# The European Register for Protected Data held by the Public Sector on data.europa.eu: Harvesting guidelines

Version 1.3

**Version control table**

| Version | Modified by | Modifications made | Release date |
|---|---|---|---|
| 1.0 | Leonard Mack (Fraunhofer FOKUS); Michal Kuban (DG CNECT, European Commission) | Initial release of the document. | 26/05/2023 |
| 1.1 | Leonard Mack (Fraunhofer FOKUS); Michal Kuban (DG CNECT, European Commission) | Added further clarifications to section 4.<br><br>Added further mandatory metadata property dct:distribution to dataset class in section 4.1.<br><br>Corrected usage note of dct:rights in section 4.2.<br><br>Amended CKAN metadata model description in section 5.2.2 according to changes in section 4.1. | 07/06/2023 |
| 1.2 | Torben Jastrow (Fraunhofer FOKUS); Simon Dutkowski (Fraunhofer FOKUS) | Replaced example in section 5.1.3. | 03/07/2023 |
| 1.3 | Leonard Mack (Fraunhofer FOKUS); Svetlana Marchenko (Fraunhofer FOKUS); | Changed „European Single Access Point" to "European Register for Protected Data held by the Public Sector" and "ESAP" to "ERPD" in line with revised official terminology. | 29/08/2023 |

# Table of Contents

# 1. Introduction

The Data Governance Act[1] (hereinafter 'DGA') is a regulation of the European Parliament and the Council to facilitate data sharing within the EU's Single Market. It came into force on 23rd June 2022 and covers private companies, citizens, and public sector bodies. The Act will be applicable from 24th September 2023. The Data Governance Act includes mechanisms to foster the reuse of public sector data that, for certain reasons, cannot be made available as open data. This could be health or mobility data that are, according to Article 3 of the Data Governance Act, protected on grounds of statistical or commercial confidentiality, intellectual property rights of third parties, or privacy.

The most basic lever to promote data reuse is simply to increase its discoverability, enabling potential users to understand which data public sector institutions hold. For this, the publication of metadata containing information about existing datasets is required. Accordingly, Article 8 of the Data Governance Act instructs Member States to establish so-called **National Single Information Points (NSIPs)**. NSIPs are intended to serve as national one-stop-shop allowing any user, such as citizens, entrepreneurs, or researchers, to search for and find information about the affected public sector data in their Member States. Because the relevant data, such as health or mobility data, is protected, users will not be able to readily access that same data. Instead, they will have to issue an access request. Therefore, NSIPs must not just enable users to search for protected data, but they must also offer means for users to understand how they can gain access and, if applicable, offer a direct procedure to issue access requests to the relevant public sector bodies.

**The title "European Register for Protected Data held by the Public Sector" (abbreviated as ERPD) will replace the previous title "European Single Access Point" (ESAP) (referred to in Article 8 (4) of the Data Governance Act) with immediate effect**. This decision has been taken to pre-empt any confusion with another upcoming legislation, also foreseeing a European Single Information Point for publicly available information financial services, capital markets, and sustainability.

Additionally, the DGA also specifies that the European Commission shall establish a **European Register for Protected Data held by the Public Sector (ERPD)** to collect, partially mirror, and render discoverable the NSIPs' data in a European register. According to recital 26 of the DGA, NSIPs and the ERPD can be implemented either as entirely new infrastructures or as part of existing (open) data infrastructures.

This document provides an overview of how the ERPD will be implemented by the European Commission. For this, it specifies how metadata must be provided and structured by the NSIPs to interface with the ERPD. Member States may also draw from this document their own conclusions on how to align, by design, their own NSIP with the ERPD, thus facilitating harmonisation across NSIPs. However, the primary purpose of these harvesting guidelines is to formulate specific requirements for the unidirectional harvesting of metadata from NSIPs into the ERPD – and to advise Member States on how to establish their NSIP accordingly.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R0868&from=EN

## 2. The European Register for Protected Data held by the Public Sector on data.europa.eu

Based on a common decision by the Publications Office and DG CNECT, **the European Register for Protected Data held by the Public Sector will be integrated into and implemented as part of data.europa.eu**. This means that the ERPD will leverage the existing infrastructure and general approach to data harvesting and management. However, the advantages of integrating the ERPD into data.europa.eu go well beyond mere technicalities: This approach allows users to easily search and find metadata relating to open data (from open data portals, already available on data.europa.eu) and non-open data (from NSIPs, newly added) in one place, fostering data uptake and reuse along a single value chain. Data.europa.eu offers rich, multilingual search features, it is well known in the community thanks to its manyfold activities around data, and it is constantly with new, useful features added on an ongoing basis. This means that the ERPD and ERPD users will benefit from a head start and from future developments of data.europa.eu.

The integration of the ERPD in data.europa.eu can go also go ahead with relative ease, offering advantages to NSIPs and data.europa.eu. The implementation on data.europa.eu will ensure that all NSIP data included in the ERPD are arranged in a dedicated catalogue structure and is searchable via filters. This ensures that, on the portal, the non-open NSIP data can be separated from other open data. From the perspective of data publishers, the implementation along the lines of data.europa.eu's existing operating model offers various advantages, too. In particular, the existing harvesting approach and overall data architecture can be adapted relatively easily. This means that relevant national data publishers can build on their experience with data.europa.eu's proven, production-level system rather than having to adapt to an entirely new development.

The Data Governance Act does not provide concrete technical instructions for the implementation of the NSIPs and the ERPD, neither on the technology or data level. The harvesting guidelines presented here are therefore based on an interpretation of the relevant articles of the Data Governance Act.[2]

The harvesting guidelines presented in this document build largely on the DCAT-AP data schema. Over the past decade, DCAT-AP has become the preferred choice for EU Member States to describe their open data. It is a well-documented, comprehensive, and flexible data schema that is the most convenient choice for data publishers and users, thanks to its widespread adoption and high interoperability. All required metadata presented in section 4 of these guidelines can be represented via existing DCAT-AP properties. Differences are only present in the hierarchy of required metadata.[3]

In keeping with data.europa.eu's general data governance, the harvesting guidelines for the European Register for Protected Data held by the Public Sector demand that metadata is provided in a of DCAT-AP-compliant manner. Metadata that can be identified as being mandatory from the Data Governance Act are therefore mapped into DCAT-AP properties and must be structured correspondingly by NSIPs to

[2] This concerns in particular Articles 5-8 of the Data Governance Act.

[3] On the level of required metadata, precisely four additional metadata properties are added. These include information on publishers and conditions for the re-use of data (both relating to the dataset metadata level; dct:publisher and dct:rights) as well as information on individual the format and size of individual distributions (dct:format and dcat:byteSize). For further information, please see section 4.

enable harvesting (as laid out in sections 4 and 6). This relates to the requested information on the titles of datasets, their descriptions, their publisher, conditions for reuse and access procedure, format, and size. NSIPs are also required to separate provided metadata into datasets and distributions, in line with the DCAT-AP logic. A dataset can have multiple distributions, e.g. offering a file with the same data in either XML, JSON, or CSV.

The ERPD data architecture follows data.europa.eu's known hierarchical structure with minor modifications. Once harvested, the NSIP metadata on datasets and distributions are automatically grouped into country specific NSIP catalogues in the ERPD. These catalogues are referred to as NSIP country catalogues and are labelled according to the following schema "*Country* National Single Information Point". This means that for each country, only one NSIP catalogue exists, e.g. "Italy National Single Information Point".[4] Furthermore, to ensure a consistent classification of NSIP data aside of data.europa.eu's open-data-related metadata, all NSIP data are sorted into a meta-catalogue labelled "European Register for Protected Data held by the Public Sector". This data structure allows users to target their search and retrieve metadata either across the entire range of the ERPD (i.e. discovering data across all NSIPs) or limiting their search to only one NSIP. Furthermore, users will also be able to search across the entire data range of data.europa.eu, locating both open data and ERPD data alongside each other.

The following sections provide detailed technical instructions on the required metadata, formatting, as well as supported interfaces.

[4] The publisher of this catalogue will automatically be the relevant NSIP. Unlike in the case of data.europa.eu's existing open data registry, there is no option to create catalogues for individual institutions below the NSIP level (e.g. a specific catalogue for the statistical office of a given country).

# 3. Requirements

The following requirements specify which and how data must be provided the NSIPs to enable harvesting by the ERPD.

## 3.1 Segregation of NSIP data

As stated before, the Data Governance Act states in recital 26 that MS could either implement NSIPs as entirely new, standalone infrastructures or that they might repurpose the existing open data (portal) infrastructure to also register the metadata on non-open public sector data.

The requirements stated in this document do not favour either solution. It is evident that, if an NSIP is implemented as a standalone solution, the metadata from that NSIP is delivered by a standalone endpoint as well.

However, if your NSIP is implemented as part of an existing (national) open data infrastructure, it is essential that you offer a means by which NSIP data can be separated from other, open data on the same portal. For this, your NSIP must offer:

- A dedicated endpoint exclusively for the harvesting of NSIP data, **OR**
- A filter mechanism for the endpoint that allows the retrieval of NSIP data only. For the avoidance of doubt, this implies that any filtering must take place at the endpoint. Furthermore, the endpoint must allow filtering for NSIP data only as well as filtering for open data only.

These requirements must be met to ensure that data from your NSIP can be correctly ingested into the ERPD on data.europa.eu. The correct ingestion of data by data.europa.eu relies on a mode of data provisioning that makes, in some way, NSIP data separable from other, open data that might be provided via the same portal.

## 3.2 Technical requirements

The harvester accesses the NSIP endpoints on a weekly basis. Depending on the total size of data provided by each NSIP as well as depending on other factors such as available resources, other harvesting intervals can be negotiated on an individual basis. Metadata data is processed overnight. Every incoming non-DCAT-AP-format will be transformed to DCAT-AP 2.1.1. The harvester is configured individually for each harvested NSIP.

## 3.3 Access to harvested sites

### 3.3.1 Authentication

Some source sites require authentication, this means we need a login name and password before we can access the data. If this applies to your portal, please state this in your message when using our contact form[5] (see section 6 for further instructions).

### 3.3.2 API access to harvested site

To be able to be harvested, the source site needs to expose an endpoint from which the data can be gathered.

[5] https://data.europa.eu/en/contact-us?type=feedback-suggestions

This endpoint should, as described in 3.1, offer the ability to only provide the NSIP data. Either by being a dedicated endpoint or by offering a filter which includes only NSIP data.

The harvested NSIP data should have one of the formats described in section 6 and the endpoint should use one of the protocols described in section 6. The preferred combination is DCAT-AP via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), but others are also possible.

### 3.3.3 FTP access to harvested site

The ERPD on data.europa.eu does not support FTP for downloading datasets from a source site.

## 3.4 Operational requirements

### 3.4.1 Harvesting frequency

Due to the high volume of metadata that will be harvested from a growing list of data suppliers and the required runtime for the harvesting processes, data supplier sites are harvested weekly by default. Furthermore, the harvesting processes must be clustered and scheduled on a fixed time schedule (e.g. during the night) in order to avoid any load impacts on the harvested sites during regular business hours usage. Other factors and circumstances permitting, harvesting intervals that are more or less frequent can be agreed individually.

### 3.4.2 Data source site API / endpoint

The data source endpoint should accept queries with, for example, offset / limit parameters for resumption, partitioning, and pagination of the datasets to be harvested.

### 3.4.3 Ensuring uniquely identifiable datasets

Only when the same dataset always has the same unique id it can be ensured that it will be recognized as the same dataset on data.europa.eu and that it will not be duplicated.

# 4. Required metadata

The following metadata is required for NSIPs and the ERPD according to Articles 5-8 of the Data Governance Act: title, description, publisher, conditions for re-use / access procedure, format, and size. To enable a comprehensive harvesting of required metadata, equivalent metadata is mandatory for ERPD harvesting according to these guidelines.

The following tables show how the metadata required by the Data Governance Act are modelled to corresponding DCAT-AP properties based on DCAT-AP (version 2.1.1). As explained in section 2, the required metadata must be modelled into datasets and distributions. Because the Data Governance Act requires metadata that pertains to the datasets *and* distribution classes, metadata on both classes must be provided as specified below to fully meet the Act's legal requirements. The data service class can be used in specific cases to specify metadata on data endpoints. Metadata properties that are generally mandatory according to the DCAT-AP specification are marked by an asterisk (*) in the tables below.

Both dataset and distribution level metadata must be made available for harvesting. Sections 4.1 and 4.2 list mandatory metadata required for dataset and distribution classes. For the avoidance of doubt, information on the underlying distribution(s) is also required via the dcat:Distribution property in section 4.1.

## 4.1 Datasets

According to DCAT-AP, a dataset is a collection of data, published or curated by a single agent. Data comes in many forms including numbers, words, pixels, imagery, sound and other multi-media, and potentially other types, any of which might be collected into a dataset.

The following metadata is mandatory for NSIP datasets:

| Property | URI | Range | Usage note | Cardinality |
|---|---|---|---|---|
| Title (M) | dct:title * | rdfs:Literal | This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name. | 1..n |
| Description (M) | dct:description * | rdfs:Literal | This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description. | 1..n |
| Publisher (M) | dct:publisher | foaf:Agent | This property refers to an entity (organisation) responsible for making the Dataset available. | 1..1 |
| Access rights (M) | dct:accessRights | dct:RightsStatement | This property refers to information that indicates whether the dataset is open data, has access restrictions, or is not public. From the controlled vocabulary of the Publications Office of the EU[6], the following codes should be used for NSIP data: "non-public" or "restricted". "open" is prohibited for NSIP data. | 1..1 |
| Distribution (M) | dct:distribution | dcat:Distribution | Distribution(s) available for a dataset. | 1..n |

[6] https://op.europa.eu/en/web/eu-vocabularies/concept-scheme/-/resource?uri=http://publications.europa.eu/resource/authority/access-right

Further properties that are either recommended or optional according to the DCAT-AP specification can be provided at the NSIP's or data publisher's discretion.[7]

## 4.2 Distribution

A distribution according to DCAT-AP represents an accessible form of a dataset such as a downloadable file.

The following metadata is mandatory for distributions:

| Property | URI | Range | Usage note | Cardinality |
|---|---|---|---|---|
| Format (M) | dct:format | dct:MediaTypeOrExtent | This property refers to the file format of the Distribution.<br>You can only specify one format per Distribution. If an NSIP offers the same data in different formats, each format must be specified as a separate distribution. | 1..1 |
| Size (M) | dcat:byteSize | rdfs:Literal | The size in bytes can be approximated (as a decimal) if the precise size is not known.[8] | 1..1 |
| Access procedure (M) | dcat:accessURL * | rdfs:Resource | A URL of a Website that enables either access to the described data or that contains information on how to request the data. | 1..n |
| Conditions for re-use (Rights) (M) | dct:rights | dct:RightsStatement | This property refers to a statement that specifies rights associated with the Distribution. | 1..1 |

Further properties that are either recommended or optional according to the DCAT-AP specification can be provided at the NSIP's or data publisher's discretion.

If your NSIP includes metadata on endpoints (e.g. and APIs) that are directly accessible[9], we recommend that you use the DCAT-AP DataService class in addition to the Distribution. If you use the DataService class,

[7] As specified in section 3.1, data.europa.eu will differentiate open data and NSIP-related data in one of two ways: Either NSIP-related data is delivered via a dedicated endpoint or, if an endpoint delivers both open data and NSIP-data, data publishers must specify a mechanism that allows data.europa.eu to filter (and therefore separate) NSIP-data from open data. The exact filter logic can be agreed individually as part of the harvesting on-boarding process. Both mechanisms enable data.europa.eu to differentiate and then label the relevant data as non-open data that has originated from an NSIP.
Ongoing discussions in SEMIC might lead to future extensions or modifications of DCAT-AP that specify additional properties or vocabularies, e.g. to indicate the legislation that led to the release of a dataset. If such additions were to be adopted by SEMIC and implemented by Member States, data publishers will of course be able to provide such data to the ERPD. data.europa.eu's triple store backend will accept and save such data.
However, if NSIP data and open data are provided via the same endpoint, it should be stressed that such tagging will not invalidate the need for a filter mechanism as described in section 3.1.

[8] If data is offered via an API or other endpoint, size should refer to the overall size of the underlying dataset.

[9] This means that this information should only be provided if an endpoint exists that can be accessed by external users without e.g. registration. In these cases, we recommend that the endpoint URL is provided in addition to the information on how to request access (as expressed by dcat:accessURL as part of Distributions). For cases where the relevant endpoints cannot be accessed without prior registration, information on the access URL of the relevant distribution is sufficient.

you must include information on the properties endpointURL[10] and title[11]. Further recommended or optional DataService properties can be added on discretion.

[10] I.e. The root location or primary endpoint of the service (an IRI), according to DCAT-AP 2.1.1. https://github.com/SEMICeu/DCAT-AP/blob/v2.1.1/releases/2.1.1/dcat-ap_2.1.1.pdf

[11] I.e. a name given to the Data Service. This property can be repeated for parallel language versions of the name. https://github.com/SEMICeu/DCAT-AP/blob/v2.1.1/releases/2.1.1/dcat-ap_2.1.1.pdf

# 5. Supported formats and protocols

DCAT-AP via OAI-PMH is the preferred way of data harvesting. We can also accept data provided via CKAN APIs. However, we recommend that this is solution is only used for legacy systems, i.e. only in cases where NSIPs are implemented as part of existing, CKAN-based (open) data infrastructures.

## 5.1 DCAT-AP

Providing metadata as DCAT-AP is the officially recommended method and will always be preferred for harvesting.

### 5.1.1 General remarks

DCAT-AP is a metadata specification for describing public sector datasets in Europe. It's based on the data catalogue vocabulary[12]. The datasets are provided as linked data and can be represented in multiple ways. For the harvesting process, any common representation like rdf/xml, n-triples, or turtle is allowed.

### 5.1.2 Metadata model

For general information on the metadata model, please refer to the official documentation[13]. The respective qualifiers (mandatory, recommended, and optional) need to be adhered to and, going beyond the core requirements of DCAT-AP, mandatory metadata as specified in section 4 of this document must be included.

### 5.1.3 Example

The following is an example dataset with all mandatory properties in rdf/xml.

```
<?xml version="1.0" encoding="utf-8" ?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

    xmlns:dcat="http://www.w3.org/ns/dcat#"

    xmlns:dc="http://purl.org/dc/terms/"

    xmlns:foaf="http://xmlns.com/foaf/0.1/"

    xmlns:vcard="http://www.w3.org/2006/vcard/ns#"

    xmlns:locn="http://www.w3.org/ns/locn#"

    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">


 <dcat:Dataset rdf:about="http://data.europa.eu/88u/dataset/ded24b58-a5ab-4d34-8603-23ded830bab2">

   <dc:publisher>

    <foaf:Agent rdf:about="http://publications.europa.eu/resource/authority/corporate-body/CCC">
```

[12] http://www.w3.org/TR/vocab-dcat/

[13] https://github.com/SEMICeu/DCAT-AP

<foaf:name>Customs Cooperation Council</foaf:name>

</foaf:Agent>

</dc:publisher>

<dcat:contactPoint>

<vcard:Kind>

<rdf:type rdf:resource="http://www.w3.org/2006/vcard/ns#Individual"/>

<vcard:hasEmail rdf:resource="mailto:john@doe.de"/>

<vcard:fn>John Doe</vcard:fn>

</vcard:Kind>

</dcat:contactPoint>

<dcat:keyword>example</dcat:keyword>

<dcat:theme rdf:resource="http://publications.europa.eu/resource/authority/data-theme/ENVI"/>

<dc:title>NSIP example metadata dataset</dc:title>

<dc:temporal>

<dc:PeriodOfTime>

<dcat:endDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-06-09T00:00:00</dcat:endDate>

<dcat:startDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-06-09T00:00:00</dcat:startDate>

</dc:PeriodOfTime>

</dc:temporal>

<dc:spatial>

<dc:Location rdf:about="https://piveau.eu/def/example-location">

<locn:geometry rdf:datatype="http://www.opengis.net/ont/geosparql#gmlLiteral"><gml:Envelope srsName="http://www.opengis.net/def/EPSG/0/4326"><gml:lowerCorner>53.1485 12.915</gml:lowerCorner><gml:upperCorner>53.1985 12.9983</gml:upperCorner></gml:Envelope></locn:geometry>

```
    </dc:Location>
  </dc:spatial>

  <dc:identifier>ded24b58-a5ab-4d34-8603-23ded830bab2</dc:identifier>
  <dc:description>This is an minimal example dataset to showcase the metadata to be offered by an NSIP</dc:description>
  <dc:accessRights>
    <dc:RightsStatement rdf:about="http://publications.europa.eu/resource/authority/access-right/RESTRICTED"/>
  </dc:accessRights>

  <dcat:distribution>
    <dcat:Distribution rdf:about="http://data.europa.eu/88u/distribution/a5be938b-a5ab-4d34-8603-cabf323af6ee">
      <dc:format>
        <dc:MediaTypeOrExtent rdf:about="http://publications.europa.eu/resource/authority/file-type/PDF"/>
      </dc:format>

      <dc:title>NSIP example metadata distribution</dc:title>
      <dc:description>An Example Distribution for the NSIP Example Dataset</dc:description>
      <dc:identifier>https://nsip.data.example.com/dataset/ded24b58-a5ab-4d34-8603-23ded830bab2/resource/a5be938b-a5ab-4d34-8603-cabf323af6ee</dc:identifier>
      <dcat:accessURL rdf:resource="https://nsip.data.example.com/dataset/ded24b58-a5ab-4d34-8603-23ded830bab2/access"/>
      <dcat:byteSize rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">18006.0</dcat:byteSize>
      <dc:rights>
        <dc:RightsStatement rdf:about="http://example-rights.com"/>
      </dc:rights>
```

```
    </dcat:Distribution>
  </dcat:distribution>
 </dcat:Dataset>


</rdf:RDF>
```

### 5.1.4 Requests

The harvester currently supports harvesting from an OAI-PMH compliant source or from reading a dump file containing the RDF/XML representation of the datasets or directly reading DCAT-AP from a SPARQL endpoint. If datasets are provided as a dump file, it is recommended to split the file into pages, for example, by using the hydra core vocabulary 3.

For OAI-PMH-compliant sources, only the verb 'ListRecords' is used.

### 5.1.5 Responses

As indicated above, the response must be DCAT-AP-compliant to be understood by the harvesting component.

### 5.1.6 Error handling

The OAI-PMH protocol provides methods for error handling that the harvester can understand. When using this protocol, these error methods should be used.

### 5.1.7 Service information for integration

As stated above, a categorisation mapping should be provided. Apart from that, the URL for the OAI-PMH endpoint or the dump file is needed.

## 5.2 CKAN API

The open-source data portal platform CKAN is still used by various open data portals. Its RPC-style API (action API) is supported as an interface for data suppliers of data.europa.eu. This support will also apply to the ERPD on data.europa.eu. The following options for using that interface are available.

- The data supplier uses CKAN for providing its NSIP metadata. It is important that the used CKAN version supports the action API. The legacy APIs of CKAN are not supported.
- The data supplier offers a CKAN compliant API, where the necessary endpoints reproduce the exact API behaviour.

### 5.2.1 Requests and responses

Only the 'package_search' API endpoint is needed to harvest the metadata. Its specifications are described in detail in the official documentation. This endpoint is used to get the metadata in a paginated way. Therefore, it accepts query parameters in a request and returns a dictionary with datasets as a result. The high-level use of this endpoint must be offered as specified in the CKAN documentation.

Example call: GET http://singleinformationpoint.tld/api/3/action/package_search?rows=50

### 5.2.2 Metadata model

Although the CKAN API can be used as is, the basic CKAN data schema was extended and modified to meet the requirements of the underlying data structure (DCAT-AP) of the data.europa.eu. The response of the 'package_search' action exposes a 'results' field, which holds a list of dictised datasets. The data structure of such a dataset differs from the one of a plain CKAN installation.

Please note:

- Fields marked with an asterisk (*) are CKAN standard. Further information in the official documentation.[14]
- Type specifications according to official JSON standard (http://json.org/).
- Besides the mandatory fields, the field names and types are not strict, but data suppliers must make sure an obvious mapping is possible.
- For a detailed explanation of each field, refer to the DCAT-AP specifications.

Just like metadata provided in DCAT-AP, CKAN metadata must be structured into datasets and distributions. For usage notes of the relevant DCAT-AP properties, please see sections 4.1 and 4.2.

#### *Dataset schema*

The following fields are mandatory for datasets.

| Field | Type | DCAT-AP dataset equivalent |
|---|---|---|
| Title * | string | dct:title |
| Notes * | string | dct:description |
| Publisher | object | dct:publisher |
| accessRights | object | dct:accessRights |
| Resources | object | dct:distribution |

#### *Distribution schema*

The following fields are mandatory.

| Field | Type | DCAT-AP distribution equivalent |
|---|---|---|
| url | string | dcat:accessURL |
| size | number | dcat:byteSize |
| Format | string | dct:format |
| rights | object | dct:rights |

## 5.3 Example

A result of the 'package_search' action looks like this.


```
{
  "help":"http://example.eu/data/api/3/action/help_show?name=package_search",
  "success":true,
  "result":{
    "count":113948,
    "sort":"score desc, metadata_modified desc",
```


[14] https://docs.ckan.org/en/2.9/

```
"facets":{
},
"results":[
  {
    "issued":"2011-10-20T00:00:00Z",
    "id":"525abe30-ef60-4bf9-824e-916368c1fad8",
    "metadata_created":"2015-09-15T12:08:54.860742",
    "metadata_modified":"2015-09-15T13:17:51.405474",
    "temporal":[
      {
        "start_date":"2011-10-19T22:00:00Z",
        "end_date":"2011-10-19T22:00:00Z"
      }
    ],
    "state":"active",
    "type":"dataset",
    "resources":[
      {
        "package_id":"525abe30-ef60-4bf9-824e-916368c1fad8",
        "id":"7166a1fa-d994-4d88-8e76-3378930b1e16",
        "state":"active",
        "format":"XHTML",
        "mimetype":"application/xhtml+xml",
        "name":"Example",
        "created":"2015-09-15T14:39:43.865240",
        "url":"http://example.com"
      }
    ],
    "tags":[
      {
        "vocabulary_id":null,
        "state":"active",
        "display_name":"Example Tag",
        "id":"06993102-a2ee-4e40-b9e4-ed3e4b86e943",
        "name":"example-tag"
      }
    ],
    "groups":[
      {
        "display_name":"Economy and finance",
        "description":"",
        "title":"Economy and finance",
        "id":"128d0956-4526-440e-a951-f153c190d890",
        "name":"economy-and-finance"
      }
    ],
    "creator_user_id":"0ab3c2ec-c2a2-4eef-b70f-ed093e028063",
    "publisher":{
      "resource":"http://example.com "
    },
    "organization":{
      "description":"Example Organization",
      "created":"2015-09-15T13:56:32.985936",
      "title":"Example Organization",
      "name":"example-orag",
      "is_organization":true,
      "state":"active",
      "image_url":"",
      "revision_id":"ea70fb1f-29a8-4e7b-8527-809e4792a75b",
      "type":"organization",
      "id":"0897b420-3c3d-4a19-9c2c-a9815e2db2be",
      "approval_status":"approved"
    },
    "name":"example-dataset",
    "notes":"Example",
    "owner_org":"0897b420-3c3d-4a19-9c2c-a9815e2db2be",
    "modified":"2011-10-20T00:00:00Z",
    "url":"",
    "title":"Example Dataset",
    "identifier":[
```

```
            "http://example-ident.com"
          ]
        }
      ],
      "search_facets":{
      }
    }
}
```

# 6. First steps for getting started

The ERPD on data.europa.eu harvests metadata about non-open-data published by public sector bodies of European Union Member States' NSIPs. To initiate the onboarding of your NSIP onto the ERPD on data.europa.eu, you will need to undertake two sequential steps: check that your NSIP is suitable for harvesting and issue a harvesting request via the data.europa.eu contact form.

## 6.1 Required information to prepare for harvesting

The very first step is to go through the following checklist to gather all the required information. The purpose of this checklist is to guide you in gathering and summarising the main requirements to enable the successful harvesting of a data supplier site and to assure a certain quality level of harvested datasets. Before contacting the ERPD on data.europa.eu please make sure that you can answer all listed questions. Of course, if anything is unclear, you can always reach out to us via the contact form.

Please remember that the preferred harvesting interface is OAI-PMH.

| | **Requirement** | **Value** |
|---|---|---|
| 1 | Which country does your NSIP cover? | Free text |
| 2 | Is your NSIP already being harvested by another portal? | Free text |
| 3 | Does your National Single Information Point provide the metadata listed in section 4 of this documentation? | See section 4.<br>Only metadata can be harvested, not the data itself. |
| 4 | What is the Uniform resource locator (URL) to your NSIPs interface / endpoint? | URL |
| 5 | If your NSIP is integrated in your existing open data infrastructure **AND** if your NSIP cannot be harvested via a dedicated endpoint:<br>How can NSIP data be filtered from other (i.e. open) data that is part your infrastructure? | Free text.<br>This is essential to correctly distinguish data provided under the DGA from other data. |
| 6 | What is the default language of the datasets from your National Single Information point? | Free text |
| 7 | Which metadata standard is supported by your NSIP? | DCAT-AP or CKAN (mapped to DCAT-AP) |
| 8 | Which representation of the metadata is used? | XML, JSON, or any RDF representation |
| 9 | Which type of API is used to retrieve the data? | OAI-PMH (Recommended)<br>RDF dump file<br>CKAN<br>SPARQL endpoint |
| 10 | Is authentication required for you to access your API? | yes/no |
| 11 | Does your data include complete vocabulary for categorisation, or other fields that use a defined vocabulary (for example update frequency)? | Free text.<br>Please provide vocabularies with translations, if available. |

| | | |
|---|---|---|
| 12 | Does your data use standard date/time formats as specified by the ISO8601? | Yes/no.<br>(Please note: Using the ISO8601 standard is mandatory.) |
| 13 | How often can/should the site be harvested? | E.g. daily, weekly, monthly.<br>The default harvesting interval is weekly. |
| 14 | Are there any times when the site should not be harvested (e.g. scheduled maintenance)? | Free text |
| 15 | Who is the publisher of the NSIP (name and email address)? | Free text. e.g. Federal Open Data Agency; info@open-data.gv.example |
| 16 | What is the URL to the homepage of the NSIP? | URL |

## 6.2 Harvesting request via contact form

Once you have gathered all answers to the checklist, the second step is to contact the ERPD to initiate the harvesting onboarding of your NSIP. Please submit a request via the form[15] and please select 'Get harvested by data.europa.eu' as the issue type. In the contact request, please provide information on all questions listed in the checklist.

Once we receive your request, we will assess it and keep you informed on progress.

[15] https://data.europa.eu/en/contact-us