrom or prov:wasAttributedTo or prov:wasGeneratedBy (resp.)
qb:dataSet
dqv:isMeasurementOf
dcterms:conformsTo
odrl:target
dqv:QualityMeasurementDataset
dqv:Metric
dcterms:Standard
dqv:QualityPolicy
dqv:QualityMetadata
qb:DataSet
dqv:Quality Annotation
dqv:inDimension
qb:structure
dqv:Dimension
dqv:QualityCertificate
dqv: UserQualityFeedback
rdfg:Graph
qb:DataStructureDefinition
dqv:inCategory
duv: UserFeedback
oa:Annotation
dqv:Category
containment
rdfs:subClassOf
classes from Data Usage
classes from RDF Data Cube
classes from DCAT
classes from DCTERMS
classes from DQV
classes from PROV
classes from OA

# D4.1 Vokabularspezifikation: RDF Beispiel



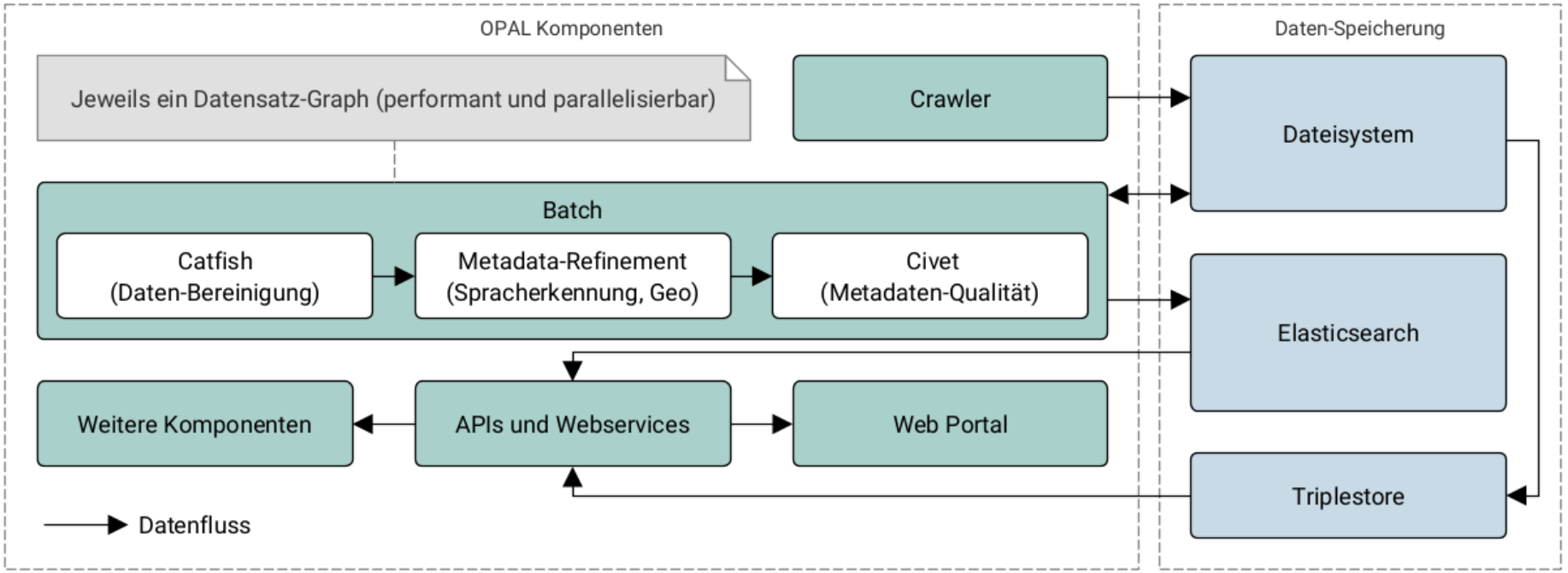

- Beispiel aus D3.2 Qualitätsanalyse-Komponente Civet
- Deliverable als PDF-Datei ✓

# D4.2 Konvertierungskomponente: Converter




- Publish-Subscribe Microservices via Spring Cloud
- Code: Converter
- Konfiguration: Converter-configuration

# D4.2 Konvertierungskomponente: Batch



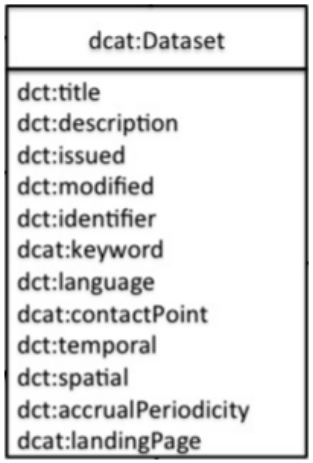

- Sequentielle Behandlung von Datensatz-Graphen
- Elasticsearch
- Code: OPAL Batch, OPAL Catfish
- Deliverable als PDF-Datei ✓

# D4.3 Prototyp Indexstrukturen und Entitätserkennung

## D4.3: Entitätserkennung

- D3.3 Metadatenextraktionskomponente
  - **Named Entity Recognition** (FOX)
- D3.4 Topic-Extraktionskomponente
  - **Topic**-Extraction (**Ort**, **Datum**)
  - Klassifizierung **DCAT Kategorien** (themes)
- D3.5 Metadatenextraktions-Komponente
  - LauNuts (**Geo Daten**)
- D4.4 Indizierungskomponente
  - **Disambiguierung** (AGDISTIS/MAG)

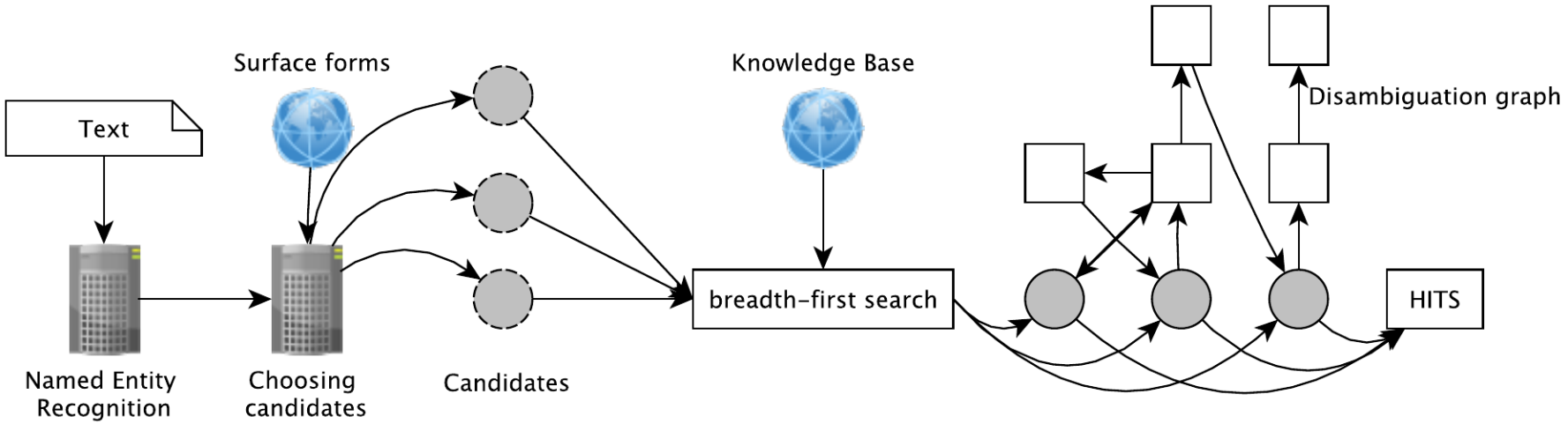
# D4.3: Indexstrukturen




```
"mappings": {
    "dynamic":"false",
     "properties": {
        "http://purl.org/dc/terms/title": {"type": "text"},
        "http://purl.org/dc/terms/description": {"type": "text"},
        "http://www.w3.org/ns/dcat#keyword": {"type": "text"},
        "http://purl.org/dc/terms/issued": {"type": "text"},
        "http://purl.org/dc/terms/modified": {"type": "text"},
        "http://purl.org/dc/terms/publisher":{"type":"keyword"},
        "http://www.w3.org/ns/dqv#hasQualityMeasurement":{"type":"keyword"},
        "http://www.w3.org/ns/dcat#distribution":{"type":"keyword"},
        "http://purl.org/dc/terms/accrualPeriodicity":{"type":"keyword"},
        "http://purl.org/dc/terms/spatial":{"type":"text"},
        "http://purl.org/dc/terms/identifier":{"type":"keyword"},
        "http://xmlns.com/foaf/0.1/isPrimaryTopicOf":{"type":"text"}
    }
}
```

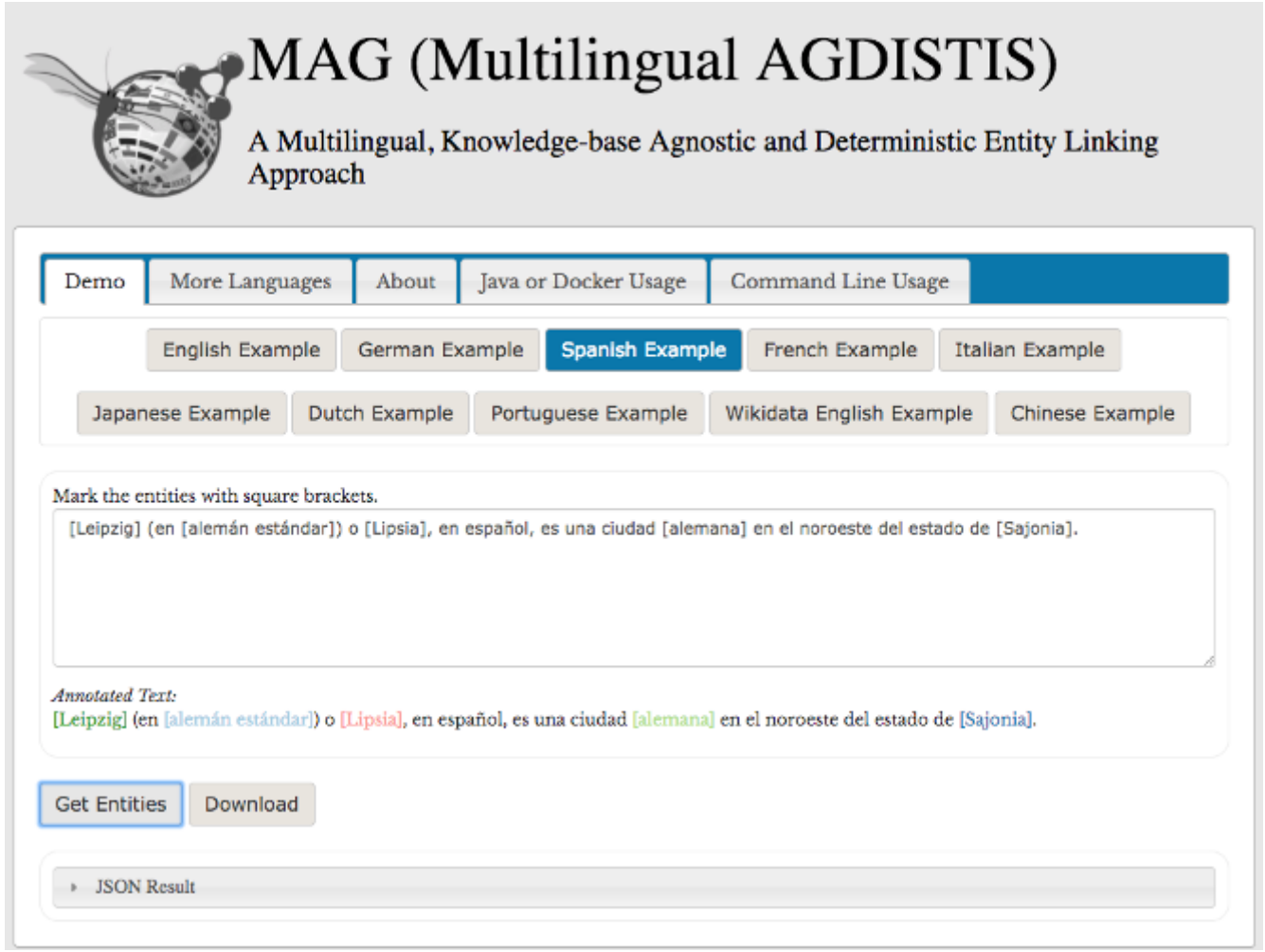
- RDF DCAT → Elasticsearch Mappings
- Deliverable als PDF-Datei ✓

# D4.4 Indizierungskomponente




- **Disambiguierung** / Linking Entitäten, mehrsprachig
- Graph basiertes Verfahren (HITS algorithmus)
- Erweitert in LIMBO: Elasticsearch
- Integration in **OPAL**: **geografische DB LauNuts**

# D4.4 Indizierungskomponente




Deliverable (Code): AGDISTIS ✓

# Arbeitspaket 5: Datenintegration

## Arbeitspaket 5:
## Datenintegration

**Ziel**: "Metadaten verschiedener Datensätze sollen **automatisiert verknüpft** werden, um miteinander **in Relation stehende Daten** zu erkennen. Damit soll OPAL es ermöglichen, dass Nutzer für komplexere Anwendungsfälle die dafür geeignete Menge an Datensätzen auffinden."

## Arbeitspaket 5:
## Datenintegration

- D5.1 Erste Version der Verknüpfungskomponente
- D5.2 Lernalgorithmen für Linkspezifikationen auf Metadaten
- D5.3 Lizenzintegrationskomponente
- D5.4 Erweiterte Lernalgorithmen für Linkspezifikationen auf Metadaten
- D5.5 Finale Verknüpfungskomponente

# D5.1 Erste Version der Verknüpfungskomponente


- Deliverable (Code): LIMES 1.5.0
- Im Folgenden verwendet

## D5.1 Verknüpfungskomponente: LIMES (1/3)

- **LIMES**: Link Discovery Framework for Metric Spaces
- **Überschneidungen** in versch. Wissensgraphen **finden**
  - z.B. **Erstellung Verknüpfungen** zwischen gleichen Ressourcen in verschiedenen Wissensgraphen
  - z.B. **geografische Ähnlichkeiten** finden
- Verwendung von **Metriken**, die Inhalte vergleichen
- Verwendet Dreiecksungleichung (aus der Geometrie / Mathematik).
- Verringert Anzahl benötigter Vergleiche.
- LIMES-Framework: Grundlage maschinellen Lernens

## D5.1 Verknüpfungskomponente: WOMBAT (2/3)

- WOMBAT: A Generalization Approach for Automatic Link Discovery
- Ansatz des Maschinellen Lernens (ML) zur Link Discovery
- **Supervised Learning**, benötigt ausschließlich positive Lernbeispielen
- **Unsupervised Learning**, ohne Lernbeispiele
- Algorithmus : Zwei aufeinander aufbauende Teile
  - Einzelner (atomic) Vergleich von Bestandteilen zweier Graphen
  - Kombinierung (generalization)

## D5.1 Verknüpfungskomponente: DRAGON (3/3)

- DRAGON: Decision Tree Learning for Link Discovery
- Verwendet **Entscheidungsbäume** zum Erlernen von Link Spezifikationen
- Beim rekursiven Erstellen von Entscheidungsbäumen kann dabei
  - entweder die Trefferquote (recall) unter Verwendung des lokalen Gini Index, oder
  - die Genauigkeit (precision) unter Verwendung des globalen F-Maß fokussiert werden
- Im Rahmen von OPAL entwickelt (Springer)
  u.a. auch: LimesWebUI ✓

## D5.2 Lernalgorithmen für Linkspezifikationen auf Metadaten

- 5 Experimente zur Deduplizierung
- Linkspezifikationen für OPAL/DCAT Datensätze
- jeweils 30 GB Arbeitsspeicher
- Deliverable (Code): OPAL Datenintegration
- Deliverable als PDF-Datei

## D5.2: WOMBAT unsupervised complete (1/5)

- Linkspezifikation: dcat:Dataset und
  23 optionale weitere Eigenschaften
- Anzahl erkannter Duplikate:
  - 0 (Schwellenwert: 0,95)
  - 0 (Schwellenwert: 0,9)
- Aufgrund Größe der Eingabedaten bei der Ausführung mit 30 GB Arbeitsspeicher vorzeitig beendet.
  java.lang.OutOfMemoryError: GC overhead limit exceeded

## D5.2: WOMBAT unsupervised simple II (2/5)

- Linkspezifikation: dcat:Dataset und
  23 optionale weitere Eigenschaften
- Anzahl erkannter Duplikate:
  - 0 (Schwellenwert: 0,9)
  - 23.744.536 (Schwellenwert: 0,8)
- keine Ergebnisse und
  zu viele Ergebnisse

## D5.2: WOMBAT unsupervised simple I (3/5)

- Linkspezifikation: dcat:Dataset, Distributionen dcat:downloadURL und **10 optionale** weitere Eigenschaften von
- Anzahl erkannter Duplikate:
  - 888 (Schwellenwert: 0,9)
  - 18.373 (Schwellenwert: : 0,5)
- **888 Duplikate von Datensätzen**

## D5.2: LIMES (Dataset) (4/5)

- Linkspezifikation: dcat:Dataset,
  Distributionen dcat:downloadURL und
  10 weitere optionale Eigenschaften von Distributionen
- Anzahl erkannter Duplikate: 888 Datensätze

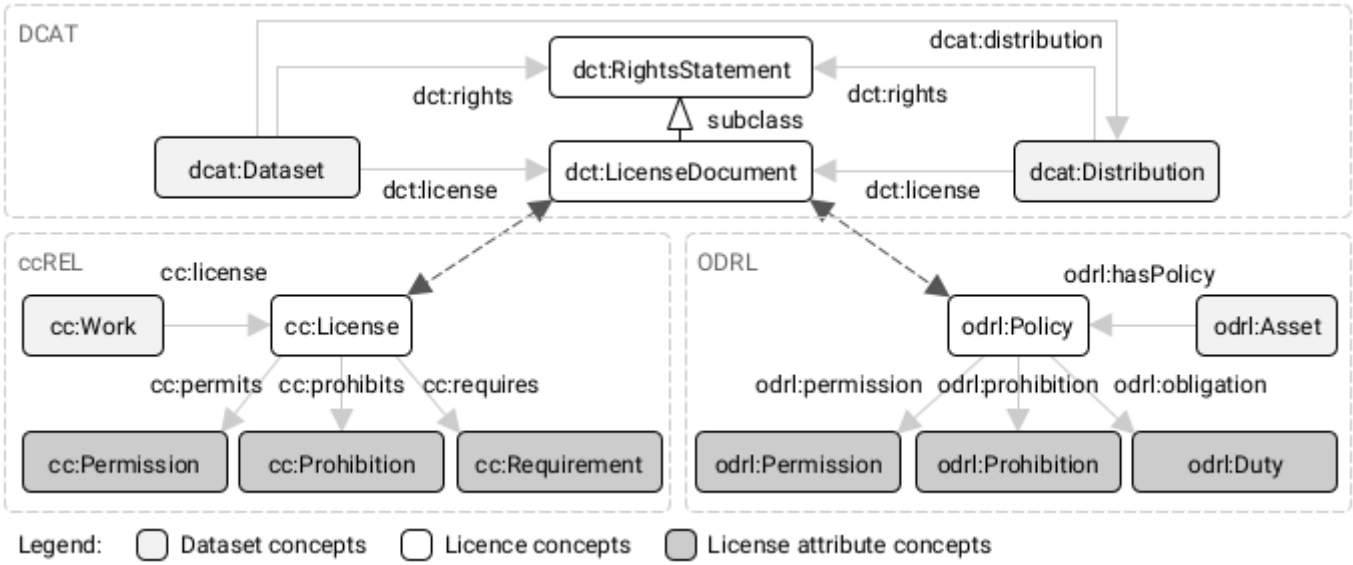
## D5.2: LIMES (Distribution) (5/5)

- Linkspezifikation: dcat:Distribution
  und dcat:downloadURL
- Anzahl erkannter Duplikate: 1.833 Distributionen ✓

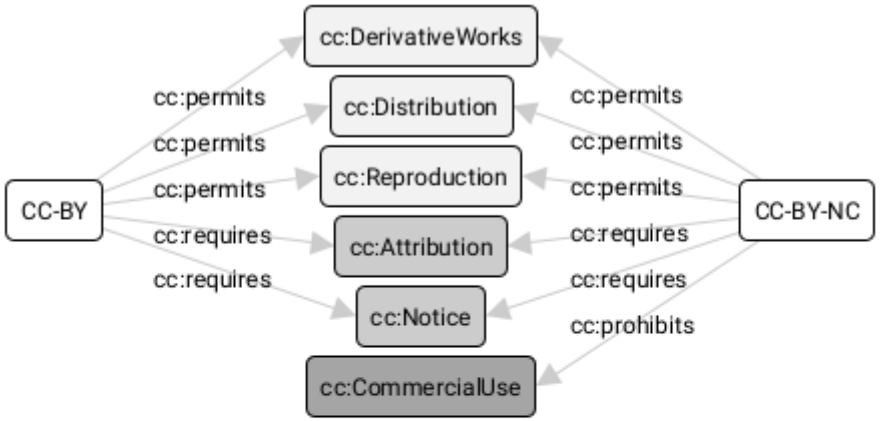
## D5.3 Lizenzintegrationskomponente

- Kombinierung von Datensätzen: Wahl aus erlaubten Lizenzen
- Deliverable (Code): licences

# D5.3 Lizenzen: Vokabulare



# D5.3 Lizenzen: Attribute




| | DerivativeWorks | Distribution | Reproduction | Attribution | Notice | ShareAlike | CommercialUse |
|---|---|---|---|---|---|---|---|
| CC-BY | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| CC-BY-NC | 0 | 0 | 0 | 1 | 1 | 0 | 1 |

# D5.3 Lizenzen: Kombination

| | DerivativeWorks | Distribution | Reproduction | Attribution | Notice | ShareAlike | CommercialUse |
|---|---|---|---|---|---|---|---|
| CC-BY | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| CC-BY-NC | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| ρ (CC-BY, CC-BY-NC) | 0 | 0 | 0 | 1 | 1 | 0 | **1** |

| | DerivativeWorks | Distribution | Reproduction | Attribution | Notice | ShareAlike | CommercialUse | |
|---|---|---|---|---|---|---|---|---|
| ρ (CC-BY, CC-BY-NC) = b | 0 | 0 | 0 | 1 | 1 | 0 | 1 | |
| CC-BY = $b^1$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | ✗ |
| CC-BY-NC = $b^2$ | 0 | 0 | 0 | 1 | 1 | 0 | 1 | ✓ |
| CC-0 = $b^3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ✗ |

# D5.3 Lizenzen: Evaluierung

| | PUBLIC DOMAIN | PUBLIC DOMAIN | BY | BY SA | BY NC | BY ND | BY NC SA | BY NC ND |
|---|---|---|---|---|---|---|---|---|
| PUBLIC DOMAIN | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| PUBLIC DOMAIN | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| BY | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| BY SA | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| BY NC | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| BY ND | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| BY NC SA | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| BY NC ND | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |

"Choose two works you wish to **combine or remix**. [...]
Use at least the **most restrictive licensing** of the two
(use the license **most to right or down state**)
for the new work." CC wiki

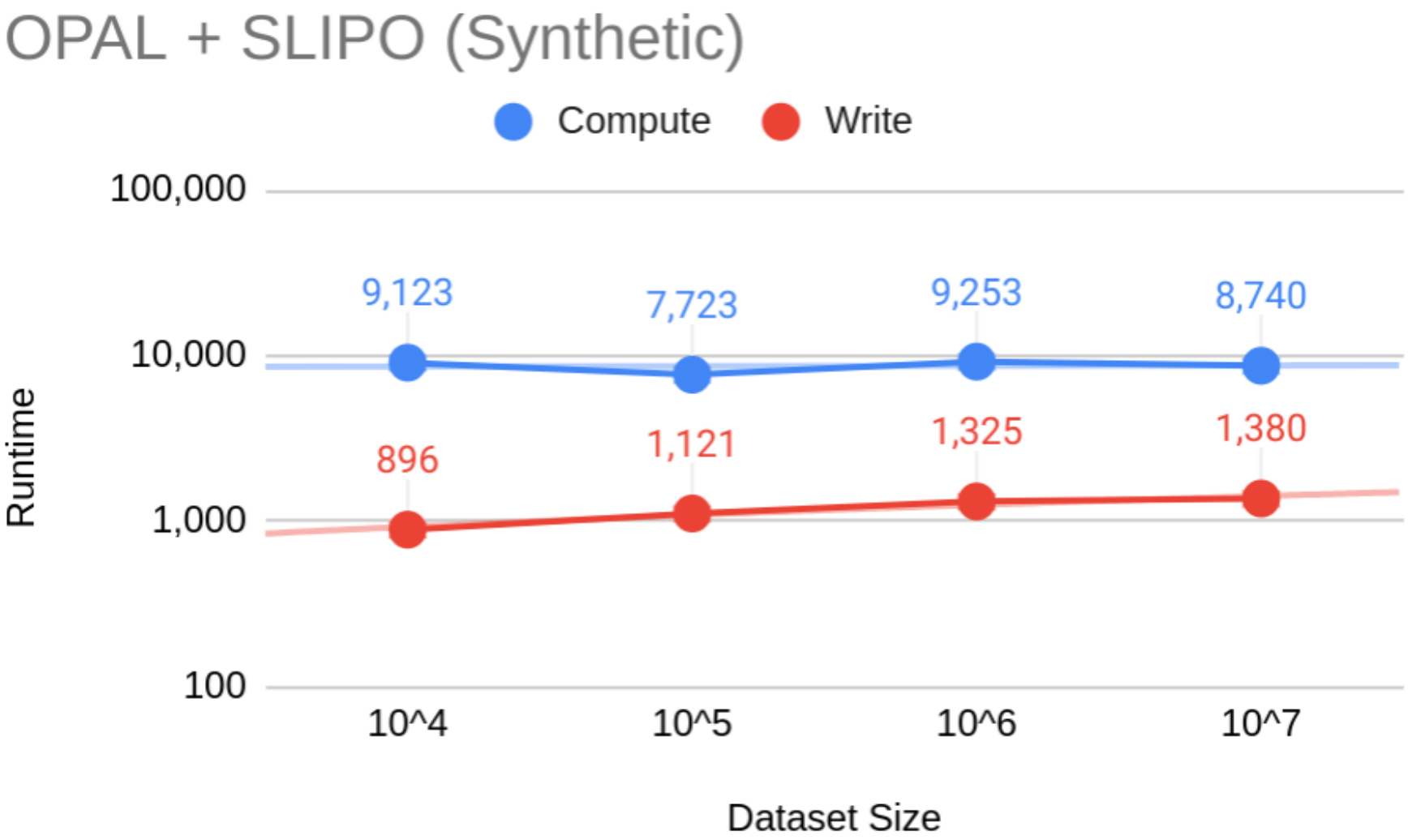
# D5.3 Lizenzen: Evaluierung

| | PD | CC0 | BY | BY-SA | BY-NC | BY-NC-SA |
|---|---|---|---|---|---|---|
| PD | all | all | BY, BY-NC, BY-NC-ND, BY-NC-SA, BY-ND, BY-SA | BY-SA | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC-SA |
| CC0 | all | all | BY, BY-NC, BY-NC-ND, BY-NC-SA, BY-ND, BY-SA | BY-SA | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC-SA |
| BY | BY, BY-NC, BY-NC-ND, BY-NC-SA, BY-ND, BY-SA | BY, BY-NC, BY-NC-ND, BY-NC-SA, BY-ND, BY-SA | BY, BY-NC, BY-NC-ND, BY-NC-SA, BY-ND, BY-SA | BY-SA | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC-SA |
| BY-SA | BY-SA | BY-SA | BY-SA | BY-SA | - | - |
| BY-NC | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC, BY-NC-ND, BY-NC-SA | - | BY-NC, BY-NC-ND, BY-NC-SA | BY-NC-SA |
| BY-NC-SA | BY-NC-SA | BY-NC-SA | BY-NC-SA | - | BY-NC-SA | BY-NC-SA |

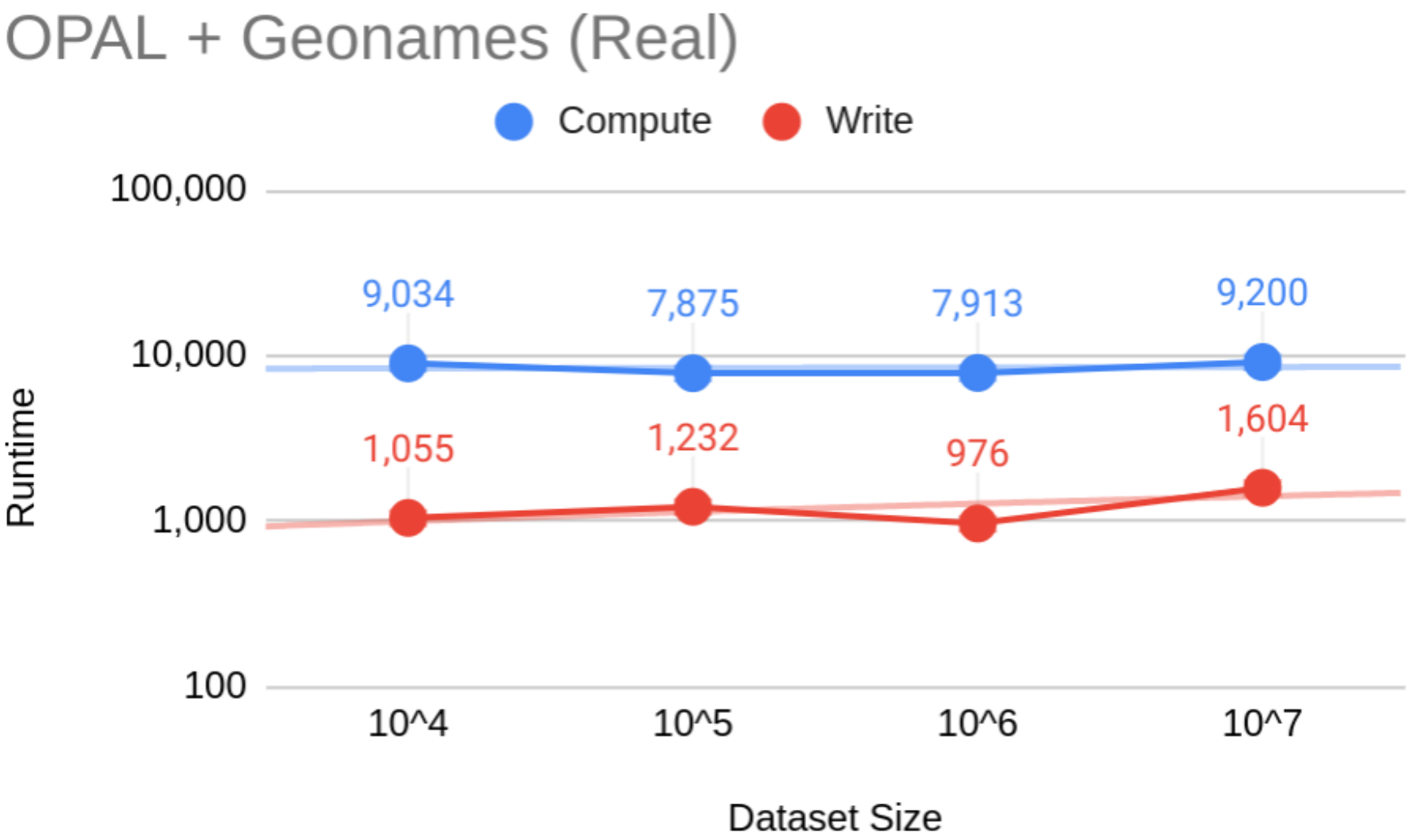
- Auflistung aller kompatiblen Lizenzen zur Re-Lizensierung von kombinierten Datensätzen.
- Als **Artikel** in IEEE International Conference on Semantic Computing (**ICSC 2021**) Resource Track ✓

## D5.4 Erweiterte Lernalgorithmen für Linkspezifikationen auf Metadaten

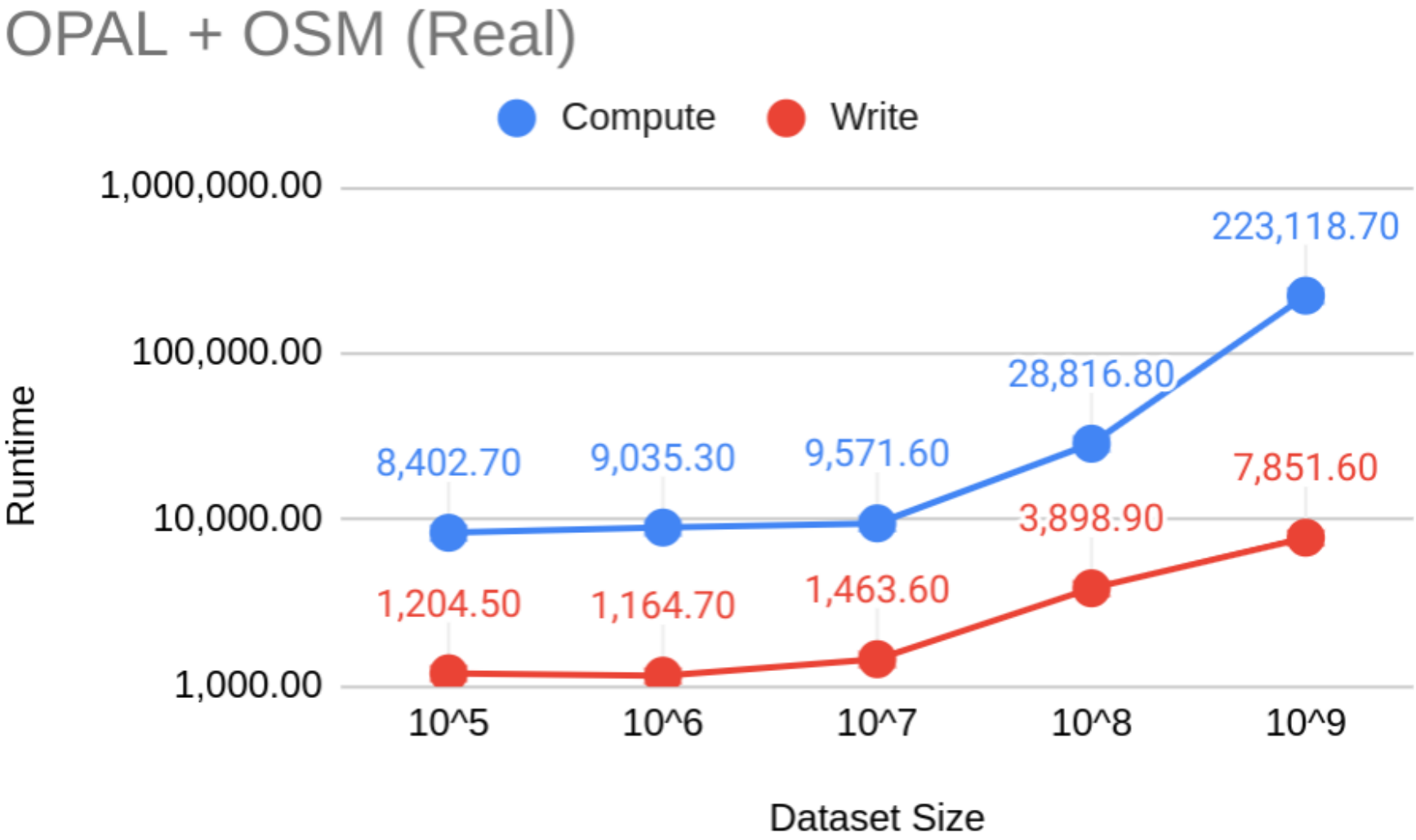
- Daten:
  - OPAL LauNuts: 84.000 Koordinaten-Punkte
  - LinkedGeoData.org (OpenStreetMap als RDF)
- ORCHID - Reduction-Ratio-Optimal Computation of Geo-Spatial Distances for Link Discovery PDF Springer


OPAL + SLIPO (Synthetic)
Compute
Write
100,000
10,000
1,000
100
Runtime
9,123
7,723
9,253
8,740
896
1,121
1,325
1,380
10^4
10^5
10^6
10^7
Dataset Size

# D5.4: Geonames



# D5.4: OSM




Deliverable (Code): LIMES Spark ✓

# D5.5 Finale Verknüpfungskomponente


- RDF Dataset Enrichment Framework: DEER
- Deliverable (Code): LIMES 1.7.4 ✓

# Arbeitspaket 6:
# Datenselektion

## Arbeitspaket 6:
## Datenselektion

Ziel: "**Relevante Teile** eines Datensatzes lassen sich anhand von Prädikaten und Relationen sowie **räumlichen Relationen** auswählen, um den Umfang der übertragenen Daten möglichst zu minimieren."

# Arbeitspaket 6:
# Datenselektion

- D6.1 Linked-Data-Slicing-Komponente
- D6.2 Räumliches Slicing

# D6.1 Linked-Data-Slicing-Komponente

## Ansatz 1: ElasticTriples

- Import: **90 Millionen Triple** in **77 Minuten**
  (16.3 GB im N-Triples Format)
- Splitting: Eine Anfrage dauert **2-3 Sekunden**.
  Beispiel: 1 aus 1 Millionen **DCAT Datensätzen**
  mit **206 zugehörigen Triplen**
  nutzt intern 2.281 Anfragen
- Deliverable (Code): ElasticTriples - Elasticsearch powered triple storage

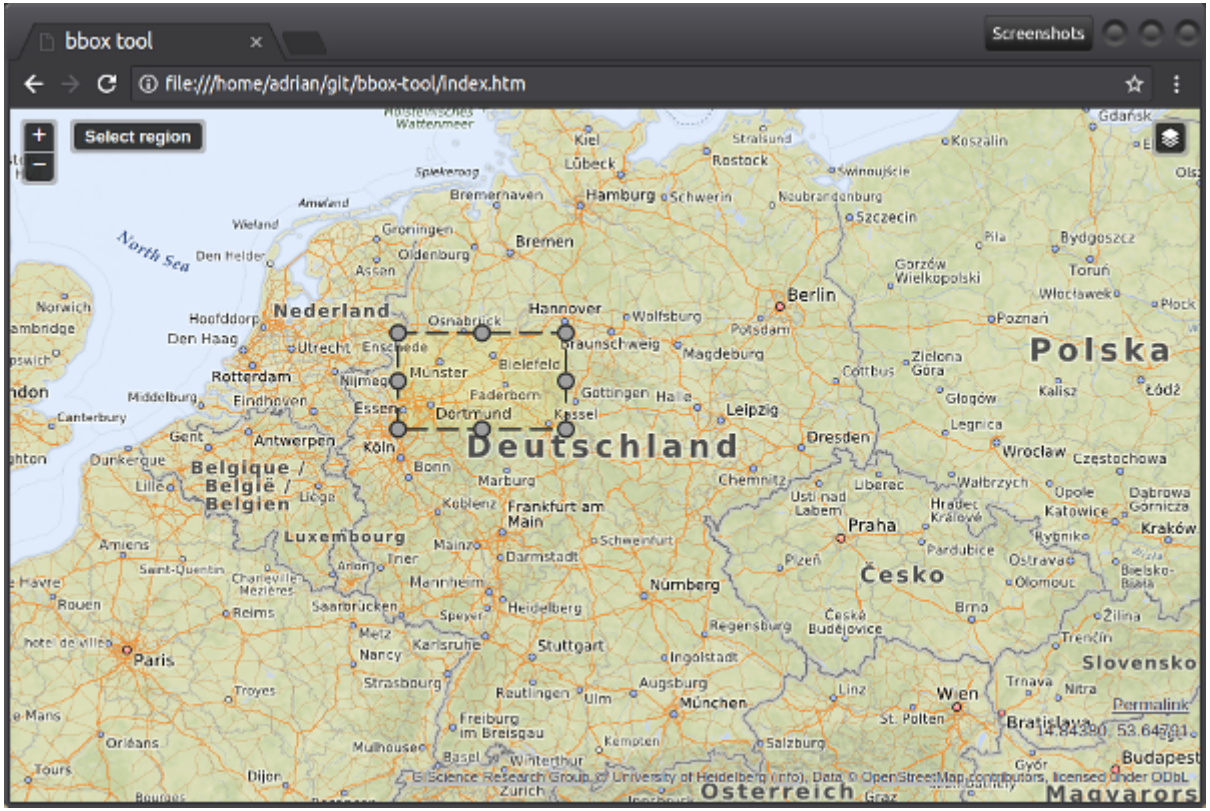
## D6.1 Ansatz 2: OPAL Slicer

- Nutzt **Muster im SPARQL Format**
  um Untermengen von **Wissensgraphen zu extrahieren**.
- Basiert auf RDFSlice Projekt
- Beispiel:

```
-source input.ttl
-patterns "Select * where {?d a <http://www.w3.org/ns/dcat#Dataset>}"
-out datasets.ttl
```

Deliverable (Code): OPAL Slicer ✓

# D6.2 Räumliches Slicing




- Integriert in OPAL Portal Demo
- Deliverable (Code): Spatial Slicing ✓

# Arbeitspaket 7: Anwendungsfälle

## Arbeitspaket 7:
## Anwendungsfälle

**Ziel**: "Die **Anwendbarkeit** des Linked-Data-Ansatzes zur Extraktion und Verwaltung von Metadaten offener Datensätze soll anhand der **Suchfunktion** als zentraler Komponente eines Datenportals sowie weiterer **Demonstratoren** validiert werden."

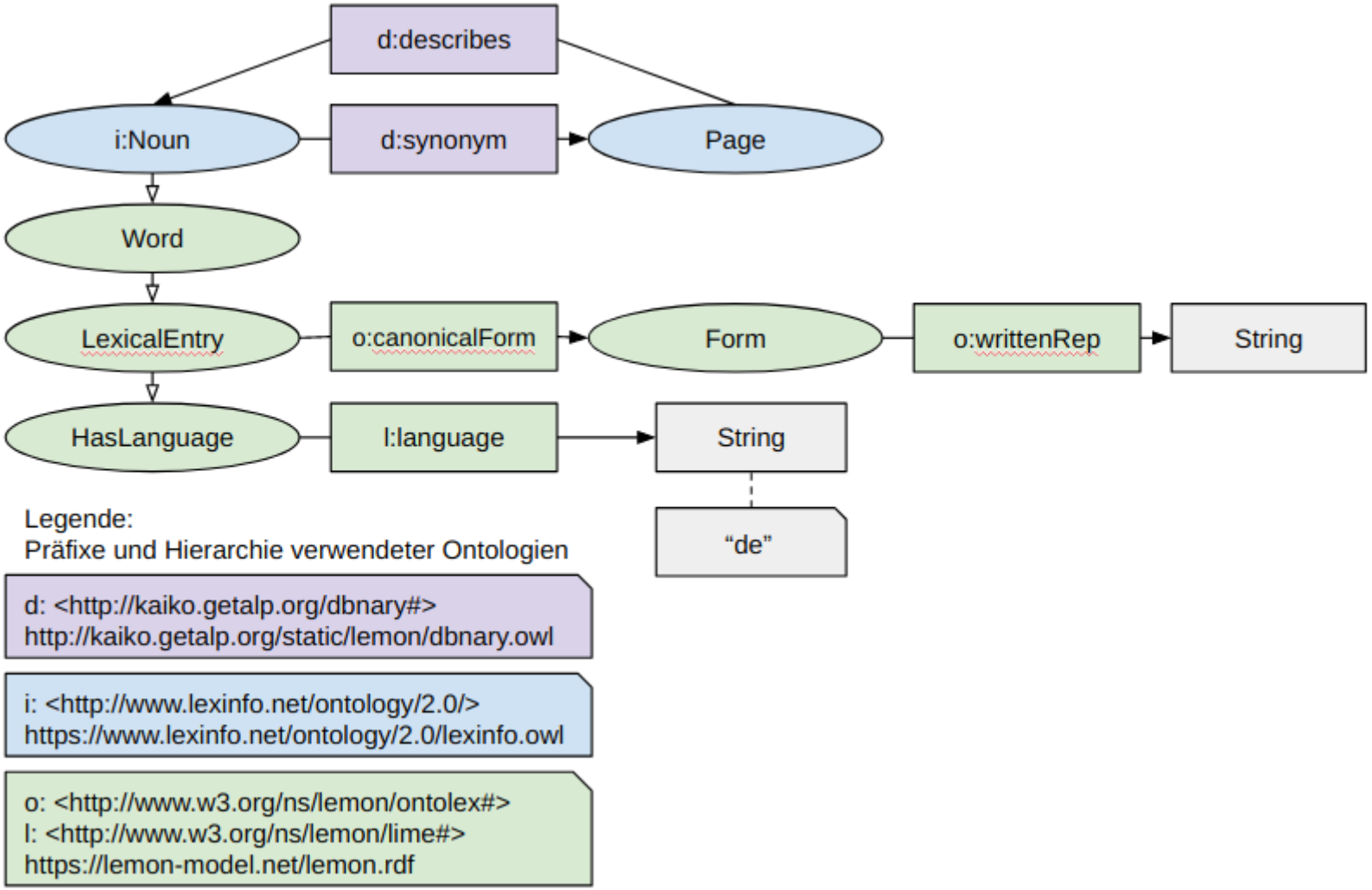
## Arbeitspaket 7:
## Anwendungsfälle

- D7.1 Suchkomponente
- D7.2 Benchmarking der Suchkomponente
- D7.3 City-App Demonstrator
- D7.4 Social Media Bot Demonstrator

## D7.1 Suchkomponente: Elasticsearch

- Ansatz hier:
  **Generierung von Synonym-Listen zur Auffindbarkeit**
- Ergebnis:
  - **6.668 deutschsprachige Nomen**,
    für die Synonyme bekannt sind
  - **21.634 Synonyme** zu den entsprechenden Nomen
  - Auswahl von Nomen aus Titeln und
    Beschreibungestexten aus mCLOUD und GovData
    → **1.497 Nomen und entsprechende Synonyme**

# D7.1 Suchkomponente: Elasticsearch



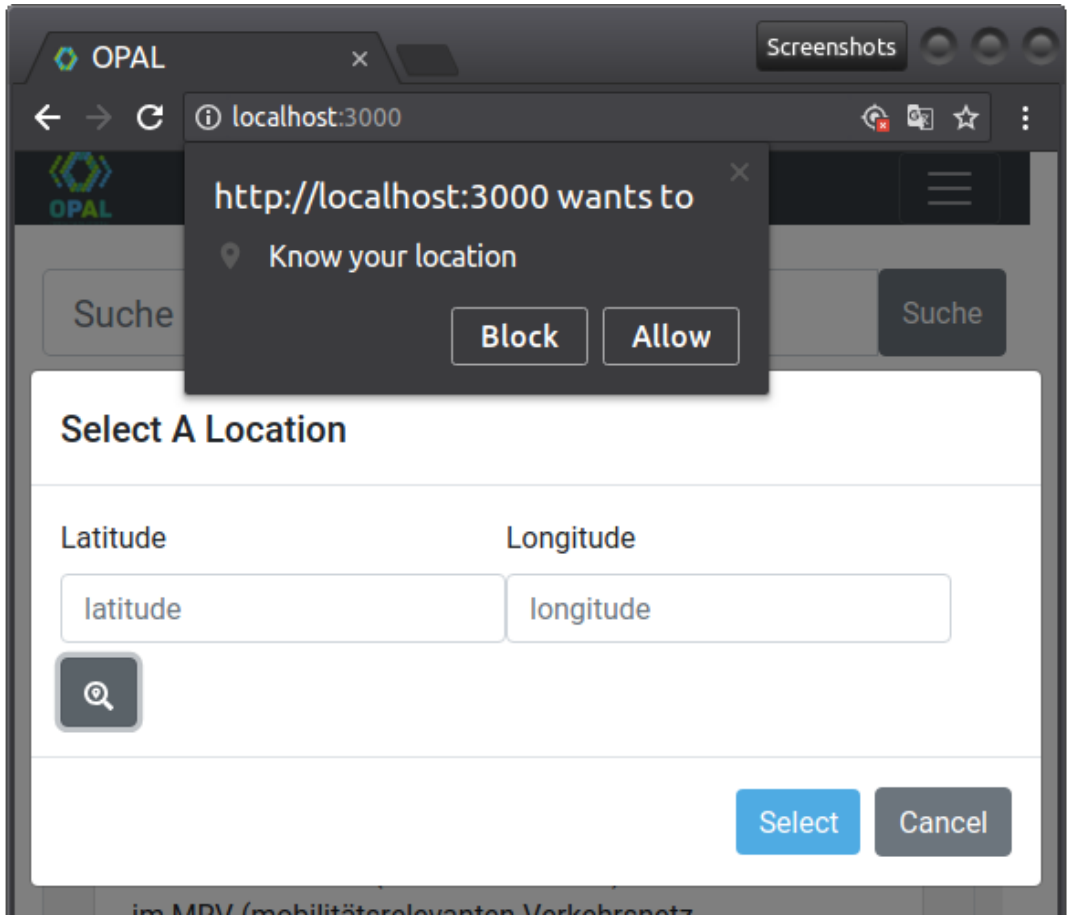

Deliverable als PDF-Datei ✓

# D7.2 Benchmarking der Suchkomponente

| Query | Elasticsearch | Fuseki |
|---|---|---|
| SELECT (COUNT(distinct ?s) AS ?num)<br>WHERE { GRAPH ?g {?s a dcat:Dataset } } | **.030** +- .002 | **.118** +-0.028 |
| SELECT (COUNT(distinct ?s) AS ?num)<br>WHERE { GRAPH ?g {<br>?s a dcat:Dataset . ?s dct:title ?o .<br>FILTER isLiteral(?o)<br>FILTER contains(STR(?o), "Berlin") }} | **.023** +- .002 | **.349** +- .156 |
| SELECT (COUNT(distinct ?s) AS ?num)<br>WHERE { GRAPH ?g {<br>?s a dcat:Dataset . ?s dct:description ?o .<br>FILTER isLiteral(?o)<br>FILTER contains(STR(?o), "Baustelle") }} | **.023** +- .002 | **.329** +- .058 |
| SELECT (COUNT(distinct ?s) AS ?num) WHERE{<br>GRAPH ?g {<br>?s a dcat:Dataset . ?s dcat:keyword ?o .<br>FILTER isLiteral(?o)<br>FILTER contains(STR(?o), Bahnhof)}} | **.240** +- .002 | **1.234** +- .03564 |
| SELECT (COUNT(distinct ?s) AS ?num)<br>WHERE { GRAPH ?g {<br>?s a dcat:Dataset . ?s dct:description ?o .<br>FILTER isLiteral(?o)<br>FILTER contains(STR(?o), "Berlin Flughafen") }} | **.023** +- .002 | **0.4635** +- .1547 |

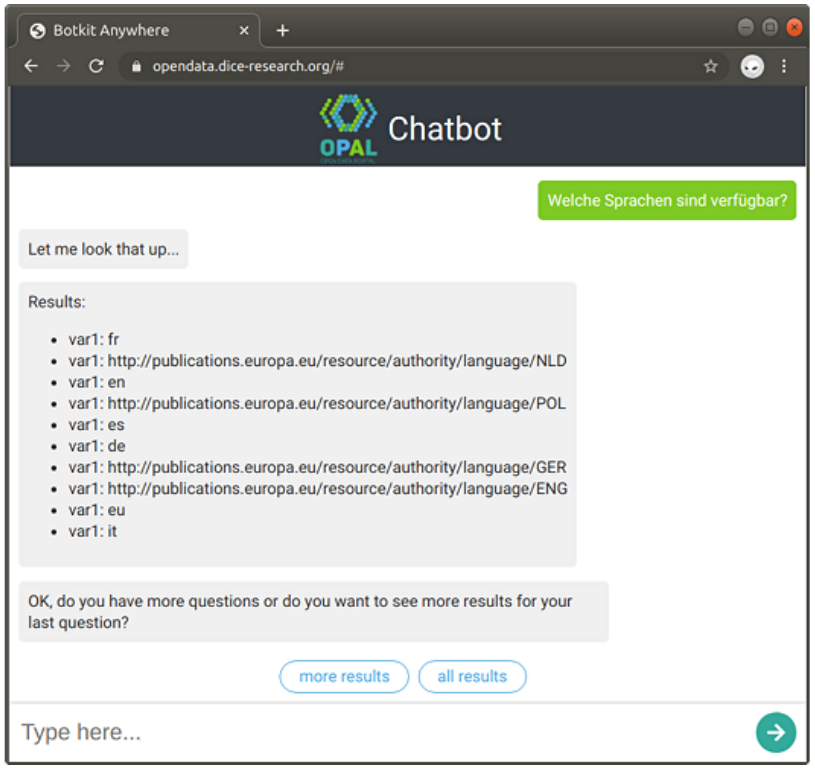
- Deliverable (Code): Search-Component-Benchmark
- Deliverable als PDF-Datei ✓

# D7.3 City-App Demonstrator




- Responsives Webdesign
- W3C Geolocation API Specification
- Deliverable (Code): OPAL Web User Interface ✓

# D7.4 Social Media Bot Demonstrator




- Demo
  - Welche Datenformate gibt es?
  - How many Datasets are available in RDF?
- Dokumentation Bachelorarbeit Marten Schmidt
- Deliverable (Code): DCAT QA ✓

# Arbeitspaket 8: Portalentwicklung

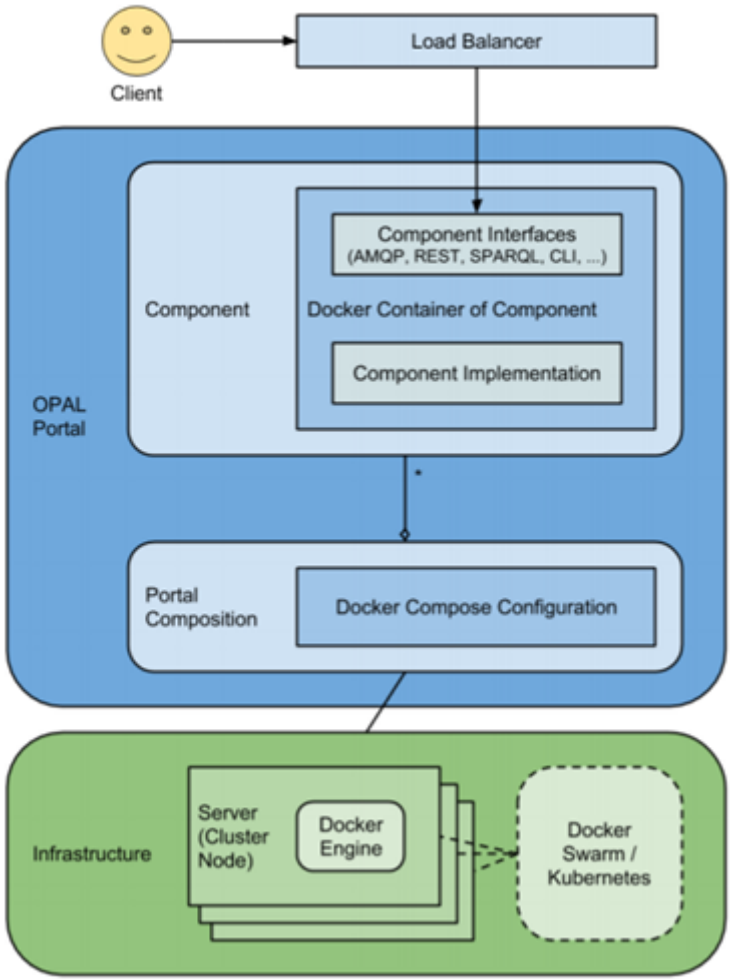
## Arbeitspaket 8:
## Portalentwicklung

**Ziel**: "In Arbeitspaket 8 wird das Open Data Portal Germany als erweiterbare komponentenbasierte **Webanwendung** entwickelt."

## Arbeitspaket 8:
## Portalentwicklung

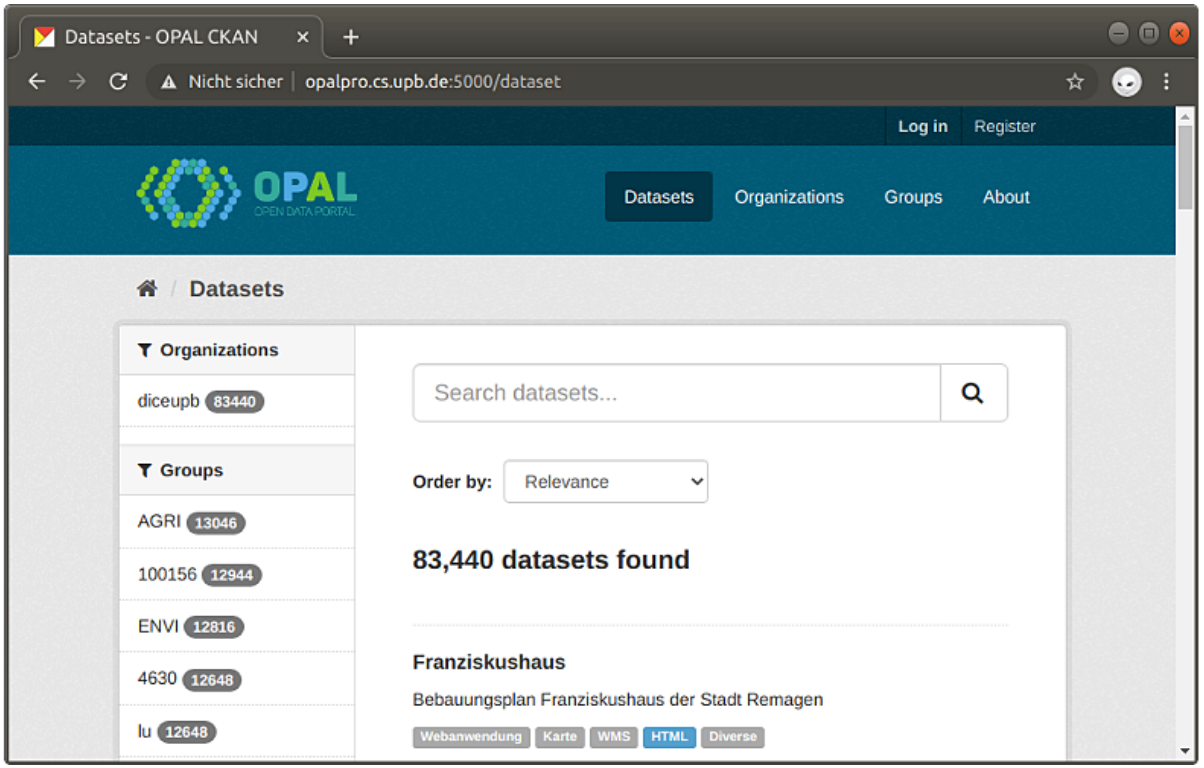
- D8.1 Portalinfrastruktur
- D8.2 Erster Portalprototyp
- D8.3 Erweiterter Portaldemonstrator
- D8.4 Finales OPAL-Portal
- D8.5 Anwenderdokumenation zum OPAL-Portal
- D8.6 Evaluierungsergebnisse

# D8.1 Portalinfrastruktur (2018)



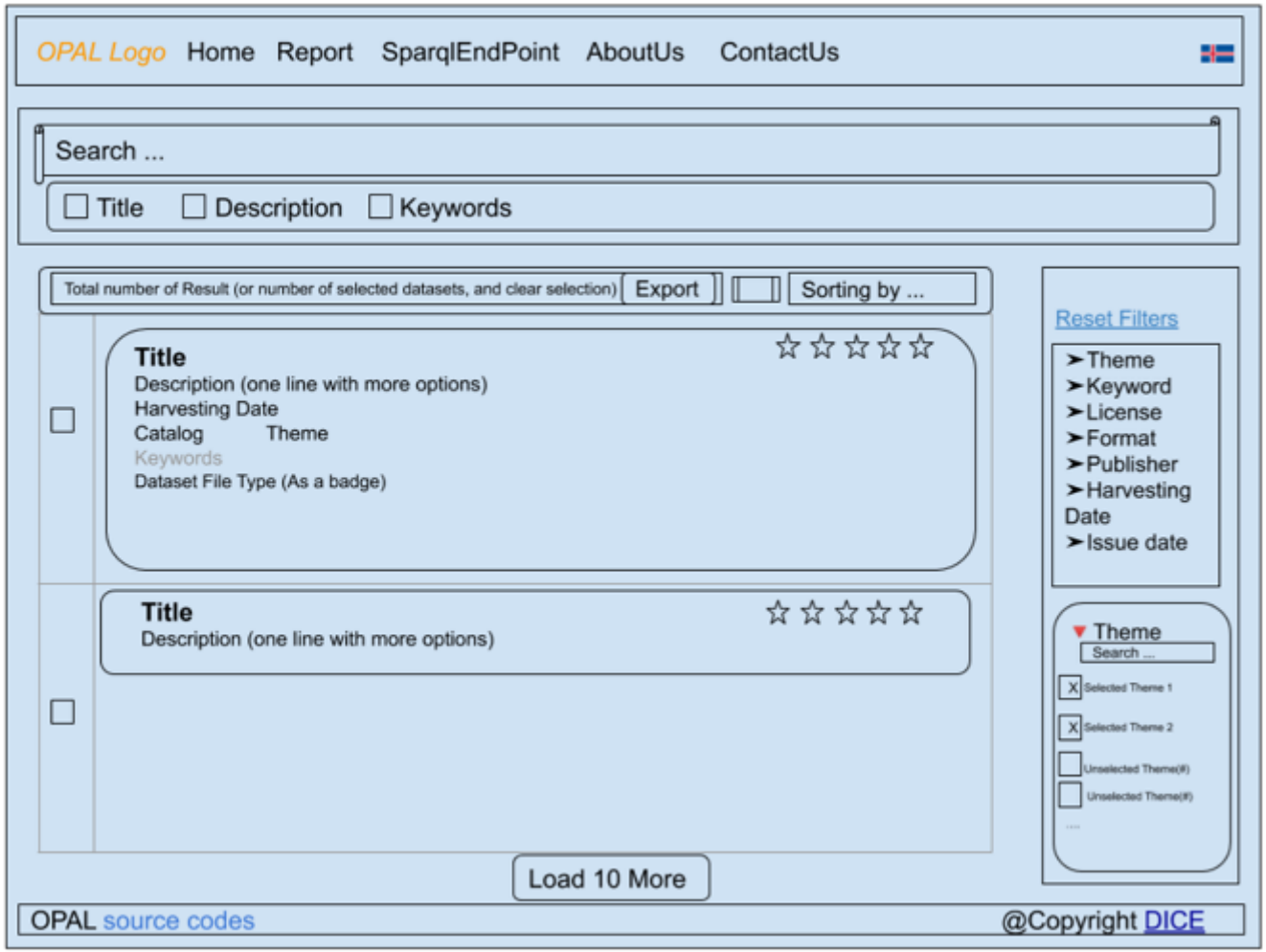

- Deliverable als PDF-Datei ✓

# D8.2 Erster Portalprototyp (2018/19)



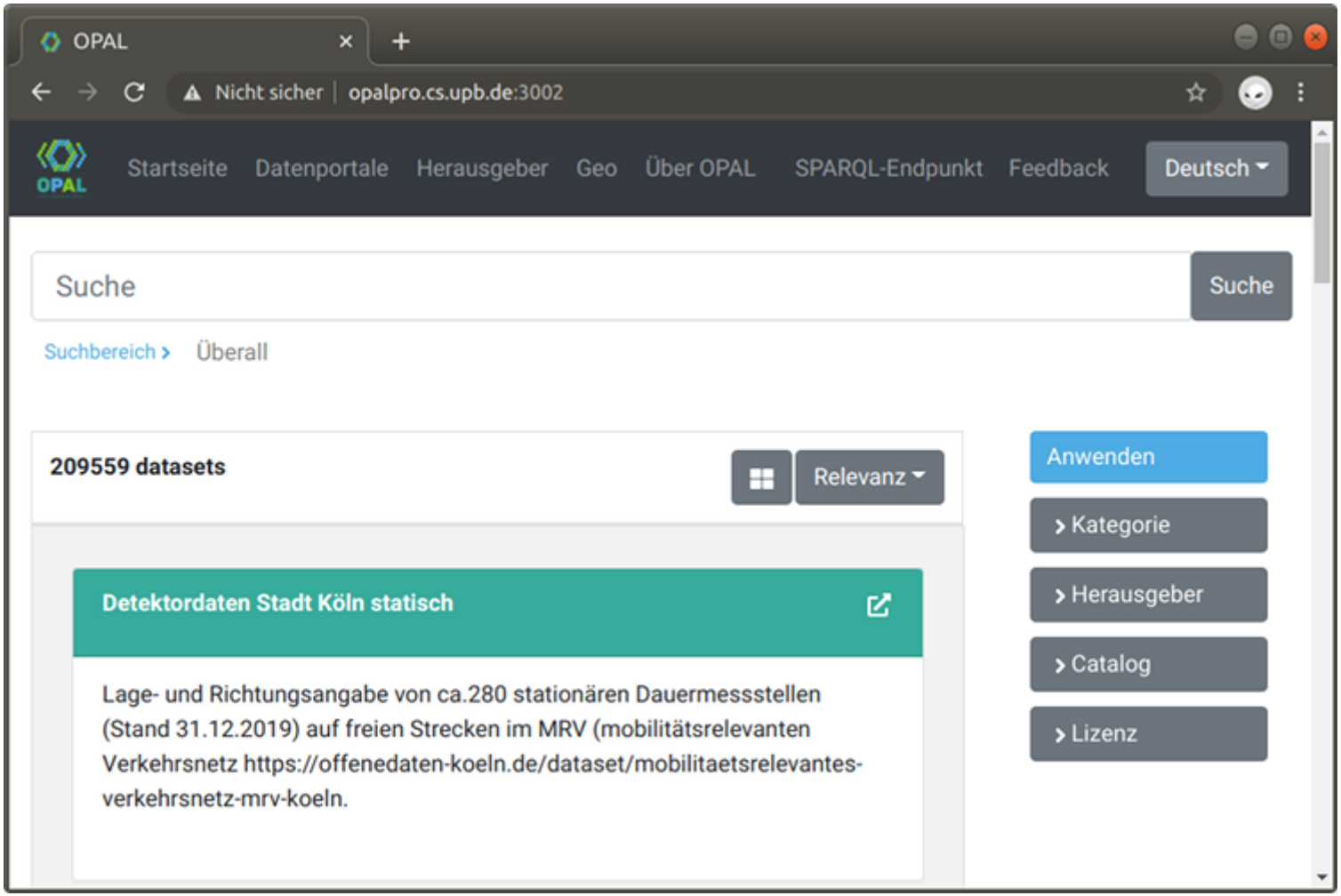

- Konfiguration (Code): opal-ckan-docker und opal-docker-compose
- Deliverable als PDF-Datei ✓

# D8.3 Erweiterter Portaldemonstrator (Juni 2019)




Mockup aus Deliverable als PDF-Datei ✓

# D8.4 Finales OPAL-Portal (2020)

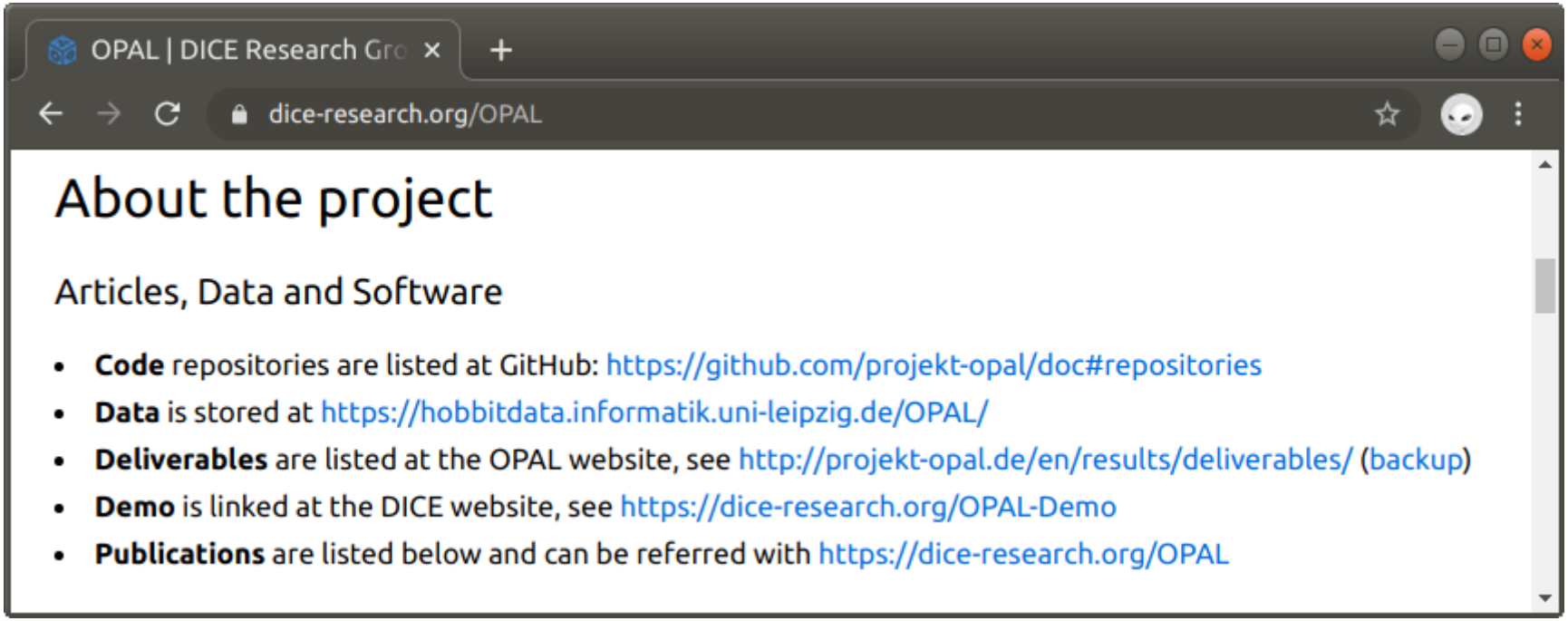



Deliverable (Code): OPAL Demo ✓

## D8.5 Anwenderdokumenation zum OPAL-Portal

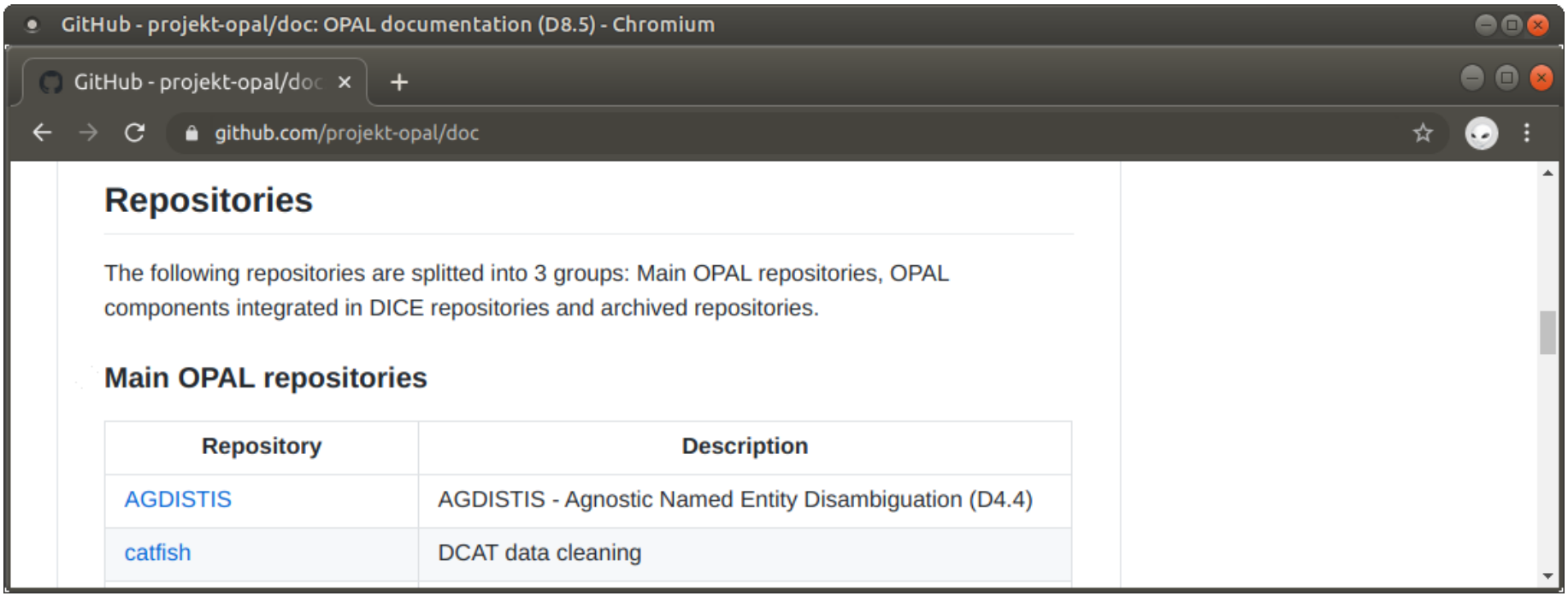
- Anwender: Nutzer der Projektergebnisse
- Übersicht auf **Projektwebseite** (folgt gleich)
- **Code** Dokumentation (folgt gleich)
- Beschreibung der **20 Komponenten** (Code) auf 40 Seiten: Deliverable als PDF-Datei

# D8.5 Dokumentation: Projektwebseite




- Links zu allen Ergebnissen
- Mittel- und langfristige Auffindbarkeit
- dice-research.org/OPAL

# D8.5 Dokumentation: Code

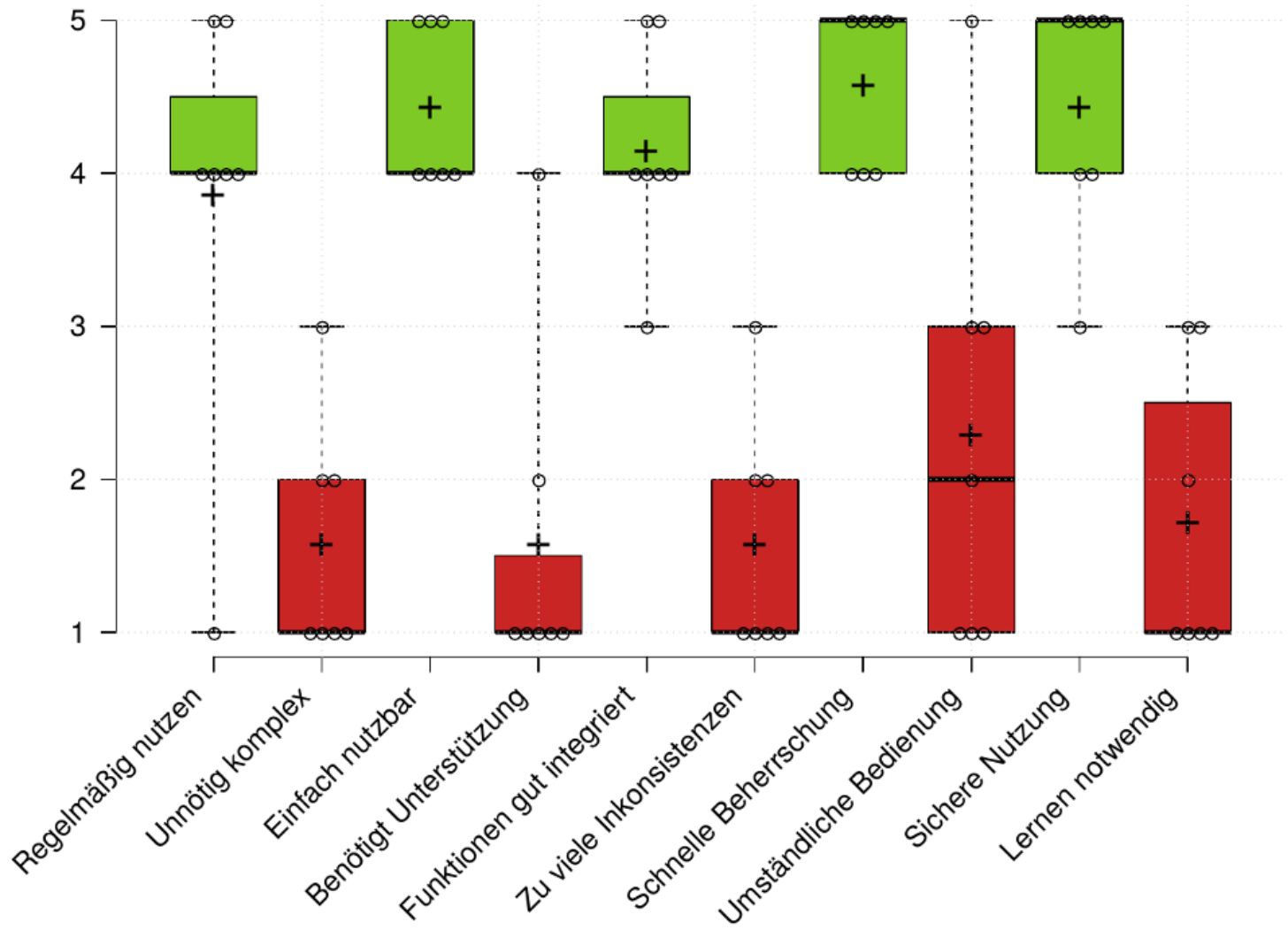



- Rund **50 Repositorien**
- Aufteilung Haupt- und Nebenprodukte
- Kurzbeschreibungen
- README Dateien und im Code
- github.com/projekt-opal/doc ✓

## D8.6 Evaluierungsergebnisse

- System Usability Scale (SUS)
- Bewertung der Gebrauchstauglichkeit von Systemen
- Einfacher und technologieunabhängiger Fragebogen
- 10 Fragen nach der Likert-Skala
  - Stimme überhaupt nicht zu (1)
  - Stimme voll und ganz zu (5)
- Kommentare und Feedback

# D8.6 Evaluierungsergebnisse




Deliverable als PDF-Datei ✓

# Arbeitspaket 9: Projektmanagement

## Arbeitspaket 9:
## Projektmanagement

Ziel: "Dieses Arbeitspaket beinhaltet die **Koordination**, **Dokumentation** und **Überwachung** der Meilensteine des Projekts, die Kommunikation mit dem Projektträger sowie die **Kommunikation** mit Externen aus Industrie und Forschung."

## Arbeitspaket 9

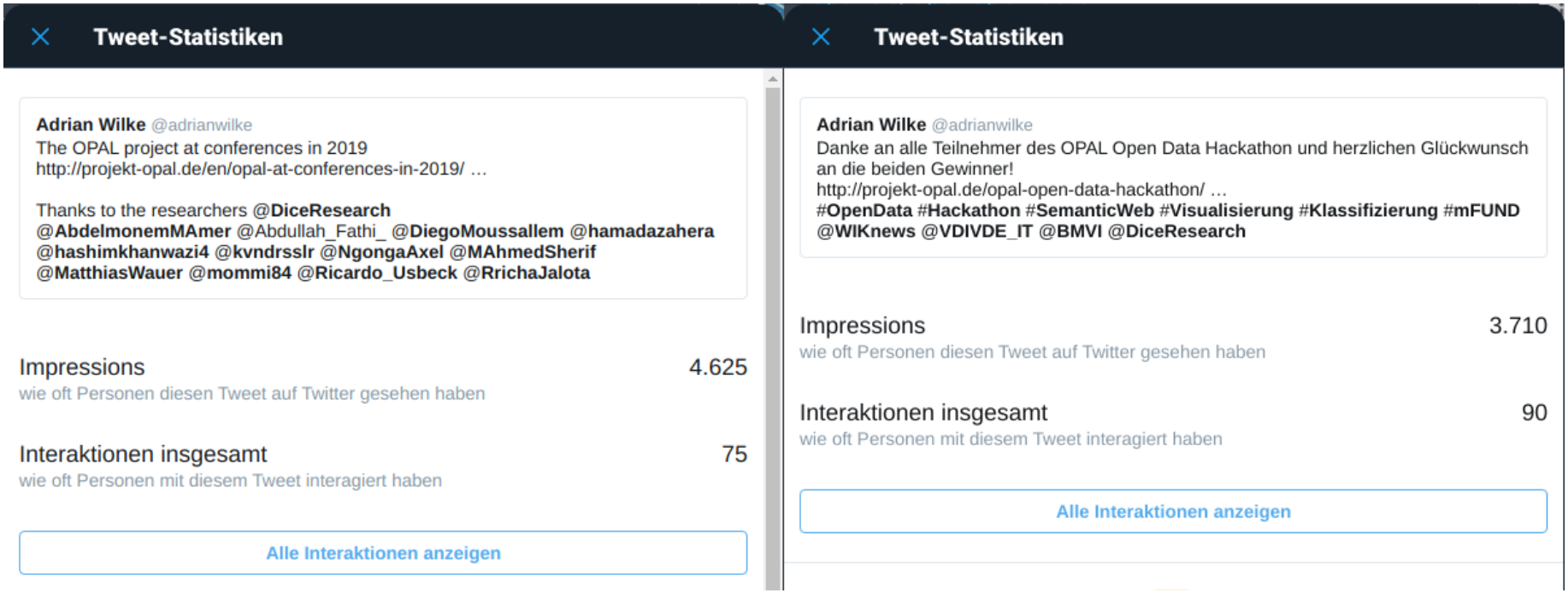
- D9.1 Kommunikations- und Disseminationsplan
- D9.2 Projektbericht Jahr 1
- D9.3 Projektbericht Jahr 2
- D9.4 Projektbericht Jahr 3

## D9.1 Kommunikations- und Disseminationsplan

- Projekt-Webseite: dice-research.org/OPAL
  - **OPAL Portal Demo**
  - **Deliverables** (PDF)
  - **Code** (Github, Open Source)
  - **Daten** (Web, FTP)
  - **Wissenschaftliche Publikationen**
- Deliverable als PDF-Datei

# D9.1 Dissemination: Social Media

## Twitter: OPAL Tweets (Auszug)



## D9.1 Dissemination: Blog und Publikationen

- **Blog**, z.B.: OPAL Open Data Hackathon
  - Zweisprachig
  - Studierende
- **Konferenzen und Publikationen**:
  - OPAL Konferenzbeiträge 2019
  - In **2019** in mind. **11 Beiträgen**: "This work has been supported by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) within the project OPAL under the grant no 19F2028A."
  - Liste auf der OPAL Projektwebseite
  - Vollständige Liste im Abschlussbericht ✓

## D9.2/D9.3/D9.4 Projektberichte

- D9.2 Projektbericht Jahr 1 ✓
- D9.3 Projektbericht Jahr 2 ✓
- D9.4 Projektbericht Jahr 3 ✓
- Abschlussbericht (bis März 2021)
- mFUND Projektsteckbrief

# Abschluss

## Langzeitverfügbarkeit

- Projektwebseite DICE (dice-research.org/OPAL)
  - Dokumente & Code & Daten
- Projektwebseite BMVI (www.bmvi.de...opal.html)
- Demo OPAL Portal (dice-research.org/OPAL-Demo)
- Demo Social Media Bot

## Zusätzliche Entwicklungen (Auswahl)

### OPAL Export

- Exportiert **RDF** ins **CSV** Format (Code)
- In Zusammenhang mit mCLOUD/mFUND entwickelt

### Apache Jena

Open Source Projekt: Erweiterung um **DCAT Version 2** (Code)

**Vielen Dank!**

Diese Präsentation online:
projekt-opal.github.io/doc/final-presentation/Praesentation/

OPAL Projektwebseite der DICE Fachgruppe:
dice-research.org/OPAL