Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP)

November 2019
Purpose

The following document has been developed to define the metadata needed for the Innovation Gateway MVP stage of the onboarding process.

The specification is initially focused on Summary and Business level metadata (see proposed metadata levels in principles below)

The metadata specification has been created and generated from:
https://hdronboard.metadata.works/#/122894/dataClass/123142/main?path=122894-all-123142

<table>
<thead>
<tr>
<th>Document Control Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version</strong></td>
</tr>
</tbody>
</table>

| **Minimum Retention Period** | Permanent |
| **Disposal Required** | Archive when new version created |
| **Classification** | PUBLIC |
| **Next Review due:** | 22/11/2019 |
Contents

METADATA SPECIFICATION TO SUPPORT INNOVATION GATEWAY MINIMUM VIABLE PRODUCT (MVP) ….. 1
NOVEMBER 2019 .......................................................................................................................... 1

PURPOSE ..................................................................................................................................... 2

CONTENTS ..................................................................................................................................... 3

INTRODUCTION ............................................................................................................................ 5

SUMMARY METADATA .................................................................................................................. 6
IDENTIFIER ...................................................................................................................................... 6
TITLE ............................................................................................................................................. 6
ABSTRACT ...................................................................................................................................... 6
PUBLISHER .................................................................................................................................... 7
CONTACT POINT ............................................................................................................................ 7
ACCESS RIGHTS ........................................................................................................................... 7
GROUP ............................................................................................................................................ 8

BUSINESS METADATA .................................................................................................................. 9
REQUIRED ....................................................................................................................................... 9
DESCRIPTION ............................................................................................................................... 9
RELEASE DATE ............................................................................................................................ 9
ACCESS REQUEST COST .............................................................................................................. 10
ACCESS REQUEST DURATION ....................................................................................................... 10
DATA CONTROLLER ..................................................................................................................... 10
DATA PROCESSOR ........................................................................................................................ 10
LICENSE ........................................................................................................................................ 11
DERIVED DATASETS .................................................................................................................... 11
LINKED DATASET ......................................................................................................................... 12
RECOMMENDED ........................................................................................................................ 13
COVERAGE AND DETAIL ............................................................................................................ 13
GEOGRAPHIC COVERAGE ........................................................................................................... 13
PERIODICITY ................................................................................................................................ 13
DATASET END DATE ................................................................................................................... 14
DATASET START DATE ................................................................................................................. 14
JURISDICTION .............................................................................................................................. 15
POPULATION TYPE ..................................................................................................................... 15
STATISTICAL POPULATION .......................................................................................................... 15
AGE BAND ..................................................................................................................................... 15
PHYSICAL SAMPLE AVAILABILITY ............................................................................................... 16
KEYWORDS .................................................................................................................................... 16
FORMAT AND STRUCTURE .......................................................................................................... 16
CONFORMS TO ............................................................................................................................. 16
CONTROLLED VOCABULARY ....................................................................................................... 16
Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP)

LANGUAGE .................................................................................................................. 17
FORMAT ......................................................................................................................... 18
FILE SIZE ......................................................................................................................... 18
ATTRIBUTION ..................................................................................................................... 18
CREATOR ............................................................................................................................. 18
CITATIONS ......................................................................................................................... 19
DOI ........................................................................................................................................ 19

TECHNICAL METADATA ............................................................................................... 20
TABLE NAME ..................................................................................................................... 20
TABLE DESCRIPTION ........................................................................................................ 20
COLUMN NAME ............................................................................................................... 20
COLUMN DESCRIPTION .................................................................................................... 20
DATA TYPE ....................................................................................................................... 20
SENSITIVE .......................................................................................................................... 20
Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP)

Introduction

As part of the development of the Innovation Gateway Minimum Viable Product (MVP), Health Data Research UK (HDR UK) must complete an initial onboarding activity to collect high level information that describes the datasets that are available for research and innovation across members of the UK Health Data Research Alliance and the seven health data research hubs. This information must be human and machine-readable allowing researchers to understand the data that is available and allowing the Gateway to index the information to return relevant search results.

HDR UK’s initial requirements have been used as a baseline specification and iterated based on feedback and suggestions from IBM (supporting MVP portal), Oxford University (providing the Metadata catalogue) and HDR UK. This is part of a MVP. We have tried to provide coverage for organisation who are already collecting metadata and would like to provide it to the gateway, but we expect as the project progresses there will be further iterations of the specification.

The following principles have governed the work and decisions made:

- Specification that will be distributed in first iteration will be “good enough MVP” rather than perfect
- Metadata should be based on a user need and a user story
- Metadata should use standard terminologies whenever possible and 2.1. DCAT is the de-Facto standard and links to schema.org should be provided where available
  - If an identified need does not have an equivalent DCAT entry, another standard should be selected, including ONS standards
  - Only if no standard can be found should we create a new metadata entry
- Naming convention: Metadata "titles" will be lowercase, underscore separated (note that in the human readable specification we are using Capitalised for ease of reading)
- Prioritisation of Metadata has been based on Impact vs Effort (effort needs to consider potential for automation in the longer-term vs manual in nearer term) however, this will be easier to quantify as the project progresses

The following different levels of metadata quality have been defined:

- Summary Metadata (Mandatory)
- Business Metadata
  - Required
  - Recommended
- Technical Metadata (variable level metadata i.e. data elements, types etc.)
  - Required
  - Recommended
- Enhanced Metadata (not in scope for initial on boarding)
  - Profiling
  - Quality
  - Other

As the project progresses metadata will be categorised accordingly based on use need.
## Summary metadata

Summary metadata must be completed for data custodians onboarding metadata into the Innovation Gateway MVP.

<table>
<thead>
<tr>
<th>Identifier</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion Guidance:</strong> Please provide a local identifier used to identify the dataset (if available).&lt;br&gt;&lt;br&gt;<strong>Note:</strong> <em>if a DOI is available this can be provided in the Attribution section.</em>&lt;br&gt;&lt;br&gt;<strong>Definition:</strong> A unique identifier of the item.&lt;br&gt;&lt;br&gt;<strong>Note:</strong> <em>The identifier might be used as part of the URI of the item, but xs:string still having it represented explicitly is useful.</em>&lt;br&gt;&lt;br&gt;Vocabulary: Data Catalog Vocabulary (DCAT)&lt;br&gt;&lt;br&gt;RDF Property: dc:identifier&lt;br&gt;&lt;br&gt;Range: rdfs:Literal&lt;br&gt;&lt;br&gt;Source: dct:identifier dct:resource_identifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Min Occurs:</strong> 0</td>
<td><strong>Max Occurs:</strong> 1</td>
<td></td>
</tr>
</tbody>
</table>

**As A [user]**<br>I Want To<br>know the name of identifier for the dataset<br>So That<br>I can identify, refer to it

<table>
<thead>
<tr>
<th>Title</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion Guidance:</strong> This is the name of the dataset.&lt;br&gt;&lt;br&gt;<strong>Definition:</strong> A name given to the item.&lt;br&gt;&lt;br&gt;Vocabulary: Data Catalog Vocabulary (DCAT)&lt;br&gt;&lt;br&gt;RDF Property: dc:title&lt;br&gt;&lt;br&gt;Range: rdfs:Literal&lt;br&gt;&lt;br&gt;Source: dct:title&lt;br&gt;&lt;br&gt;See also: schema.org Dataset/name</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>xs:string</strong></td>
<td><strong>Min Occurs:</strong> 1</td>
<td><strong>Max Occurs:</strong> 1</td>
</tr>
</tbody>
</table>

**As A [user]**<br>I Want To<br>know the name of the dataset<br>So That<br>I can identify and find it

<table>
<thead>
<tr>
<th>Abstract</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion Guidance:</strong> Provide a short summary of the dataset (limited to 256 characters)&lt;br&gt;&lt;br&gt;<strong>Definition:</strong> A summary of the resource.&lt;br&gt;&lt;br&gt;Vocabulary: DCMI Metadata Terms&lt;br&gt;&lt;br&gt;RDF Property: rdf-syntax-ns#Property&lt;br&gt;&lt;br&gt;Refines: dc:description&lt;br&gt;&lt;br&gt;Source: dc:abstract&lt;br&gt;&lt;br&gt;See also: schema.org Dataset/abstract</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>xs:string</strong></td>
<td><strong>Min Occurs:</strong> 1</td>
<td><strong>Max Occurs:</strong> 1</td>
</tr>
</tbody>
</table>

**As A [user]**<br>I Want To<br>read a plain English summary of the dataset<br>So That<br>I can decide if I want to find out more
**Publisher**

**Completion Notes:** This is the organisation that is the data custodian. They are responsible for the data access request process, as well as publishing and maintaining the metadata.

**Definition:** The entity responsible for making the item available.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dc:publisher

**Source:** dct:creator

See also: schema.org Dataset/publisher

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>know the publisher</td>
</tr>
<tr>
<td>So That</td>
<td>I can understand the context of the dataset</td>
</tr>
</tbody>
</table>

**Contact Point**

**Completion Guidance:** A URL with the information for and/or the email address of, the person or role within the data custodian organisation, who is responsible for the data access process and metadata maintenance.

**Definition:** Relevant contact information for the catalogued resource. Use of vCard is recommended [VCARD-RDF].

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dct:contactPoint

**Range:** vcard:Kind

**Source:** dct:contactPoint

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>contact person</td>
</tr>
<tr>
<td>So That</td>
<td>I can understand the datasets and request access</td>
</tr>
</tbody>
</table>

**Access Rights**

**Completion Guidance:** Text or link to website/documentation where data access request process and/or guidance is provided.

**Definition:** Information about who can access the resource or an indication of its security status.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dct:accessRights

**Range:** dct:RightsStatement

Usage note: Access Rights MAY include information regarding access or restrictions based on privacy, security, or other policies.

**Source:** access rights

See also: schema.org Dataset/conditionsOfAccess

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>know what I can do with the data</td>
</tr>
<tr>
<td>So That</td>
<td>I have a legal and ethical basis for analysing it in my research</td>
</tr>
</tbody>
</table>
**Group**

**Completion Guidance**: Please complete if the dataset is part of a group family or collection i.e. Hospital Episode Statistics has several constituents

**Definition**: A collection is an entity that provides a structure to some constituents that must themselves be entities. These constituents are said to be member of the collections.

More specifically, Influence is the capacity of an entity, activity, or agent to have an effect on the character, development, or behaviour of another by means of usage, start, end, generation, invalidation, xs:string communication, derivation, attribution, association, or delegation.

**Vocabulary**: [PROV-O: The PROV Ontology](https://www.w3.org/TR/prov-o/)

**RDF Property**: prov:Collection

**Sub-property of**: prov:qualifiedInfluence

**Domain**: prov:Entity

**Range**: prov:Entity

**Source**: schema.org Dataset/isPartOf

---

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>know whether the dataset is part of a wider group</td>
</tr>
<tr>
<td>So That</td>
<td>I can find other datasets of the group</td>
</tr>
</tbody>
</table>
Business Metadata

Business metadata provides context for datasets that improve their discoverability through the Innovation Gateway. HDR UK has two categories of business metadata they ask data custodians to provide.

- **Required**: this information is required as part of the onboarding process and is key to making your data more discoverable for researchers
- **Recommended**: this information is recommended for the onboarding process as it gives researchers key contextual information and will improve the dataset rank

### Required

This information is required as part of the onboarding process and is key to making your data more discoverable for researchers

<table>
<thead>
<tr>
<th><strong>Description</strong></th>
<th><strong>Completion Guidance</strong></th>
<th><strong>Vocabulary</strong></th>
<th><strong>RDF Property</strong></th>
<th><strong>Range</strong></th>
<th><strong>Source</strong></th>
<th><strong>Usage Note</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion Guidance</strong></td>
<td>A free-text account of the data (limited to 50000 characters) and/or a resolvable URL of a document (webpage, pdf, word) that describing the dataset. This provides detailed context about the dataset.</td>
<td>Data Catalog Vocabulary (DCAT)</td>
<td>dc:description</td>
<td>rdfs:Literal</td>
<td>dcat:record_description</td>
<td>This property SHOULD be set using the first known date of issuance.</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>A free-text account of the record.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>Data Catalog Vocabulary (DCAT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RDF Property</strong></td>
<td>dc:description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>rdfs:Literal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>dcat:record_description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>See</strong></td>
<td>schema.org Dataset:record</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As A [user] Researcher

I Want To know the description

So That I can understand information about the dataset

<table>
<thead>
<tr>
<th><strong>Release Date</strong></th>
<th><strong>Completion Guidance</strong></th>
<th><strong>Vocabulary</strong></th>
<th><strong>RDF Property</strong></th>
<th><strong>Range</strong></th>
<th><strong>Source</strong></th>
<th><strong>Usage Note</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion Guidance</strong></td>
<td>Date of the latest release of the dataset. If this is a regular release i.e. quarterly, please complete this alongside periodicity, Periodicity will be used to determine when the next update is expected. <strong>Definition</strong>: Date of formal issuance (e.g., publication) of the distribution.</td>
<td>Data Catalog Vocabulary (DCAT)</td>
<td>dcat:issued</td>
<td>xs:dateTime</td>
<td>dcat:distribution_release_date</td>
<td>This property SHOULD be set using the first known date of issuance.</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>Date of formal issuance (e.g., publication) of the distribution, encoded using the relevant ISO 8601 Date and Time compliant string [DATETIME] and typed using the appropriate XML Schema datatype [XMLSCHEMA11-2] (xsd:gYear, xsd:gYearMonth, xsd:date, or xsd:dateTime).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>Data Catalog Vocabulary (DCAT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RDF Property</strong></td>
<td>dcat:issued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>rdfs:Literal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>dcat:distribution_release_date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>See</strong></td>
<td>schema.org Dataset/datePublished</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As A [user] Researcher

I Want To publication date
**Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP)**

**Completion Guidance**: Please provide a free text description of the cost, or an indication of the range/calculation of costs, for processing a data access request.

**Definition**: Indication of cost (in GBP) for processing each data access request by the data custodian.

Source: Schema.org Offer:price

Min Occurs: 0 Max Occurs: 1

**As A [user]**  
Researcher

**I Want To**  
know how much it will cost to process my data access request

**So That**  
So that I can budget and decide accordingly

**Completion Guidance**: Please provide an indicative value and/or estimate of the typical processing time.

**Definition**: Indication of the typical duration of a data access request.

Source: Schema.org Offer:deliveryLeadTime

Min Occurs: 0 Max Occurs: 1

**As A [user]**  
Researcher

**I Want To**  
know how much time it will take to process my data access request

**So That**  
I can plan and decide accordingly

**Data Controller**

Data Controller means a person who (either alone or jointly or in common with other persons) determines the purposes for which and the way any Data Subject data, specifically personal data or are to be processed.

Please complete if the data custodian is not the Data Controller.

Source: ICO Key Terms: Controllers and Processors, Article 4.7 of GDPR

Vocabulary: Data Privacy Vocabulary

Source: dpv:DataController

Min Occurs: 0 Max Occurs: 1

**As A [user]**  
Processor

**I Want To**  
know who has ultimately responsibility for the dataset (if not the publisher)

**So That**  
I can continue to process their data

**Data Processor**

A Data Processor, in relation to any Data Subject data, specifically personal data, means any person (other than an employee of the data controller) who processes the data on behalf of the data controller.

Please complete if the data custodian is the Data Processor.

Source: ICO Key Terms: Controllers and Processors, Article 4.7 of GDPR

Vocabulary: Data Privacy Vocabulary

Min Occurs: 0 Max Occurs: *

10
As A [user]  Administrator
I Want To  know who the data processor is
So That  I can understand the legal basis for processing

License
Definition: A legal document under which the distribution is made available.
Usage note: Information about licenses and rights SHOULD be provided on the level of Distribution. Information about licenses and rights MAY be provided for a Dataset in addition to but not instead of the information provided for the Distributions of that Dataset. Providing license or rights information for a Dataset that is different from information provided for a Distribution of that Dataset SHOULD be avoided as this can create legal conflicts.
Vocabulary: Data Catalog Vocabulary (DCAT)
RDF Property: dct:license
Range: dct:LicenseDocument
Source: dcat:distribution_license
Source: Dataset/license

As A [user]  Researcher
I Want To  know the license that applies to the data
So That  I know the legal information that describes how the distribution is made available

Derived Datasets
Completion Guidance: indicate if derived datasets are available and type of derivation available
Definition: A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity. If this is Entity :e, then it can qualify how it was derived using prov:qualifiedDerivation [a prov:Derivation; prov:entity :e; :foo :bar ].
Vocabulary: PROV-O: The PROV Ontology
RDF Property: prov:qualifiedDerivation
Sub-property of: prov:qualifiedInfluence
Domain: prov:Entity
Range: prov:Derivation
PROV-DM term: prov:wasDerivedFrom
See also: Schema.org  Dataset/isBasedOn

As A [user]  Researcher
I Want To  Know if there are any derivations
So That  I can reuse pre-processed information
**Completion Guidance:** if applicable, describe other datasets that have previously been linked to this dataset and their availability. If no datasets have currently been linked, please indicate if there is the possibility to link the datasets as part of the data access request process.

**Definition:** Link to a description of a relationship with another resource.

*Usage note:* Used to link to another resource where the nature of the relationship is known but does not match one of the standard [DCTERMS] properties (dct:hasPart, dct:isPartOf, dct:conformsTo, dct:isFormatOf, dct:hasFormat, dct:isVersionOf, dct:hasVersion, dct:replaces, dct:isReplacedBy, dct:references, dct:isReferencedBy, xs:string, dct:requires, dct:isRequiredBy) or [PROV-O] properties (prov:wasDerivedFrom, prov:wasInfluencedBy, prov:wasQuotedFrom, prov:wasRevisionOf, prov:hadPrimarySource, prov:alternateOf, prov:specializationOf).

**Vocabulary:** [Data Catalog Vocabulary (DCAT)]

**RDF Property:** dcat:qualifiedRelation

**Sub-property of:** prov:qualifiedInfluence

**Domain:** dcat:Resource

**Range:** dcat:Class:Relationship

**Source:** dcat:resource_qualified_relation

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>I Want To</th>
<th>So That</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher</td>
<td>know if there are any datasets that have already been linked</td>
<td>I can use more pre-processed data in my research</td>
</tr>
</tbody>
</table>
**Recommended**

This information is recommended for the onboarding process as it gives researchers key contextual information and will improve the dataset rank:

### Coverage and detail

<table>
<thead>
<tr>
<th>Definition</th>
<th>The geographical area covered by the dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage note:</td>
<td>The spatial coverage of a dataset may be encoded as an instance of dct:Location or may be indicated using a URI reference (link) to a resource describing a location. It is recommended that links are to entries in a well-maintained gazetteer such as Geonames.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vocabulary:</th>
<th>Data Catalog Vocabulary (DCAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF Property:</td>
<td>dct:spatial</td>
</tr>
<tr>
<td>Range:</td>
<td>dct:Location (A spatial region or named place)</td>
</tr>
<tr>
<td>Source:</td>
<td>dcat:dataset_spatial</td>
</tr>
</tbody>
</table>

See also: schema.org Dataset/spatialCoverage

| xs:string | Min Occurs: 0 | Max Occurs: 1 |

**As A [user]** | **Researcher** |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>know the geographic area covered by the dataset</td>
</tr>
<tr>
<td>So That</td>
<td>I can contextualise my research</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completion guidance</th>
<th>please indicate the frequency of publishing and/or the rate that the whole dataset is updated.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition:</td>
<td>The frequency at which dataset is published.</td>
</tr>
<tr>
<td>Usage note:</td>
<td>The value of dct:accrualPeriodicity gives the rate at which the dataset-as-a-whole is updated. This may be complemented by dct:temporalResolution to give the time between collected data points in a time series.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vocabulary:</th>
<th>Data Catalog Vocabulary (DCAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF Property:</td>
<td>dct:accrualPeriodicity</td>
</tr>
<tr>
<td>Range:</td>
<td>dct:Frequency (A rate at which something recurs)</td>
</tr>
<tr>
<td>Source:</td>
<td>dcat:dataset_frequency</td>
</tr>
</tbody>
</table>

See also: Schema.org Dataset/temporalCoverage

| xs:string | Min Occurs: 0 | Max Occurs: 1 |

**As A [user]** | **Researcher** |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>know the frequency of the dataset</td>
</tr>
<tr>
<td>So That</td>
<td>I know when the next release is</td>
</tr>
</tbody>
</table>
**Dataset end date**

**Completion Guidance:** The end of the time period that the dataset provides coverage for.

**Definition:** The end of the period.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dcat:endDate

**Domain:** dct:PeriodOfTime

**Source:** dcat:period_end_date

**Range:** rdfs:Literal encoded using the relevant ISO 8601 Date and Time compliant string [DATE/TIME] and typed using the appropriate XML Schema datatype [XMLSCHEMA11-2] (xsd:gYear, xsd:gYearMonth, xsd:date, or xsd:dateTime).

**Source:** Schema.org Dataset/temporalCoverage

As A [user] Researcher

I Want To Know the time period for which data is available

So That I can use data that is within the time period I am interested in /know there is enough history.

**Dataset start date**

**Completion Guidance:** The start of the time period that the dataset provides coverage for.

**Definition:** The start of the period.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dcat:startDate

**Domain:** dct:PeriodOfTime

**Source:** dcat:period_start_date

**Range:** rdfs:Literal encoded using the relevant ISO 8601 Date and Time compliant string [DATE/TIME] and typed using the appropriate XML Schema datatype [XMLSCHEMA11-2]

**Source:** Schema.org Dataset/temporalCoverage

As A [user] Researcher

I Want To Know about potential constraints on the data due to jurisdiction

So That I can tailor my request and research accordingly (or realise the data isn’t suitable).
### Jurisdiction

<table>
<thead>
<tr>
<th>Completion Guidance</th>
<th>Definition</th>
<th>Note</th>
<th>Vocabulary</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please use ISO country code for country under whose laws the data subjects' data is collected, processed and stored</td>
<td>A named and identified geospatial area with defined borders which is used for exercising the action of the Rule. An IRI MUST be used to represent this value.</td>
<td>A code value for the area and source of the code must be xs:string presented in the Right Operand. Example: the [iso3166] Country Codes or the Getty Thesaurus of Geographic Names. Narrower terms: spatialCoordinates.</td>
<td>ODRL Vocabulary &amp; Expression</td>
<td>odrl:spatial</td>
</tr>
<tr>
<td>See also: Schema.org Dataset/locationCreated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As A [user] Researcher
I Want To know about potential constraints on the data due to jurisdiction
So That I can tailor my request and research accordingly (or realise the data isn't suitable)

### Population Type

<table>
<thead>
<tr>
<th>Completion Guidance</th>
<th>Usage Note</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please provide a description(s) of the primary population type within the dataset i.e. participants in a study, or images with certain characteristics.</td>
<td><em>Usage Note: Used with Statistical Population, which is the number of instances of a certain population type that satisfy some set of constraints.</em></td>
<td>Schema.org Population Type</td>
</tr>
</tbody>
</table>

As A [user] Researcher
I Want To Know the population type within the data
So That I can understand whether data is suitable for my need

### Statistical Population

<table>
<thead>
<tr>
<th>Completion Guidance</th>
<th>Usage Note</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please provide the primary population size within the dataset i.e. x number of participants in a study, or y number of images with certain characteristics.</td>
<td><em>Usage Note: Used with Population Type, which specifies the type of the population in the dataset. Statistical Population is the number of instances of a certain population type that satisfy some set of constraints.</em></td>
<td>Schema.org Statistical Population</td>
</tr>
</tbody>
</table>

As A [user] Researcher
I Want To Know the population size and type within data
So That I can understand whether data is suitable for my need

### Age Band

<table>
<thead>
<tr>
<th>Completion Guidance</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmonised Principle Age Band Groups 1 &amp; 2</td>
<td>gss.civilservice.gov.uk</td>
</tr>
</tbody>
</table>

As A [user] Researcher
**I Want To**  
Know the age range of individuals included within data

**So That**  
I can understand whether data is suitable for my need

<table>
<thead>
<tr>
<th>Physical Sample Availability</th>
<th>Availability of physical samples associated with the dataset as Yes/No. If Yes provide descriptions of the samples available and process for access.</th>
<th>xs:string</th>
<th>Min Occurs: 0</th>
<th>Max Occurs: 1</th>
</tr>
</thead>
</table>

**As A [user]**  
Researcher

**I Want To**  
Know whether physical samples are available

**So That**  
I can understand if they can be requested for further research

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Definition: A keyword or tag describing the resource. Vocabulary: Data Catalog Vocabulary (DCAT) RDF Property: dcat:keyword Range: rdfs:Literal Source: dcat:resource_keyword</th>
<th>xs:string</th>
<th>Min Occurs: 0</th>
<th>Max Occurs: 1</th>
</tr>
</thead>
</table>

**As A [user]**  
Researcher

**I Want To**  
know the keywords

**So That**  
I can navigate to it easily

**Format and Structure**

| Conforms To | Completion Guidance: data standard such as OMOP or FHIR  
Definition: An established standard to which the described resource conforms.  
“A basis for comparison; a reference point against which other things can be evaluated." [DCTERMS]]  
Usage note: This property SHOULD be used to indicate the model, schema, ontology, view or profile that the catalogued resource content conforms to.”  
Vocabulary: Data Catalog Vocabulary (DCAT) RDF Property: dct:conformsTo Range: dct:Standard Source: dcat:resource_conforms_to | xs:string | Min Occurs: 0 | Max Occurs: 1 |
|-------------|---------------------------------------------------------------------------------------------------------------------------------|----------|---------------|---------------|

**As A [user]**  
Researcher

**I Want To**  
Know which standards the data set conforms to

**So That**  
I can use it to link to other datasets that use the same standards

<table>
<thead>
<tr>
<th>Controlled Vocabulary</th>
<th>Completion Guidance: Standard terminology / Ontology / Controlled Vocabulary such as ICD 10 Codes, NHS Data Dictionary or SNOMED CT International</th>
<th>xs:string</th>
<th>Min Occurs: 0</th>
<th>Max Occurs: 1</th>
</tr>
</thead>
</table>
**Definition:** The nature or genre of the resource.

*Usage note:* The value SHOULD be taken from a well governed and broadly recognised controlled vocabulary, such as: DCMI Type vocabulary [DCTERMS] [ISO-19115-1] scope codes DataCite resource types [DataCite] PARSE. Insight content-types used by re3data.org [RE3DATA-SCHEMA] (see item 15 contentType) MARC intellectual resource types Some members of these controlled vocabularies are not strictly suitable for datasets or data services (e.g. DCMI Type Event, PhysicalObject; [ISO-19115-1] CollectionHardware, CollectionSession, Initiative, Sample, Repository), but might be used in the context of other kinds of catalogues defined in DCAT profiles or applications. *Usage note:* To describe the file format, physical medium, or dimensions of the resource, use the dct:format element.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dct:type

**Sub-property of:** dc:type

**Range:** rdfs:Class

**Source:** dcat:resource_type

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>Know which vocabulary does the dataset use or provide</td>
</tr>
<tr>
<td>So That</td>
<td>I can use it to link to other datasets that use the same vocabularies</td>
</tr>
</tbody>
</table>

**Language**

Definition: A language of the item. This refers to the natural language used for textual metadata (i.e. titles, descriptions, etc) of a catalogued resource (i.e. dataset or service) or the textual values of a dataset distribution.

*Resources defined by the Library of Congress (ISO 639-1, ISO 639-2) SHOULD be used. If an ISO 639-1 (two-letter) code is defined for language, then its corresponding IRI SHOULD be used; if no ISO 639-1 code is defined, then IRI corresponding to the ISO 639-2 (three-letter) code SHOULD be used. Usage note: Repeat this property if the resource is available in multiple languages. Usage note: The value(s) provided for members of a catalogue (i.e. dataset or service) override the value(s) provided for the catalogue if they conflict. Usage note: If representations of a dataset are available for each language separately, define an instance of dct:Distribution for each language and describe the specific language of each distribution using dct:language (i.e. the dataset will have multiple dct:language values and each distribution will have just one as the value of its dct:language property).*

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dct:language

**Range:** dct:LinguisticSystem

**Source:** dcat:language

<table>
<thead>
<tr>
<th>As A [user]</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Want To</td>
<td>Know the language of the dataset</td>
</tr>
<tr>
<td>So That</td>
<td>I understand the contents using that language</td>
</tr>
</tbody>
</table>
**Completion Guidance:** If multiple formats are available please specify

**Definition:** The file format of the distribution.

Usage note: `dct:mediaType` SHOULD be used if the type of the distribution is defined by IANA [IANA MEDIA-TYPES].

Vocabulary: Data Catalog Vocabulary (DCAT)

RDF Property: `dct:format`

Range: `dct:MediaTypeOrExtent`

Source: `dcat:format`

As A [user] Researcher
I Want To know the format
So That I know how to analyse it

**Completion Guidance:** Typical extract file size range i.e. typically 1TB or 10GB or 100MB etc.

Usage note: The size in bytes can be approximated (as a decimal) when the precise size is not known.

**Related Definition:** The size of a distribution in bytes.

Vocabulary: Data Catalog Vocabulary (DCAT)

RDF Property: `dct:byteSize`

Domain: `dct:Distribution`

Range: `rdfs:Literal` typed as `xsd:decimal`

Source: `byteSize`

See also: Dataset/contentSize

As A [user] Researcher
I Want To know the file size
So That I can make provisions for storage, compute, transfer etc.

**Completion Guidance:** Please provide the text that you would like included as part of any citation that credits this dataset.

**Definition:** The entity responsible for producing the resource.

Usage note: Resources of type `foaf:Agent` are recommended as values for this property. See also: § 6.11 Class: Organization/Person

Vocabulary: Data Catalog Vocabulary (DCAT)

RDF Property: `dct:creator`

Range: `foaf:Agent`

Source: `dcat:resource_creator`

See also: Schema.org Dataset/resource_creator

As A [user] Researcher
I Want To know the creator
So That I can understand the context and for attribution
**Citations**

**Completion Guidance:** Please provide a list of known citations, if available.

**Definition:** A related resource, such as a publication, that references, cites, or otherwise points to the catalogued resource.

*Usage note:* In relation to the use case of data citation, when the catalogued resource is a dataset, the dct:isReferencedBy property allows to relate the dataset to the resources (such as scholarly publications) that cite or point to the dataset. Multiple dct:isReferencedBy properties can be used to indicate the dataset has been referenced by multiple publications, or other resources. *Usage note:* This property is used to associate a resource with the resource (of type dct:Resource) in question. For other relations to resources not covered with this property, the more generic property dct:qualifiedRelation can be used.

**Vocabulary:** Data Catalog Vocabulary (DCAT)

**RDF Property:** dct:isReferencedBy

**Source:** dcat:resource_is_referenced_by

**As A [user] Researcher**

**I Want To** know how to cite the asset

**So That** I can appropriately cite the resource in my research publications

**DOI**

**Completion Guidance** Please provide a Digital Object Identifier if available.

**Definition** Digital Object Identifier. Provides an actionable, interoperable, persistent link Actionable – through use of identifier syntax and network resolution mechanism (Handle System®) Persistent – through combination of supporting improved handle infrastructure (registry database, proxy support, etc) and social infrastructure (obligations by Registration Agencies) Interoperable – through use of a data model providing semantic interoperability and grouping mechanisms.

**Vocabulary:** DOI Data Dictionary

**Source:** DOI

**As A [user] Researcher**

**I Want To** know the persistent link

**So That** I can cite and discover the asset
## Technical Metadata

In the phase 1 we will ask sites to provide a data dictionary of technical metadata that includes the following information:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Type</th>
<th>Min Occurs</th>
<th>Max Occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table Name</strong></td>
<td>Name of the table in the dataset. Use a fully qualified name if appropriate.</td>
<td>xs:string</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Table Description</strong></td>
<td>Description of the table in the dataset.</td>
<td>xs:string</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Column Name</strong></td>
<td>Name of the column in the table dataset.</td>
<td>xs:string</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Column Description</strong></td>
<td>Description of the column in the table dataset.</td>
<td>xs:string</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Data Type</strong></td>
<td>Type of data contained in the column.</td>
<td>xs:string</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Sensitive</strong></td>
<td><strong>Completion guidance</strong>: Please indicate (True / False) whether the information must be treated as sensitive and may need additional constraints / removal / anonymisation / masking through the data access request process.</td>
<td>xs:boolean</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>An ODRL conformant policy expressing the rights associated with the resource.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Usage note</strong></td>
<td>Information about rights expressed as an ODRL policy [ODRL-MODEL] using the ODRL vocabulary [ODRL-VOCAB] MAY be provided for the resource.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>ODRL Vocabulary &amp; Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RDF Property</strong></td>
<td>odrl:hasPolicy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>odrl:Policy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>See also guidance at</strong></td>
<td>§ 8 License and rights statements, § 6.4.19 Property: license, § 6.4.1 Property: access rights, § 6.4.20 Property: rights</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>