



# Beschreibung der Importschnittstelle

Dieses Dokument beschreibt die Abfragen der verschiedenen Schnittstellen im Harvesting Prozess der mCLOUD.

Die Abbildung der Daten aus der Schnittstellen Antwort ist in der mCLOUD Mapping Übersicht beschrieben<sup>1</sup>.

---

<sup>1</sup> <https://mcloud.de/web/guest/informationen#documents>

## CKAN

Der CKAN Harvester verwendet die API Version 3 von CKAN<sup>2</sup>.

Zunächst wird über den `package_list` Request die Anzahl der verfügbaren Datensätze ermittelt. Das Harvesting erfolgt dann entweder über den `current_package_list_with_resources` oder den `package_search` Request. Die bevorzugte Variante ist der `package_search` Request.

### Verwendete Endpoints

- `/api/3/action/package_list`
- `/api/3/action/current_package_list_with_resources`
- `/api/3/action/package_search`

### Verwendete Parameter

`package_search`:

- `rows`: Kann konfiguriert werden, Standard-Wert ist 100
- `start`: Anfangswert kann konfiguriert werden, erhöht sich immer um den Wert von `rows`
- `fq`: Filterung auf Groups, Tags oder Ids kann konfiguriert werden. Weitere Filter können frei eingegeben werden.

`current_package_list_with_resources`:

- `limit`: Kann konfiguriert werden, Standard-Wert ist 100
- `offset`: Anfangswert kann konfiguriert werden, erhöht sich immer um den Wert von `rows`

### Beispiel Request

```
https://data.deutschebahn.com/api/3/action/package_search?  
rows=100&start=0
```

---

<sup>2</sup> <http://docs.ckan.org/en/ckan-2.7.3/api/>

## CSW

CSW Schnittstellen werden über den GetRecords Request geharvested. Die Anfrage kann über HTTP-POST oder HTTP-GET erfolgen.

Als OutputSchema wird <http://www.isotc211.org/2005/gmd> verwendet.

## Verwendete Parameter

- maxRecords
- startPosition
- Query/Constraint/Filter

## Beispiel Anfrage per POST-Request

```
<csw:GetRecords xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
  xmlns="http://www.opengis.net/cat/csw/2.0.2" xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:ogc="http://www.opengis.net/ogc" xmlns:dct="http://purl.org/dc/terms/"
  xmlns:ows="http://www.opengis.net/ows" xmlns:gml="http://www.opengis.net/gml"
  xmlns:gmd="http://www.isotc211.org/2005/gmd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" service="CSW" version="2.0.2"
  maxRecords="50" startPosition="1" resultType="results" outputFormat="application/xml"
  outputSchema="http://www.isotc211.org/2005/gmd"
  xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2 http://schemas.opengis.net/csw/2.0.2/
CSW-discovery.xsd">
  <csw:DistributedSearch/>
  <csw:Query typeNames="gmd:MD_Metadata">
    <csw:ElementSetName>full</csw:ElementSetName>
    <Constraint version="1.1.0">
      <ogc:Filter>
        <ogc:PropertyIsEqualTo>
          <ogc:PropertyName>subject</ogc:PropertyName>
          <ogc:Literal>opendata</ogc:Literal>
        </ogc:PropertyIsEqualTo>
      </ogc:Filter>
    </Constraint>
  </csw:Query>
</csw:GetRecords>
```

## DCAT-AP.de

Das Harvesting von DCAT-AP.de erfolgt über einen DCAT Katalog. Dieser sollte den Spezifikationen auf DCAT-AP.de<sup>3</sup> entsprechen. Über die Katalog-Url können zusätzliche Parameter angegeben werden.

### Paging

Für ein Harvesting über mehrere Seiten eines DCAT Katalog wird eine PagedCollection mit den Angaben nextPage und lastPage benötigt.

### Verwendete Parameter

- page: Wird aus der PagedCollection der vorherigen Antwort befüllt

### Beispiel Request

<https://www.govdata.de/ckan/catalog/catalog.rdf>

### Beispiel PagedCollection

```
<hydra:PagedCollection rdf:about="https://ckan.govdata.de/catalog.rdf?page=1">
  <hydra:nextPage>https://ckan.govdata.de/catalog.rdf?page=2</hydra:nextPage>
  <hydra:totalItems rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">45063</
hydra:totalItems>
  <hydra:itemsPerPage rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">100</
hydra:itemsPerPage>
  <hydra:lastPage>https://ckan.govdata.de/catalog.rdf?page=451</hydra:lastPage>
  <hydra:firstPage>https://ckan.govdata.de/catalog.rdf?page=1</hydra:firstPage>
</hydra:PagedCollection>
```

---

3 <https://www.dcat-ap.de/def/>

## OAI Protocol for Metadata Harvesting

### (OAI-PMH)

Ein Harvesting über das OAI Protocol ist über den ListRecords Aufruf möglich.  
Als Format muss vom Server „iso19139“ unterstützt werden.

Auswahl über Parameter „Set“

Paging über resumptionToken

### Verwendete Parameter

- verb: ListRecords
- metadataPrefix: iso19139
- set: Konfigurierbar
- resumptionToken: Wird aus vorheriger Antwort übernommen

### Beispiel Request

```
https://oai.dwd.de/oai/provider?  
verb=ListRecords&metadataPrefix=iso19139&set=mCLOUD
```

## SPARQL

Für ein Harvesting per SPARQL muss eine SPARQL Query angegeben werden, die Werte aus dem Datenbestand auf mCLOUD Datenfelder abbildet

Als mCLOUD Datenfelder stehen zur Verfügung:

- id
- title
- description
- publisher
- distribution\_url
- distribution\_title
- source\_link
- issued
- modified
- license



## Beispiel Anfrage

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dataid: <http://dataid.dbpedia.org/ns/core#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dcatx: <http://www.example.org/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?id ?publisher ?source_link ?title ?description ?distribution_title
?distribution_url ?issued ?modified ?license
(GROUP_CONCAT(DISTINCT ?keyword; SEPARATOR=", ") AS ?keywords)
(GROUP_CONCAT(DISTINCT ?dcat_theme; SEPARATOR=", ") AS ?dcat_theme)
WHERE {
    ?limbo_ds dataid:group ?group;
        dataid:artifact ?artifact;
        dct:identifier ?id;
        dcat:distribution ?limbo_dist;
        dct:issued ?issued;
        dct:license ?license;
        dct:publisher ?publisher;
        prov:wasDerivedFrom ?mcloud_dist.

    ?limbo_dist dcat:downloadURL ?distribution_url ;
        dcatx:localId ?distribution_title.

    ?mcloud_ds dcat:distribution ?mcloud_dist ;
        dct:title ?mcloud_title;
        dct:description ?mcloud_description;
        dcat:keyword ?keyword;
        dcat:theme ?dcat_theme;
        dcat:landingPage ?mcloud_landingPage .

    BIND (CONCAT("Limbo-Daten f&uuml;r <a href=\"",?mcloud_landingPage , "\">",?mcloud_title,"</a>
\r\n\r\nOriginal:\r\n",?mcloud_description) as ?description).
    BIND (CONCAT("LIMBO - ",?mcloud_title) as ?title).
    BIND (?limbo_ds as ?source_link)
    BIND (?issued as ?modified)

FILTER (?group = 'org.limbo')
}
GROUP BY ?id ?publisher ?source_link ?title ?description ?distribution_title ?distribution_url ?
issued ?modified ?license
```