OPAL
OPEN DATA PORTAL

Deliverable D1.2
Datenanalyse

Autoren: Matthias Wauer

Reviewer: Adrian Wilke

Veröffentlichung	Öffentlich
Fälligkeitsdatum	31.12.2017
Fertigstellung	15.05.2018
Arbeitspaket	AP1
Typ	Bericht
Status	Final
Version	1.0

Kurzfassung:
In der initialen Datenanalyse wurde eine Liste von relevanten Datenquellen identifiziert. Diese setzt sich aus Meta-Datenportalen, z.B. mCloud und MDM, zusammen. Die Datenquellen wurden anschließend nach multi-methodischem Ansatz analysiert. Als Ergebnis dieser Analyse lässt sich festhalten, dass das Crawling der Datenquellen (bei MDM eingeschränkt) möglich ist und eine Integration und Fusion der Metadaten aus den primären und externen Quellen durchführbar und sinnvoll erscheint.

Schlagworte:
Open Data, Portale, Datenanalyse

mFUND
Das Startkapital für die Mobilität 4.0

Bundesministerium
für Verkehr und
digitale Infrastruktur

Inhalt

1 Einleitung **3**

2 Fragestellungen zur Datenanalyse **3**

3 Betrachtete Datenquellen **4**

4 Ansatz **4**

5 Analyseergebnisse der technischen und statistischen Untersuchung **5**

5.1 Untersuchung von mCLOUD 5

5.1.1 Datensätze 5

5.1.2 Zugriffsmöglichkeiten 5

5.1.3 Metadaten 6

5.1.4 Format der Metadaten 6

5.1.5 Anbieter 6

5.1.6 Veränderlichkeit 6

5.1.7 Fehleranalysen 6

5.2 Untersuchung von MDM 7

5.2.1 Datensätze 7

5.2.2 Zugriffsmöglichkeiten 7

5.2.3 Metadaten 7

5.2.4 Format der Metadaten 8

5.2.5 Anbieter 9

5.2.6 Veränderlichkeit 9

5.3 Weitere Portale 9

5.3.1 GovData 9

5.3.2 OffeneDaten.de 9

5.3.3 European Data Portal 10

6 Detaillierte Analyse zufällig ausgewählter Datensätze **10**

6.1 Methodik 10

6.2 Auswahl der Datensätze von mCLOUD 11

6.3 Untersuchung der Datensätze 11

6.3.1 Datensatz 209 “Jährliche Raster von Winterraps…” 11

6.3.2 Datensatz 292 “Monatliche Sonnenscheindauer” 12

6.3.3 Datensatz 435 “Serviceeinrichtungen” 12

6.3.4 Datensatz 378 “RadwegeGis Hamburg” 13

6.3.5 Datensatz 133 “Grundwassergleichen Max 2008” 14

6.3.6 Datensatz 49 “Bund: Farbrelief des Wasserlaufs” 14

6.3.7 Datensatz 600 “VBB-Fahrplan 2013” 14

6.3.8 Datensatz 245 “Urbane Räume: Lufttemperatur und Luftfeuchte stündlich” 15

6.3.9 Datensatz 84 “Digitale Bundeswasserstraßenkarte im Maßstab 1:1.000.000” 15

6.3.10 Datensatz 606 “VBB-Fahrplandaten August 2017 bis Dezember 2017” 15

6.4 Zusammenfassung der untersuchten Datensätze 15

7 Zusammenfassung und Ausblick 17

1 Einleitung

Im Rahmen des OPAL-Projektes sollen Metadaten offener Datensätze aus verschiedenen Datenquellen, insbesondere den Portalen mCLOUD und MDM, in einem ganzheitlichen Datenportal integriert und miteinander verknüpft werden. Eine zentrale Voraussetzung dafür ist eine Spezifikation dieser Datenquellen und eine konkrete Betrachtung der Eigenschaften der jeweiligen Datensätze.

In diesem Deliverable sollen daher bestehende Datenquellen hinsichtlich verfügbarer Metadaten untersucht werden. Aufbauend auf Erkenntnissen der vorhergehenden Anforderungsanalyse (D1.1) werden hierzu zunächst Fragen abgeleitet, die anschließend zur Umsetzung des OPAL-Portals hinsichtlich der Datenquellen beantwortet werden. Nach Auswahl der zu betrachtenden Datenquellen werden statistische Daten zu den entsprechenden Datensätzen erhoben, um die vorher festgelegten Fragestellungen zu beantworten. Mittels repräsentativem Sampling werden zusätzlich einzelne Datensätze der Datenbasis ausgewählt und jeweils manuell untersucht, um genauere Informationen zu Datenqualität, Zugriffsmöglichkeiten und notwendigen spezifischen Anforderungen an Extraktions-, Transformations- und Integrationsmethoden des Portals zu spezifizieren.

2 Fragestellungen zur Datenanalyse

Ziel der Datenanalyse ist es, eine Strategie für, sowie spezifische Anforderungen an die Crawlerkomponente, Extraktion und Datenintegration im Rahmen von OPAL zu entwickeln. Hierfür wurden folgende Fragestellungen herausgearbeitet:

1. Welche Datenquellen sollen im Rahmen von OPAL betrachtet werden?
2. Wie viele Datensätze enthalten die Datenquellen?
3. Welche Zugriffsmöglichkeiten bietet die Datenquellen …
    a. … hinsichtlich Zugriffsprotokollen?
    b. … hinsichtlich Datenrepräsentationen (z.B. Datenformate, Strukturiertheit der jeweiligen Metadaten, Verfügbarkeit von Identifikatoren wie URIs)?
    c. … hinsichtlich Interaktionsmustern (z.B. Request-Response, Publish-Subscribe)?
    d. … hinsichtlich direktem Zugriff oder indirekter Einbindung mittels Katalogsystemen (im Fall von Meta-Datenportalen wie z.B. mCLOUD)?
4. Welche Metadaten können für die Datenquellen angegeben werden, und wie häufig enthalten die Datensätze die jeweiligen Metadaten statistisch?
5. In welchem Format liegen die jeweiligen Metadaten vor (z.B. Struktur, Einheitlichkeit, Identifikatoren)?
6. Auf wieviele Anbieter verteilen sich die Datensätze einer Datenquelle?
7. Wie veränderlich ist die Datenquelle (Häufigkeit/Umfang der Änderung der Anzahl von Datensätzen)?
8. Wie veränderlich sind die Metadaten der Datensätze innerhalb der Datenquelle
    a. … hinsichtlich der durchschnittlichen Aktualisierunghäufigkeit?
    b. … hinsichtlich der verwendeten Metadatenattribute?
    c. … hinsichtlich der Veränderlichkeit des Aktualisierungsintervalls (z.B. immer wöchentlich oder in nicht vorhersagbaren Zeitabständen)?

3 Betrachtete Datenquellen

Von zentraler Bedeutung ist die Auswahl von zu betrachtenden Datenquellen (Fragestellung 1). Der OPAL-Ansatz soll die Extraktion und Integration von Metadaten beliebiger Datenquellen erlauben. Im Zuge der initialen Untersuchung im Rahmen dieses Deliverables findet aus Gründen der Praktikabilität eine Beschränkung auf exemplarische Datenquellen mit hoher Bedeutung für das Projekt statt.
Die folgende Liste der zu betrachtenden Datenquellen ergibt sich sowohl aus der Vorhabensbeschreibung des OPAL-Projektes, als auch aus den in D1.1 erfassten Anforderungen hinsichtlich der Datenquellen, die in Fragebögen und externen Projekten als relevant erachtet werden:

1. mCLOUD
    a. Daten innerhalb des mCLOUD-Portals
    b. Verlinkte Portale (in der detaillierten Analyse in Abschnitt 6 identifiziert), z.B.:
        i. http://data.deutschebahn.com/
        ii. https://cdc.dwd.de/portal
        iii. http://transparenz.hamburg.de/
        iv. https://open.nrw/
        v. http://metaver.de/portal/
        vi. https://daten.berlin.de/datensaetze
2. MDM
3. govdata.de
4. OffeneDaten.de
5. European Data Portal

Zusätzlich zu diesen Datenquellen, die im Rahmen von OPAL besonders relevant erscheinen, steht auf der Webseite https://portal.opengeoedu.de eine umfangreiche Sammlung von derzeit 244 Datenportalen zur Verfügung. Davon sind 155 Portale für den Bereich Deutschland relevant.

4 Ansatz

Zur Beantwortung der weiteren Fragestellungen aus Abschnitt 2 für die in Abschnitt 3 genannten Datenquellen betrachten wir die jeweiligen Portale mit unterschiedlichen Methoden.
Für die Fragestellungen 2 (Anzahl Datensätze) und 3 (Zugriffsmöglichkeiten) ist eine manuelle Prüfung bzw. ein Studium der Dokumentation der Datenquellen hinreichend. Für die weiteren Fragestellungen 4 bis 8 hingegen werden umfangreichere Analysen auf dem Datenbestand der jeweiligen Quelle erforderlich. Diese setzen eine Erfassung der Meta-Datensätze voraus, entweder durch vollständiges Crawling oder durch repräsentatives Random Sampling.
Die Beantwortung der Fragestellungen zur Veränderlichkeit (7 und 8) kann zu diesem Zeitpunkt nur anhand von Schätzungen erfolgen. Für eine quantitative Aussage muss ein kontinuierliches Crawling und eine entsprechende Auswertung erfolgen.

Die folgenden Punkte skizzieren die Vorgehensweise für die jeweiligen Datenquellen:

- Automatisierte Untersuchung aller Datensätze auf mCLOUD
    - mit frühem Prototyp eines Crawlers
- Manuelle Untersuchung von Random Samples der mCLOUD

    - Verfolgung zu ursprünglicher Quelle
    - Überprüfung auf weitere Metadaten (an ursprünglicher Quelle oder zusätzlicher externer Quelle)
- Manuelle Untersuchung von MDM
- Manuelle Untersuchung von govdata.de, OffeneDaten.de und European Data Portal

5 Analyseergebnisse der technischen und statistischen Untersuchung

Dem in Abschnitt 4 beschriebenen Ansatz entsprechend wurden die Portale mCLOUD und MDM sowie als weitere Datenquellen govdata, OffeneDaten und das European Data Portal zunächst manuell und anschließend statistisch überprüft.

5.1 Untersuchung von mCLOUD

5.1.1 Datensätze

Zum Zeitpunkt der Untersuchung (1. Februar 2018) enthielt mCLOUD insgesamt 652 Datensätze.

5.1.2 Zugriffsmöglichkeiten

Das Meta-Datenportal mCLOUD stellt Metadaten und Links zu offenen Daten des BMVI zur Verfügung. Die Webseite auf Basis des Bootstrap-Frameworks stellt die einzige Zugriffsmöglichkeit dar, mCLOUD bietet keine API an. Eine manuelle Überprüfung ergab, dass sowohl die Inhalte der Startseite als auch der Seiten zur Beschreibung der Datensätze vom Server statisch ausgegeben werden, d.h. nicht auf Clientseite durch Scripts modifiziert werden. Dies erleichtert die Erfassung durch Crawler. mCLOUD bietet eine paginierte Liste aller Datensätze an, die für vollständiges Crawling verwendet werden kann. Alle für OPAL relevanten Daten liegen in HTML vor und können als semistrukturierte Daten mit entsprechenden Frameworks extrahiert werden.

<table>
<tr><th>Zugriffsmöglichkeit</th><th>Unterstützt durch mCLOUD</th></tr>
<tr><td>Zugriffsprotokoll auf Metadaten</td><td>HTTP/1.1</td></tr>
<tr><td>Datenrepräsentation</td><td>HTML5
semistrukturiert (Tabellen, z.T. proprietäre CSS-Klassen)
Keine Identifikatoren
Teilweise fehlerhafte Daten (z.B. “ja” im href-Attribut eines Link-Tags)</td></tr>
<tr><td>Interaktionsmuster</td><td>Pull</td></tr>
<tr><td>Zugriff auf Daten</td><td>Metadatenkatalog, Links auf externe Daten (über unterschiedliche Protokolle) durch CSS-Klasse a.mcloud__link unterhalb von ul.download-list
<table><tr><th colspan="2">Protokolle</th></tr><tr><td>Zugriffstyp</td><td>Anzahl</td></tr></table></td></tr>
</table>

Download (HTTP)	315
FTP	229
Portal (HTTP)	232
WMS	204
WFS	98
ATOM-Feed	41
API (HTTP)	42
WMTS	9
SOS	2
(Parsingfehler)	10

Tabelle 1: Zugriffsmöglichkeiten bei mCLOUD

5.1.3 Metadaten

mCLOUD listet auf den Suchergebnisseiten für fast alle Datensätze einen Beschreibungstext (unstrukturiert, teilweise mit Links), Anbieter (teilweise mit Link, teilweise fehlerhaft), Lizenz (mit Link), Kategorie (oft mehrere) und Datenzugriffstyp (oft mehrere, teils fehlerhaft). Weitere Metadaten werden nicht oder nur implizit im Beschreibungstext oder auf der eigentlichen Datenquelle bereitgestellt.

5.1.4 Format der Metadaten

Metadatum	**Format**
Beschreibungstext	Unstrukturiert, z.T. semistrukturiert (Links)
Anbieter	Semistrukturiert (mit Link, oft fehlerhaft)
Datenzugriffstyp	Enumeration

Tabelle 2: Metadatenformate von mCLOUD

5.1.5 Anbieter

mCLOUD enthält Daten von insgesamt 60 verschiedenen Anbietern. Größter Anbieter ist der Deutsche Wetterdienst (DWD) mit 237 Datensätzen.

5.1.6 Veränderlichkeit

mCLOUD ist weitgehend statisch, sowohl hinsichtlich der neu hinzugefügten und entfernten Datensätze als auch der Änderung der Metadaten. Änderungen an den Metadatenattributen finden im beobachteten Zeitraum nicht statt. Zu bestimmten, aber nicht vorhersehbaren Zeitpunkten werden (meist mehrere) zusätzliche Datensätze hinzugefügt.

5.1.7 Fehleranalysen

Eine weitere Untersuchung ergab, dass von den 652 Datensätzen:

- 10 keine oder eine fehlerhafte URL zum Datensatz enthielten
- 40 Datensätze keine Beschreibung enthielten
- 6 Datensätze ohne Lizenzangabe vorhanden waren
- 59 Datensätze mit mehr als einer Kategorie versehen waren (kein Fehler im eigentlichen Sinne)

Zusätzlich ist auffällig, dass zum Zeitpunkt der Fertigstellung dieses Berichts alle Links zu Datenanbietern fehlerhaft waren.

5.2 Untersuchung von MDM

5.2.1 Datensätze

Zum Zeitpunkt der Untersuchung (14. Februar 2018) enthielt MDM insgesamt 119 Datensätze.

5.2.2 Zugriffsmöglichkeiten

MDM ist über die URL http://service.mdm-portal.de erreichbar. Eine Recherche von Datensätzen erfolgt über ein Formular. Eine Liste aller Datensätze kann demnach nur über einen HTTP-POST-Request bezogen werden. Die Ergebnisseite enthält eine mittels jQuery dynamisch erzeugte Pagination, weshalb das Crawling nur mit einem Crawler erfolgen kann, der die Ausführung dieser Skripte unterstützt.
Sobald diese Liste von Publikationen vorhanden ist, können die Metadaten der Datensätze über normale HTTP-GET-Requests bezogen werden. Nahezu alle für OPAL relevanten Metadaten liegen in auf den zurückgelieferten Detailseiten in statischem HTML vor. Ausnahmen:

- Geografische Ausdehnung, die über eine Popup-Karte dargestellt wird (separater HTTP-GET-Request, der die Geometrie im proprietären MDM-Georef-Format innerhalb eines Script-Tags enthält)
- Kontaktangaben (innerhalb eines Script-Tags auf der Detailseite)

Zugriffsmöglichkeit	**Unterstützt durch MDM**
Zugriffsprotokoll auf Metadaten	HTTP/1.1
Datenrepräsentation	HTML5 semistrukturiert, teilweise dynamisch (jQuery, Tabellen, z.T. proprietäre CSS-Klassen) Proprietäre Identifikatoren (Integer) für Publikationen
Interaktionsmuster	Pull
Zugriff auf Daten	Metadatenkatalog, teilweise Links auf Datengeber und Dateischema, Subskription (mit Zugriff auf eigentliche Daten) nur über Email

Tabelle 3: Zugriffsmöglichkeiten bei MDM

5.2.3 Metadaten

Eine Auswahl an Metadaten erlaubt das Portal bereits beim Formular der Suchergebnisliste. Diese kann auf dem Portal nach folgenden Kriterien eingeschränkt werden:

- Datenart
  - Parkdaten (26 Datensätze)
  - Umfelddaten (5 Datensätze)
  - Verkehrsdaten (18 Datensätze)
  - Verkehrsinformationen (49 Datensätze)
  - Verkehrslage (6 Datensätze)
- Aktualisierungsintervall
  - Ereignisbasiert (50 Datensätze)
  - 1 Minute (36 Datensätze)
  - 5 Minuten (9 Datensätze)
  - 10 Minuten (2 Datensätze)
  - 15 / 30 Minuten (je 1 Datensatz)
  - 1 Stunde (12 Datensätze)
  - 2 / 3 / 6 Stunden (keine Datensätze)
  - 12 Stunden (1 Datensatz)
  - 24 Stunden (5 Datensätze)
- Vertragsfrei / nicht vertragsfrei (41 / 76 Datensätze)
- Format
  - DATEXII (105 Ergebnisse)
  - Container (15 Ergebnisse)

Die Publikationsdetail-Seiten enthalten weitere Metadaten wie Titel, Beschreibung, Datengeber, Bezeichnung, Ansprechpartner, sowie teilweise Angaben zur Qualitätssicherung, Vertragsentwurf, Referenzdateien und Dateischema.

5.2.4 Format der Metadaten

Metadatum	**Format**
Beschreibungstext	Unstrukturiert
Anbieter (“Datengeber”)	Semistrukturiert (mit Link)
Vertragsfrei	Bool (Ja/Nein)
Datenart	Enumeration
Format	Enumeration (DATEXII/CONTAINER)
Frequenz	Enumeration
Bezeichnung	Unstrukturiert
Qualitätssicherung	Unstrukturiert und URL
Dateischema	URL

Vertragsentwurf	Link auf PDF-Dokument
Referenzdateien	ID, Name mit URL und Version

Tabelle 4: Metadatenformate von MDM

5.2.5 Anbieter

MDM enthielt zum Zeitpunkt der Untersuchung Daten von 34 Organisationen. Davon bot das Baustelleninformationssystem des Bundes und der Länder die meisten Datensätze an.

5.2.6 Veränderlichkeit

Die Dynamik von MDM hinsichtlich der Datensätze ist unterschiedlich. Während die konkreten Daten meist dynamisch sind (teilweise auch statische Daten, z.B. statistische Auswertungen), sind die angegeben Metadaten zur Beschreibung der Datensätze meist statisch. Auch die Auswahl der Metadaten in MDM ändert sich nicht.

5.3 Weitere Portale

Der Fokus im Projekt OPAL liegt auf den Portalen mCLOUD sowie MDM. Die Untersuchungsergebnisse der weiteren in Abschnitt 3 angegeben Datenquellen werden im Folgenden daher in gekürzter Form dargestellt.

5.3.1 GovData

Das Portal GovData setzt auf CKAN (Comprehensive Knowledge Archive Network) auf. Im Gegensatz zu den relativ freien Vorgaben für Metadaten in CKAN (lediglich der Name ist verpflichtend) setzt GovData auf eine eindeutige Festlegung von Metadaten nach dem JSON-Schema Open Government Data (OGD).[1] Diese Metadatenstruktur unterscheidet Eigenschaften verschiedener Ebenen, wie in Tabelle 5 zusammengefasst.

Ebene	**Metadaten**
1. “Oberste” Ebene	Titel, Bezeichner, Beschreibung, Verantwortliche, Nutzungsbestimmungen, Liste der Ressourcen
2. Ebene (Extras)	Zeitliche und räumliche Einordnung, Herkunft, Kontakte
3. Ebene	URI oder NUTS-Code für räumliche Einordnung, URL für Lizenz, Art einer Datumsangabe (z.B. erstellt oder veröffentlicht)

Tabelle 6: Metadatenunterscheidung bei GovData (OGD).

GovData enthält zum Zeitpunkt der Untersuchung (24. April 2018) insgesamt 19.754 Datensätze. Die Suche bietet Filtermöglichkeiten über Kategorien, Datenbereitsteller, Schlagwörter, Dateiformat, Lizenz, Offenheit der Lizenz sowie Raum- und Zeitbezug.
Die Datensätze selbst bieten zusätzlich neben den o.g. Metadaten der obersten Ebene eine Bewertungs- und Kommentarfunktion sowie einen Link zu allen Metadaten in RDF/XML.

[1] https://www.govdata.de/metadatenschema

5.3.2 OffeneDaten.de

Das Portal OffeneDaten.de ist ein Katalog für offene Daten aus Deutschland, das nach eigenen Angaben neben Datensätzen der öffentlichen Verwaltung zusätzlich "auch Daten aus anderen Wissensdomainen, aus Wissenschaft und Forschung oder der Privatwirtschaft sowie von Bürgern erzeugte Daten (Crowdsourcing)"[2] anbietet. Wie GovData setzt es auf CKAN auf.

Zum Zeitpunkt der Untersuchung (24. April 2018) enthält OffeneDaten.de insgesamt 28.542 Datensätze, die über Organisationen, Gruppen (entspricht Kategorien in den zuvor beschriebenen Portalen), Tags (Schlagwörter), Formate und Lizenzen gefiltert werden können.
Da viele der Daten automatisiert erfasst werden, ist die Menge und Qualität der Metadaten oftmals gering bis schlecht. Hinzu kommt, dass auch Datensätze aufgeführt werden, die in der ursprünglichen Quelle nicht mehr vorhanden sind. Die "Aktivitätsanzeige" zu jedem Datensatz zeigt, wann der Harvester die Quelle untersucht hat. Eine Diskussion zu jedem Datensatz ist über die eingebundene externe Plattform Disqus möglich.

5.3.3 European Data Portal

Obwohl sich OPAL auf offene Daten aus Deutschland konzentriert, erscheint es uns sinnvoll, als Datenquelle auch das European Data Portal (EDP) heranzuziehen, insbesondere die darin vorhandenen Daten mit Bezug zu Deutschland.
EDP sammelt Daten aus dem öffentlichen Sektor, die auf den jeweiligen Portalen der EU-Mitgliedsstaaten (und einigen weiteren Staaten) zusammengetragen werden. Darüber hinaus bietet das EDP sowohl Datenanbietern als auch -nutzern Unterstützung an. Hierzu sind sowohl Assistenzfunktionen zum Verständnis von Lizenzen sowie ein Dashboard zur Visualisierung bestimmter Aspekte der Metadaten vorhanden.
Zum Zeitpunkt der Untersuchung (24. April 2018) enthält das European Data Portal 817.755 Datensätze aus 78 Datenportalen. Davon sind 206.068 Datensätze aus Deutschland. Hauptsächlich stammen diese aus dem Geodatenportal GDI-DE (188.955 Datensätze), wobei die Metadaten zahlreicher Datensätze in mehreren Sprachen vorhanden sind.
Im Vergleich zu OffeneDaten.de enthält das EDP relativ umfangreiche Metadaten, die bereits in RDF konvertiert sind. Dabei kommen Vokabulare wie DCTerms, DCAT und ADMS (Asset Description Metadata Schema) zum Einsatz.
Durch die große Anzahl an Datensätzen wird eine detaillierte Untersuchung der Metadaten, insbesondere hinsichtlich der Metadatenqualität, Teil des Deliverables D3.1 sein.

6 Detaillierte Analyse zufällig ausgewählter Datensätze

Nachdem in Abschnitt 5 ein Überblick und statistische Auswertungen über die betrachteten Portale beschrieben wurden, wird nachfolgend eine genauere Untersuchung auf dem zentralen Portal mCLOUD durchgeführt.

6.1 Methodik

Zur weiteren Untersuchung wird mit der Methode des Random Sampling eine zufällige Auswahl bestimmter Größe an Datensätzen gewählt. Diese Teilmenge wird manuell auf bereits vorhandene Metadaten und deren konkrete Ausprägung untersucht. Anschließend werden

[2] https://offenedaten.de/

zusätzliche Metadaten für diesen Datensatz gesucht, wobei die folgenden Quellen betrachtet werden:

a) die ursprüngliche Quelle des Datensatzes
b) weitere o.g. Metadatenportale
c) über generische Internet-Suchmaschinen auffindbare zusätzliche Metadaten

6.2 Auswahl der Datensätze von mCLOUD

Für die statistische Untersuchung der mCLOUD-Metadaten wurde eine Liste von Datensätzen erstellt. Aus diesen 652 Datensätzen wurden die in Tabelle 7 dargestellten 10 Datensätze zufällig ausgewählt.

#	Daten-satz[3]	Titel	Link
1	209	Jährliche Raster von Winterraps - Beginn der Blüte in Deutschland	mCLOUD
2	292	Monatliche Sonnenscheindauer	mCLOUD
3	435	Serviceeinrichtungen	mCLOUD
4	378	RadwegeGis Hamburg	mCLOUD
5	133	Grundwassergleichen Max 2008	mCLOUD
6	49	Bund: Farbrelief des Wasserlaufs	mCLOUD
7	600	VBB-Fahrplan 2013	mCLOUD
8	245	Urbane Räume: Lufttemperatur und Luftfeuchte stündlich	mCLOUD
9	84	Digitale Bundeswasserstraßenkarte im Maßstab 1:1.000.000	mCLOUD
10	606	VBB-Fahrplandaten August 2017 bis Dezember 2017	mCLOUD

Tabelle 7: Random Sample der mCLOUD-Datensätze

6.3 Untersuchung der Datensätze

6.3.1 Datensatz 209 "Jährliche Raster von Winterraps..."

Der Datensatz in der Kategorie "Klima und Wetter" wird vom DWD unter der GeoNutzV bereitgestellt. Die Beschreibung gleicht dem Titel. mCLOUD stellt keine weiteren Metadaten bereit und verweist auf einen FTP-Endpunkt[4]. Dieser enthält 24 gz-Komprimierte ASCII-Datensätze (je einen für die Jahre 1992 bis 2017) sowie zwei PDF-Beschreibungsdokumente (in Englisch und Deutsch). Diese enthalten:

- Kontaktinformationen (Anschrift, Telefonnummer, Faxnummer, Email, *keine* persönlichen Daten)
- eine Datenbeschreibung, darin
    - räumliche Abdeckung ("Deutschland")
    - zeitliche Abdeckung ("1.1.1992 - Vorjahr")

[3] Diese Zahl ist die Zeilennummer des Datensatzes in einem Dokument mit allen sortierten Titeln.
[4] ftp://ftp-cdc.dwd.de/pub/CDC/grids_germany/annual/phenology/WRAB/

    - räumliche Auflösung (“1km x 1km”)
    - zeitliche Auflösung (“jährlich”)
    - Projektion (textuelle Beschreibung, enthält EPSG und Link zu .prj-Datei)
    - Formate (textuelle Beschreibung, enthält Parameter wie Anzahl Zeilen und Spalten sowie die Repräsentation nicht belegter Punkte (“-999”)
    - Parameter (textuelle Beschreibung der Werte)
    - Unsicherheiten (textuelle Beschreibung)
    - Qualitätsinformation (textuelle Beschreibung, in diesem Fall keine Qualitätsbewertung)
- Datenherkunft (textuelle Beschreibung der Erfassungsmethodik)
- Qualitätsabschätzung (textuelle Beschreibung)
- Literatur (hier keine Inhalte)
- Copyright (textuelle Beschreibung mit Link zu Nutzungsbedingungen)
- Revisionen (textuelle Beschreibung mit Hinweis auf letzte Aktualisierung)

Externe Metadatenquellen oder Verwendungshinweise wurden für diesen Datensatz nicht gefunden. Im übergeordneten FTP-Verzeichnis befinden sich ähnliche Beschreibungsdokumente, die zusätzlich eine Liste von Abkürzungen zu einzelnen Datensätzen dieser Art enhält (z.B. WRAB für Winterraps - Beginn der Blüte).

6.3.2 Datensatz 292 “Monatliche Sonnenscheindauer”

Wie der vorherige Datensatz steht auch dieser aus der Kategorie “Klima und Wetter” des DWD unter der GeoNutzV. Der Titel wird in der Beschreibung “Monatliche Sonnenscheindauer (Deutschland, gerastert)” leicht konkretisiert. Auch in diesem Fall sind erweiterte Metadaten in Form nur in Form eines PDF-Dokuments im verlinkten FTP-Verzeichnis zu finden. Dessen Struktur entspricht der in 6.3.1 genannten. Zusätzliche Metadaten dieses Dokuments sind:

- Hinweise für Anwendungen (textuelle Beschreibung mit einem Link auf eine Visualisierungs-Webanwendung dieser Daten)
- Literatur (Bibliographie, Liste von 5 Veröffentlichungen, jeweils ohne dereferenzierbare Identifikatoren wie DOI oder URL)

Für ähnliche Datensätze, jedoch nicht exakt diesen, stellt der DWD Informationen über sein CDC-Portal zur Verfügung[5]. Neben den o.g. Metadaten in der PDF-Beschreibung stehen hier zusätzlich in HTML-”Steckbriefen” folgende Metadaten zur Verfügung:

- Kurzname (Identifikator wie “FX_MN003”)
- Kategorien (mit Unterkategorien für Zugriff, Datenherkunft, Parameter, Einheit usw.)
- Beschreibung (kompakte textuelle Beschreibung)
- Datensatzbeschreibung (Links zu PDF in o.g. Format sowie XML-Daten im INSPIRE- und einem OPENDATA-Format[6])
- Zugriff (verschiedene Links zum CDC-Portal, FTP-Server, WMS- und WFS-Diensten wenn vorhanden)

Weiterhin liefert der DWD in den mCLOUD-Beschreibungen dieser Datensätze einen Link zu umfangreichen Metadaten im GMD-Format.[7]

[5] https://cdc.dwd.de/portal
[6] bei Untersuchungen immer nicht abrufbar (HTTP 404)
[7] http://www.isotc211.org/schemas/2005/gmd/

6.3.3 Datensatz 435 "Serviceeinrichtungen"

Dieser Datensatz der Kategorien "Bahn" und "Infrastruktur" wird von der DB Netz AG unter CC-BY-4.0 bereitgestellt. Der Titel wird in der Beschreibung als "Serviceeinrichtungen sind z.B. Abstell- und Zugbildungsanlagen." konkretisiert. mCLOUD bietet darüber hinaus nur zwei Links: einen generellen zum Open-Data-Portal der Bahn und einen Direktlink zum Dateidownload einer ZIP-Datei über HTTP.

Diese Datei enthält XML-Daten zu den Serviceeinrichtungen sowie das zugehörige Schema (XSD). Der Datensatz ist auch über data.deutschebahn.com zu finden, jedoch nicht mit "DB Netz" getaggt (wenn man zunächst diesen Filter auswählt sieht man den Datensatz nicht). Die zugehörige Datensatzbeschreibung (offenbar bereitgestellt mit CKAN) enthält zusätzlich folgende Metadaten:

- Tags (zusätzliche Schlagwörter wie Weiche, Gleis, Oberleitung)
- Langbeschreibung (Semistrukturiert, Liste von Attributen, jeweils mit Beispiel)
- Zusätzliche Informationen (Tabelle mit Ansprechpartner, Kontakt-E-Mail, Erstmalige Bereitstellung und Aktualisierungshäufigkeit (jährlich))
- Informationen für den Eintrag unter "Daten und Ressourcen": Tabelle mit:
    - Erstellt (Datum)
    - Zuletzt aktualisiert (Datum)
    - Format (XML)

Weitere Metadaten auf externen Portalen oder Hinweise auf eine Nutzung dieser Daten konnten nicht ermittelt werden.

6.3.4 Datensatz 378 "RadwegeGis Hamburg"

Dieser Datensatz der Kategorie "Straßen" wird von der Behörde für Wirtschaft, Verkehr und Innovation, Amt für Verkehr und Straßenwesen der Stadt Hamburg unter der Datenlizenz Deutschland – Namensnennung – Version 2.0 (dl-de/by-2.0) bereitgestellt. Die Beschreibung liefert nur Informationen über die Vision und Stand ("im Aufbau"), nicht aber über enthaltene Inhalte. mCLOUD stellt vier Zugriffsoptionen zur Verfügung (Download, Portal, WMS und WFS). Der Download-Link enthält in der URL Hinweise auf den Dateityp (ZIP), allerdings nicht über das eigentliche Datenformat (GML). Der Portal-Link ist nicht datensatz-spezifisch. Im Portal sind die entsprechenden Informationen nicht unter "RadwegeGis", sondern offenbar unter "Velo- und Freizeitrouten" auswählbar.

Weitere Metadaten sind auf der Seite MetaVer[8] zu finden. Diese beinhalten:

- Allgemeines
    - Übergeordnete Objekte ("Amt für Verkehr und Straßenwesen")
    - Adressen (Herausgeber und Ansprechpartner (identisch) mit E-Mail und Telefon)
- Verweise
    - Dienste (die bereits in mCLOUD vorhandenen WFS, WMS, Portal und Download)
    - Verschlagwortung (Suchbegriffe, INSPIRE-Themen und ISO-Themenkategorien)
- Raum/Zeit
    - Raumbezug (Adminstrative Einheit, Geothesaurus-Raumbezug, Raumbezug mittels EPSG)
    - Zeitbezug (Status, Periodizität, Erstellung (Datum), Erläuterung)
- Fachbezug

[8] http://metaver.de/trefferanzeige?docuuid=AF604FB9-A1DE-4885-97DD-2AF348B385B5

    - Fachliche Grundlage (“Nicht bekannt”), Datensatz/Datenserie (“Datensatz”), Identifikator der Datenquelle (Link zu Eintrag auf gdi-de.org, zum Zeitpunkt der Untersuchung nicht erreichbar)
- Verfügbarkeit
    - Zugriffsbeschränkungen, Nutzungsbedingungen, Anwendungseinschränkungen
- Zusatzinformationen
    - Sprache des Metadaten- und Datensatzes, Rechtliche Grundlage (“Hamburgisches Transparenzgesetz”), Konformität (Tabelle mit Spezifikation, Grad und Datum), Objekt-ID und XML-Link zu den Metadaten

Der Datensatz wird außerdem gelistet im Transparenzportal Hamburg[9] (mit Link zu allen Metadaten als JSON), auf der zusätzlich v.a. Schlagwörter und ältere/neuere Versionen zu finden sind. Das Portal basiert auf CKAN und erlaubt den Zugriff über eine API.[10] Weitere Quellen sind das European Data Portal[11] und GovData.[12] Eine Bearbeitung ist auf dem ESRI-Portal zu finden.[13]

6.3.5 Datensatz 133 “Grundwassergleichen Max 2008”

Der Datensatz der Kategorie “Wasserstraßen und Gewässer” der Behörde für Umwelt und Energie (BUE), Amt für Immissionsschutz und Betriebe der Stadt Hamburg wird unter dl-de/by-2.0 veröffentlicht. Eine kurze Beschreibung benennt die Einheit (“in m NN”), zeitliche Ausprägung (“des hydrologischen Jahres 2008”) und Details (“maximale/höchste Grundwasserstände im 1. Hauptgrundwasserleiter), sowie ein Satz zur Bereitstellung als WMS/WFS. Zusätzlich werden 4 Dateidownloads angeboten. Anhand des Links lassen sich der Dateityp (ZIP) sowie eine Datumsangabe erkennen. Die damit versionierten Archive enhalten eine GML-Datei und ein XSD-Schema.
Wie beim vorhergehenden Datensatz sind auch hier weitere Metadaten auf MetaVer, European Data Portal, GovData und dem Transparenzportal Hamburg zu finden.

6.3.6 Datensatz 49 “Bund: Farbrelief des Wasserlaufs”

Der Datensatz wird durch das Informationstechnikzentrum Bund (ITZBund) unter der GeoNutzV in der Kategorie Wasserstraßen und Gewässer bereitgestellt. Die Beschreibung enthält keine weiteren Informationen. Neben einem WMS- und WFS-Link steht ein ATOM-Feed zur Verfügung. Dieser verweist wiederum auf Subfeeds für einzelne Wasserläufe mit Links auf die eigentlichen Daten (ZIP-Datei mit Höhenmodell des Wasserlaufs).
Weitere Metadaten sind auf GovData[14] (v.a. Veröffentlichungsdatum und letzte Änderung) sowie OpenNRW[15] (ebenfalls Datumsangaben, eine Kontakt-Emailadresse) vorhanden.

6.3.7 Datensatz 600 “VBB-Fahrplan 2013”

Dieser Datensatz der Verkehrsverbund Berlin-Brandenburg GmbH wird in der Kategorie Straßen unter CC-BY-4.0 als Dateidownload bereitgestellt. Die ZIP-Datei enthält mehrere Text-Dateien (Endung .txt, Format ähnelt CSV). Die Beschreibung erläutert, dass die Daten im GTFS-Format vorliegen (mit Link), die Motivation, zeitliche Gültigkeit (bis August 2013) sowie zusätzliche Lizenzierungshinweise.

[9] http://suche.transparenz.hamburg.de/dataset/radwegegis-hamburg
[10] http://transparenz.hamburg.de/hinweise-zur-api/
[11] https://www.europeandataportal.eu/data/en/dataset/af604fb9-a1de-4885-97dd-2af348b385b5
[12] https://ckan.govdata.de/fi/dataset/radwegegis-hamburg2
[13] https://opendata-esri-de.opendata.arcgis.com/datasets/hh-radwege
[14] https://www.govdata.de/daten/-/details/c7c93c36-cca3-11e4-afdc-1681e6b88ec1bkg
[15] https://open.nrw/dataset/c7c93c36-cca3-11e4-afdc-1681e6b88ec1bkg

Weitere Metadaten zum Datensatz liefert das Portal daten.berlin.de:[16]

- weitere Kategorie (Verkehr)
- Geographische Abdeckung (Berlin - nicht ganz korrekt, da der Datensatz auf Brandenburg umfasst)
- Geographische Granularität (Berlin - ebenfalls nicht korrekt)
- Zeitperiode (von - bis)
- Zeitliche Granularität (keine)
- Veröffentlichungs- und Aktualisierungsdatum
- E-Mail-Kontakt und Webseite
- Tags
- Dateinamen innerhalb des ZIP-Archivs

Auch GovData bietet weitere Metadaten.

6.3.8 Datensatz 245 "Urbane Räume: Lufttemperatur und Luftfeuchte stündlich"

Der Datensatz des DWD wird unter der GeoNutzV in der Kategorie Klima und Wetter bereitgestellt. Die kurze Beschreibung "Aktuelle stündliche Lufttemperatur und Luftfeuchte, gemessen an Stadtklimastationen, für ausgewählte urbane Räume in Deutschland" beschreibt grobe Inhalte und Kontext der Daten.
Zugriff auf die Daten besteht über einen FTP-Link. Darin ist lediglich ein Ordner "recent" zu finden, der wiederum 3 Dateien beinhaltet: zwei PDF-Dateien mit der Datensatzbeschreibung (wie bei den Datensätzen 209 und 292) und eine ZIP-Datei. Diese beinhaltet wiederum eine Text-Datei mit den Messwerten sowie eine Excel-Datei, die Stationsdaten der Messstationen enthalten soll, jedoch lediglich einen Eintrag (ohne Identifikator, mit der ein Zusammenhang zu den Messwerten hergestellt werden könnte) enthält.

6.3.9 Datensatz 84 "Digitale Bundeswasserstraßenkarte im Maßstab 1:1.000.000"

Der Datensatz wird von der Generaldirektion Wasserstraßen und Schifffahrt (GDWS) unter der GeoNutzV in der Kategorie Wasserstraßen und Gewässer zur Verfügung gestellt. Der Beschreibungstext entspricht dem Titel. Die Daten werden lediglich als WMS bereitgestellt.
Weitere Metadaten liefert die zugehörige Webseite der Bundesbehörde[17] und der Direktion[18] in weitgehend unstrukturierter Form, sowie GovData (einschließlich detaillierter Beschreibung, Schlagwörtern und Datumsangaben).

6.3.10 Datensatz 606 "VBB-Fahrplandaten August 2017 bis Dezember 2017"

Dieser Datensatz entspricht einer Aktualisierung des Datensatzes 600.

6.4 Zusammenfassung der untersuchten Datensätze

Die folgende Tabelle 8 fasst zusammen, welche zusätzlichen strukturierten und unstrukturierten Metadaten verfügbar sind und wie auf diese zusätzlichen Metadaten zugegriffen werden kann.

[16] https://daten.berlin.de/datensaetze/vbb-fahrplan-2013

[17] https://www.wsv.de/service/karten_geoinformationen/GeodatendiensteGeoanwendungen/geodatendienste/index.html

[18] http://www.gdws.wsv.bund.de/DE/service/karten/02_Geodatendienste_Geoanwendungen/01_Geodatendienste_NEU/Geodatendienste_node.html

Datensatz	Anbieter	Zusätzliche strukturierte Metadaten	Zusätzliche unstrukturierte Metadaten	Zugriffsoption auf zusätzliche Metadaten
209	DWD	keine	ja (PDF auf FTP)	Untersuchung des Datensatz-FTP-Verzeichnisses, Extraktion von PDF-Datei namens *Beschreibung*, ggf. semi-strukturierte Extraktion (vorauss. weitgehend manuell)
292	DWD	keine	ja (PDF auf FTP), ggf. semistruktu-riert auf CDC	siehe oben, bei CDC evtl. über Crawler möglich
435	Bahn	ja (extern)	ja (extern)	Nur über Crawling und Interlinking des Portals data.deutschebahn.com, da kein Direktlink auf die Dataset-Seite der Bahn durch mCLOUD, semi-strukturierte Extraktion
378	Hamburg	ja (extern)	ja (extern)	Crawling von MetaVer, Anbindung des Transparenzportals Hamburg über einen generischen CKAN-Adapter
133	Hamburg	ja (extern)	ja (extern)	wie für 378
49	ITZBund	keine	ja	Crawling der Einträge des Atom-Feeds
600	VBB	keine	ja (extern, semistrukturiert)	Extraktion der Beschreibung, Crawling von daten.berlin.de
245	DWD	keine	ja (PDF auf FTP)	siehe 209
84	GDWS	ja (extern)	ja (auf Webseite)	Crawling der Beschreibung des GDWS, GovData
606	VBB	keine	ja (Beschreibung und extern, semistrukturiert)	Extraktion der Beschreibung, Crawling von daten.berlin.de

Tabelle 8: Zusammenfassung der untersuchten Datensätze

Für praktisch alle untersuchten Datensätze sind umfangreiche Metadaten nur außerhalb von mCLOUD zu finden. In den meisten Fällen stehen diese un- oder semistrukturiert zur Verfügung. Neben fokussierten HTML-Crawlern sind Extraktoren aus PDF, Atom und CKAN-basierten Portalen von Bedeutung.

7 Zusammenfassung und Ausblick

In diesem Bericht wurden die für OPAL relevanten Datenquellen identifiziert und analysiert. Im Folgenden werden die Erkenntnisse der Fragestellungen, die in Abschnitt 2 erarbeitet wurden, zusammengefasst.

1. *Welche Datenquellen sollen im Rahmen von OPAL betrachtet werden?*

Eine Liste von Datenportalen wurde in Abschnitt 3 zusammengestellt. Eine Übersicht ist in Tabelle 9 dargestellt.

2. *Wie viele Datensätze enthalten die Datenquellen?*

Die Datenquellen wurden in Abschnitt 5 untersucht. Die Anzahl der Datensätze wird in Tabelle 9 zusammengefasst.

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

Tabelle 9: Anzahl der Datensätze der betrachteten primären Datenquellen

3. *Welche Zugriffsmöglichkeiten bietet die Datenquellen …*
   a. *… hinsichtlich Zugriffsprotokollen?*

   Alle Datenquellen erlauben den Zugriff über HTTP/1.1. Mit Ausnahme von MDM sind die Webseiten statisch, sodass ein fokussierter Crawler zum Einsatz kommen kann. Bei MDM muss entsprechend eine Technologie zum Einsatz kommen, die die Verarbeitung der jQuery-Skripte ermöglicht. Weiterhin ist bei GovData, OffeneDaten.de, EDP und einigen sekundären Portalen wie daten.berlin.de ein effizienter Zugriff über die CKAN-API möglich.

   b. *… hinsichtlich Datenrepräsentationen (z.B. Datenformate, Strukturiertheit der jeweiligen Metadaten, Verfügbarkeit von Identifikatoren wie URIs)?*

   Mit Ausnahme der Portale, die einen Zugriff über die CKAN-API anbieten (als JSON), liegen die Metadaten semistrukturiert als HTML vor. Identifikatoren müssen meist aus der URL abgeleitet oder neu festgelegt (minted) werden.

   c. *… hinsichtlich Interaktionsmustern (z.B. Request-Response, Publish-Subscribe)?*

Alle Datenquellen erfordern Pull-Zugriff (Request-Response). MDM ermöglicht teilweise Publish-Subscribe-Zugriff auf abonnierte Daten, das Abonnement erfordert jedoch eine vorherige Kommunikation per E-Mail.

   d. *… hinsichtlich direktem Zugriff oder indirekter Einbindung mittels Katalogsystem (im Fall von Meta-Datenportalen wie z.B. mCLOUD)?*

Die meisten betrachteten primären Quellen sind Meta-Datenportale.

4. *Welche Metadaten können für die Datenquellen angegeben werden, und wie häufig enthalten die Datensätze die jeweiligen Metadaten statistisch?*

Details zu mCLOUD und MDM sind in Abschnitt 5 zu finden.

5. *In welchem Format liegen die jeweiligen Metadaten vor (z.B. Struktur, Einheitlichkeit, Identifikatoren)?*

Einzelne Metadaten wie Lizenz, Bereitsteller und Titel sind einheitlich vorhanden. Viele andere Metadaten müssen aufwändig aus semi- und unstrukturierten, meist externen Quellen extrahiert und angebunden werden. Als Identifikatoren für das Interlinking zu diesen externen Quellen bietet sich die Ressourcen-URL (z.B. Download-Link oder WMS/WFS-URL) an.

6. *Auf wieviele Anbieter verteilen sich die Datensätze einer Datenquelle?*

Für mCLOUD und MDM wurde diese Frage in Abschnitt 5 beantwortet. Beide Datenquellen zeichnet eine hohe Heterogenität der Anbieter aus.

7. *Wie veränderlich ist die Datenquelle (Häufigkeit/Umfang der Änderung der Anzahl von Datensätzen)?*

Siehe 8.

8. *Wie veränderlich sind die Metadaten der Datensätze innerhalb der Datenquelle….*
   a. *…hinsichtlich durchschnittlicher Aktualisierunghäufigkeit?*
   b. *…hinsichtlich welcher Metadatenattribute?*
   c. *…hinsichtlich der Veränderlichkeit des Aktualisierungsintervalls (z.B. immer wöchentlich oder in nicht vorhersagbaren Zeitabständen)?*

Diese Fragen müssen detailliert in Arbeitspaket 2 und 3 beantwortet werden. Die initiale Datenanalyse lässt die Abschätzung zu, dass ein Großteil der Daten statisch ist, jedoch bei vielen Datensätzen in schwer abschätzbaren Zeitabständen Aktualisierungen vorgenommen und auch neue Datensätze zu verschiedenen Zeitpunkten hinzukommen. Ein kontinuierliches Crawling mit adaptiver Wartezeit zwischen Recrawls ist voraussichtlich sinnvoll.

Die Ergebnisse der Datenanalyse, die in diesem Bericht zusammengefasst sind, dienen als Grundlage der Architekturspezifikation (A1.3), Datenextraktion (AP2) und Datenanalyse (AP3). Die erfassten Metadaten (z.B. Schemata von GovData) dienen als Basis der Vokabularspezifikation und Konvertierung in AP4. Zukünftig sind einzelne Aspekte, z.B. relevante Identifikatoren, hinsichtlich der Datenverknüpfung und Fusion in AP5 von Bedeutung.