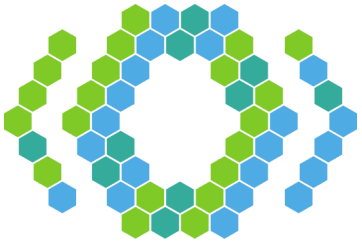

OPEN DATA PORTAL

Deliverable D3.3
Erste Metadatenextraktionskomponente

Autoren: Adrian Wilke
Reviewer: Michael Röder

Veröffentlichung	Öffentlich
Fälligkeitsdatum	30. April 2019
Fertigstellung	7. Mai 2019
Arbeitspakete	A3.2 Semistrukturierte Metadatenextraktion A3.3 Unstrukturierte Metadatenextraktion
Typ	Software
Status	Final
Version	1.0

Kurzfassung:
Die erste Version der OPAL Komponente zur Metadatenextraktion umfasst die Spracherkennung von Texten, die Eigennamenerkennung von Entitäten (Named Entity Recognition, NER) und die anschließende Disambiguierung. Außerdem werden Lizenzen für einzelne Distributionen von Metadaten extrahiert.

Schlagworte:
Metadaten, Extraktion, Entitäten, Eigennamenerkennung, Disambiguierung, Spracherkennung, Lizenzen, FOX, AGDISTIS

Inhalt

Bundesministerium für Verkehr und digitale Infrastruktur

Metadatenextraktion von Entitäten **2**
- Spracherkennung 2
- Eigennamenerkennung und Disambiguierung von Entitäten 2
- OPAL Webservices zur Metadatenextraktion 3
- Anwendungsbeispiel 4
  - Spracherkennung 4
  - Eigennamenerkennung und Disambiguierung von Entitäten 4

Metadatenextraktion von Lizenzen **5**
- Datenschema: Einordnung von Lizenzinformationen 5
- Übersicht extrahierter Lizenz-URIs 6

1 Metadatenextraktion von Entitäten

Die Extraktion von Entitäten ist dem Arbeitspaket *A3.3 Unstrukturierte Metadatenextraktion* zugeordnet. Die zugehörige OPAL Softwarekomponente zu diesem Arbeitspaket wurde als OPAL Metadata-Extraction[1] veröffentlicht.

Zur Anreicherung von Daten wird eine unterliegende Datenbasis analysiert und die resultierenden Ergebnisse dem ursprünglichen Datenmodell zurückgespiegelt. Die unterliegende Datenbasis in OPAL setzt sich mehrheitlich aus Metadaten offener Datensätze zusammen. Dabei wird für einzelne Bestandteile der Metadaten natürlichsprachlicher Text verwendet. Dies ist beispielsweise in Beschreibungstexten der Fall, die den Inhalt der referenzierten Datensätze zusammenfassen. Diese Beschreibungstexte enthalten häufig Entitäten, z.B. Städtenamen. Befinden sich solche Entitäten in weiteren Datensätzen, so können sie verwendet werden, um Benutzern eines Informationssystems Vorschläge zu weiteren Datensätzen zu unterbreiten, die ebenfalls interessant sein könnten. Hat ein Benutzer beispielsweise einen Datensatz zu einer bestimmten Stadt gefunden, interessiert er sich möglicherweise für weitere Datensätze zur selben Stadt. Die entwickelte Komponente zur Erkennung und Extraktion von Entitäten kann also als Grundlage dazu dienen, Vorschläge für Datensätze zu generieren, die die Auffindbarkeit von Daten verbessern.

Die Beschreibung der Komponente zur Metadatenextraktion setzt sich aus vier Teilen zusammen. Die (1) Spracherkennung dient dazu, ein Erkennungs-Modell auszuwählen, das für die jeweiligen Sprache erstellt wurde. Anschließend werden (2) Entitäten gesucht und eindeutige Bezeichner für diese zurückgegeben. Diese beiden Komponenten können über (3) Webservices zugegriffen werden. Hierzu wird ein (4) Anwendungsbeispiel angeführt.

1.1 Spracherkennung

Zur Erkennung, in welcher Sprache ein Text verfasst wurde, ist in der OPAL Metadatenextraktion die Software Apache OpenNLP[2] eingebunden. Im Zuge dessen wird das Language Detector Model[3] in Version 1.8.3 eingesetzt. Damit können 103 Sprachen erkannt werden. Diese werden im Modell durch das ISO 639-3[4] Format zugeordnet. In OPAL wird übergreifend das ISO 639-1 Format verwendet, welches im Data Catalog Vocabulary DCAT[5] empfohlen wird. Daher erfolgt intern eine Übersetzung zwischen den beiden ISO 639 Standards.

1.2 Eigennamenerkennung und Disambiguierung von Entitäten

Die in OPAL fokussierten Metadaten beschreiben Mobilitätsdaten. Diese beziehen sich häufig auf eine Region, z.B. eine Stadt. Die Metadaten enthalten als Beschreibung unstrukturierte, natürlichsprachliche Texte, die wiederum Entitäten beinhalten. Beispiele für Entitäten sind Städte oder Bundesländer, für die weitere Daten (z.B. Geo-Koordinaten) abgerufen werden können. Zur Erkennung von Eigennamen (z.B. Paderborn) wird in OPAL die Software FOX[6] eingesetzt. Die anschließende Disambiguierung (die Erkennung von eindeutigen Bezeichnern) von Entitäten (z.B. http://de.dbpedia.org/resource /Paderborn) geschieht über die Einbindung der Software AGDISTIS[7]. Die OPAL Komponenten zur Metadatenextraktion sind über Webservices aufrufbar.

[1] OPAL Metadata-Extraction: https://github.com/projekt-opal/metadata-extraction
[2] Apache OpenNLP: https://opennlp.apache.org/
[3] Language Detector Model: https://opennlp.apache.org/models.html
[4] ISO 639-3: https://iso639-3.sil.org/
[5] ISO 639-1 Verweis in DCAT: https://www.w3.org/TR/vocab-dcat/#Property:catalog_language
[6] FOX - Federated knOwledge eXtraction Framework: https://dice.cs.upb.de/projects/active-projects/fox/
[7] AGDISTIS - Agnostic Named Entity Disambiguation: https://github.com/dice-group/AGDISTIS

1.3 OPAL Webservices zur Metadatenextraktion

In der folgenden Tabelle sind die entwickelten Webservices zur Metadatenextraktion beschrieben. Dies umfasst eine Kurzbeschreibung der Webservices, die benötigten Eingabeparameter und jeweils ein Beispiel zum Aufruf der Webservices.

Spracherkennung	
Beschreibung:	Gibt die erkannte Sprache im Format ISO 639-3 zurück.
Parameter:	Volltext
Beispiel:	http://opal/metadata/lang?text=Sprachen lernen
Beschreibung:	Erkennt die verwendeten Sprachen und fügt sie den OPAL RDF-Datensätzen hinzu.
Parameter:	RDF-Modell im TURTLE-Format
Beispiel:	http://opal/metadata/lang/model?turtleBytes=...
Eigennamenerkennung und Disambiguierung von Entitäten	
Beschreibung:	Gibt das FOX Ergebnis im Format RDF-TURTLE zurück.
Parameter:	Volltext, Sprache
Beispiel:	http://opal/metadata/fox?text=A. Einstein was born in Ulm.&lang=en
Beschreibung:	Gibt die natürlichsprachlichen Namen von Orten im JSON Format zurück.
Parameter:	Volltext, Sprache
Beispiel:	http://opal/metadata/fox/location/names?text=Paderborn und Bad Oeynhausen sind in NRW&lang=de
Beschreibung:	Gibt die URIs von Orten im JSON Format zurück.
Parameter:	Volltext, Sprache
Beispiel:	http://opal/metadata/fox/location/uris?text=Paderborn und Bad Oeynhausen sind in NRW&lang=de

Tabelle: Entwickelte OPAL Webservices zur Metadatenextraktion

1.4 Anwendungsbeispiel

Ein Anwendungsbeispiel zur Metadatenextraktion ist der folgende Beispieltext, der stellvertretend für eine Beschreibung eines Datensatzes verwendet wird.

Beispieltext:
```
Beipielanfrage zu Paderborn und Bad Oeynhausen in NRW
```

1.4.1 Spracherkennung

Für den Beispieltext wird zunächst die verwendete Sprache bestimmt.

Aufruf Webservice Spracherkennung:
```
http://opal/metadata/lang?text=Beipielanfrage%20zu%20Paderborn%20und%20Bad%20Oeynhausen%20in%20NRW
```
Resultat:
```
deu
```

Die Rückgabe ist ein 3-stelliger CODE im ISO 639-3 Format und repräsentiert die Sprache Deutsch.

1.4.2 Eigennamenerkennung und Disambiguierung von Entitäten

Die Metadatenextraktion der Geo-Entitäten geschieht über den folgenden Aufruf.

Webservice:
```
http://opal/metadata/fox/location/names?text=Beipielanfrage%20Paderborn%20und%20Bad%20Oeynhausen%20in%20NRW&lang=de
```
Resultat:
```
["Bad Oeynhausen","NRW","Paderborn"]
```

In diesem Beispiel werden eine Stadt bestehend aus einem Wort, eine Stadt aus zwei Wörtern und die Abkürzung eines Bundeslandes erkannt. Eine weitere Variante mit Disambiguierung zeigt der folgende Aufruf:

Webservice:
```
http://opal/metadata/fox/location/uris?text=Beipielanfrage%20Paderborn%20und%20Bad%20Oeynhausen%20in%20NRW&lang=de
```
Resultat:
```
["http://de.dbpedia.org/resource/Bad_Oeynhausen",
"http://de.dbpedia.org/resource/NRW","http://de.dbpedia.org/resource/Paderborn"]
```

Bei diesem Aufruf werden URIs der Entitäten zurückgegeben. Diese sind eindeutige Bezeichner für die jeweiligen Entitäten. Dadurch können sie weiterverwendet werden, um z.B. entsprechende Geokoordinaten zu ermitteln. Auf der DBpedia Seite zur Ressource Paderborn[8] findet sich eine Übersicht abrufbarer Metadaten.

[8] DBpedia Ressource Paderborn: http://dbpedia.org/page/Paderborn

2 Metadatenextraktion von Lizenzen

Die Extraktion von Lizenzinformationen ist dem Arbeitspaket *A3.2 Semistrukturierte Metadatenextraktion* zugeordnet.

2.1 Datenschema: Einordnung von Lizenzinformationen

Der Extraktion von Metadaten aus offenen Datenportalen, die für OPAL genutzt werden, schließt sich eine Datenspeicherung im RDF-Format an. Hierzu wird u.a. das DCAT Vokabular[9] genutzt. Datenportale entsprechen dabei dem Typ `dcat:Catalog`. Die dort enthaltenen Metadatensätze werden in OPAL als `dcat:Dataset` abgebildet. Diese Datensätze können u.a. in verschieden Datenformaten (z.B. PDF, XLSX, CSV, RDF-TURTLE) zugänglich gemacht werden. Solche Daten-Versionen werden als Distribution bezeichnet und als `dcat:Distribution` klassifiziert. Für Distributionen kann über den Bezeichner `dct:license` eine URI angegeben werden, die auf die jeweils verwendete Lizenz verweist. Die folgende Abbildung veranschaulicht das verwendete DCAT Vokabular:

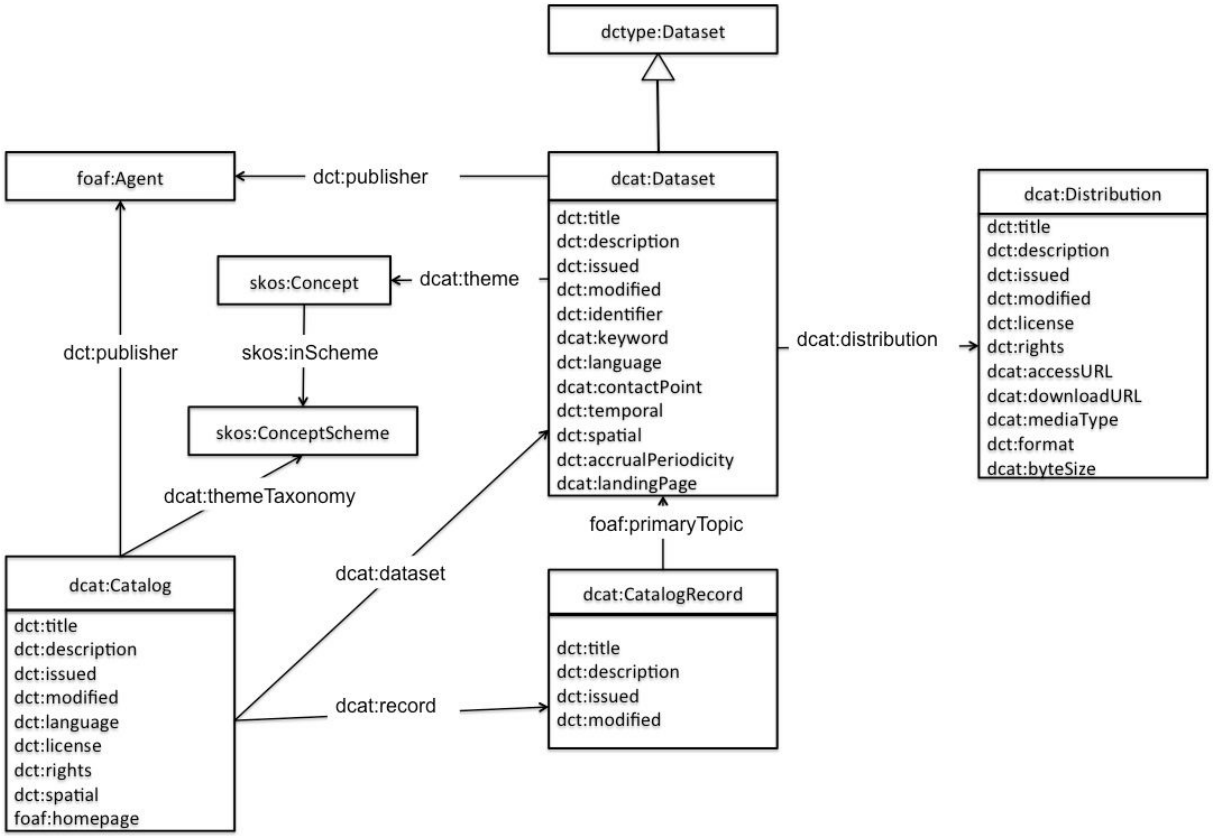

Abbildung: DCAT Vokabular
(Quelle: https://www.w3.org/TR/vocab-dcat/)

[9] DCAT: https://www.w3.org/TR/vocab-dcat/

2.2 Übersicht extrahierter Lizenz-URIs

Für die Datenportale mCLOUD[10] und European Data Portal[11] wurden Lizenz-URIs für 2.798 und 61.173 einzelne Distributionen extrahiert. Die Analyse der Lizenzen geschah über Abfragen in der Sprache SPARQL und entsprechendem Programmcode[12]. Die folgende Tabelle zeigt eine Übersicht der verwendeten Lizenzen:

Lizenz URI	Anzahl Distribu-tionen
http://europeandataportal.eu/content/show-license?license_id=OGL2.0	34.911
https://www.europeandataportal.eu/content/show-license?license_id=OGL2.0	10.645
http://europeandataportal.eu/content/show-license?license_id=FR-LO	9.493
https://www.europeandataportal.eu/content/show-license?license_id=ODC-ODbL	2.866
http://europeandataportal.eu/content/show-license?license_id=ODC-ODbL	1.420
https://www.europeandataportal.eu/content/show-license?license_id=FR-LO	1.293
file:///content/show-license?license_id=OGL2.0	870
http://www.gesetze-im-internet.de/geonutzv/index.html	758
https://www.govdata.de/dl-de/by-2-0	671
https://creativecommons.org/licenses/by/3.0/de/	562
https://creativecommons.org/publicdomain/zero/1.0/deed.de	244
https://creativecommons.org/licenses/by/4.0/deed.de	193
http://www.opendefinition.org/licenses/cc-by	140
http://www.stadtentwicklung.berlin.de/geoinformation/download/nutzIII.pdf	94
http://europeandataportal.eu/content/show-license?license_id=DL-DE-BY2.0	76
https://www.govdata.de/dl-de/zero-2-0	75
file:///content/show-license?license_id=FR-LO	67
file:///content/show-license?license_id=ODC-ODbL	41
file:///content/show-license?license_id=CC-BY4.0	40
http://www.ioer-monitor.de/fileadmin/Dokumente/PDFs/Nutzungsbedingungen_IOER-Monitor.pdf	30
file:///content/show-license?license_id=DL-DE-BY2.0	16
https://scihub.copernicus.eu/twiki/pub/SciHubWebPortal/TermsConditions/TC_Sentinel_Data_31072014.pdf	16
http://europeandataportal.eu/content/show-license?license_id=GFDL-1.3	13
https://www.europeandataportal.eu/content/show-license?license_id=CC-BY4.0	12
http://europeandataportal.eu/content/show-license?license_id=ODC-BY	7
http://europeandataportal.eu/content/show-license?license_id=CC-BY4.0	6

[10] mCLOUD: https://www.mcloud.de/
[11] European Data Portal: https://www.europeandataportal.eu/
[12] Metadatenextraktion für Statistiken: https://github.com/projekt-opal/statistics

http://opendatacommons.org/licenses/odbl/summary/	6
file:///content/show-license?license_id=DL-DE-BY1.0	5
file:///content/show-license?license_id=ODC-BY	4
http://www.opendefinition.org/licenses/cc-zero	3
http://www.opendefinition.org/licenses/odc-odbl	3
https://www.europeandataportal.eu/content/show-license?license_id=ODC-BY	3
file:///content/show-license?license_id=GFDL-1.3	2
http://europeandataportal.eu/content/show-license?license_id=PSEUL	2
http://images.vbb.de/assets/downloads/file/20928.pdf	1
http://www.opendefinition.org/licenses/cc-by-sa	1
https://www.europeandataportal.eu/content/show-license?license_id=GFDL-1.3	1
https://www.vrsinfo.de/fileadmin/Dateien/api/NutzervereinbarungODOS.pdf	1
Gesamt	64.591

Tabelle: Übersicht extrahierter Lizenzen

Für diese extrahierten Lizenz-URIs besteht in den Arbeitspaketen im weiteren Projektverlauf die Möglichkeit, eine erweiterte Datenaufbereitung durchzuführen. Es bietet sich eine Disambiguierung der Lizenzen und der verwendeten URIs an. Die DCAT-Erweiterung DCAT-AP.de [13] bietet in Version 1.0 eine Klassifizierung von 33 Lizenzen an, in denen angegeben ist, ob der jeweils verwendete Lizenztyp offen oder geschlossen ist.
DCAT-AP.de wurde beim 4. Treffen des WIK/mFUND Arbeitsforums Standardisierung/mCLOUD[14] am 20. März 2019 in der Präsentation zum Datenportal GovData vorgestellt. Die Veranstaltung fand beim Deutschen Zentrum für Luft- und Raumfahrt (DLR) in Bremen und mit Unterstützung vom Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) statt.

Weitere Dokumente zum Projekt OPAL sind auf der Webseite verfügbar:
http://projekt-opal.de/en/results/deliverables/

[13] DCAT-AP.de: https://www.dcat-ap.de/def/
[14] WIK/mFUND Arbeitsforum Standardisierung/mCLOUD: https://www.wik.org/index.php?id=943