Australian Government Linked Data Working Group

# **DRAFT URI Guidelines for publishing Linked Datasets on data.gov.au**

**(Draft for review)**

[DRAFT URI Rules for Linked Datasets](#h.mzkw4ei1q3v2)

[Conformance](#h.hg1sj27y6ufa)

[1. Introduction](#h.8x82ev9hlve4)

[2. Linked Dataset URIs](#h.evgc2ceabc0u)

[3. Domain structure](#h.yhbu9svf0ixu)

[4. URI patterns](#h.blxw9q9b2rlf)

[5. Publishing URIs](#h.wfjxm50byvq)

[6. URI naming conventions](#h.h7mypgxdds0)

[7. HOW-TO publish a Linked Dataset on data.gov.au](#h.depmzbrccojw)

[References](#h.1vlcemw9iv0h)

## Conformance

## The key words *MUST*, *MUST NOT*, *REQUIRED*, *SHOULD*, *SHOULD NOT*, *RECOMMENDED*, *MAY*, and *OPTIONAL* in this document are to be interpreted as described in [11].

## 1. Introduction

Uniform Resource Identifiers (URI) are a single global identification system used on the World Wide Web, similar to telephone numbers in a public switched telephone network. URIs are a key technology to support Linked Data by offering a generic mechanism to identify entities (‘Things’) or concepts in the world. Government departments and agencies assign identifiers to all entities ('Things') they are responsible for - e.g., hospitals, schools, roads, equipment, etc. These identifiers are then used when referring to or making statements about particular entities. For example, when referring to a road closure, the identifier (e.g. M5) will be used to inform the public. In order to publish data in a Linked Data fashion, government and governmental agencies need to define these resource identifiers using URIs. Since public sector information (PSI) is intended to be re-used by diverse applications, it is important that these resource identifier URIs remain stable.

This document provides a set of general guidelines aimed at helping government stakeholders to define and manage URIs for ‘Linked Datasets’ and the resources described within. The URI guidelines in this document are building upon the four Linked Data principles postulated by Sir Tim Berners-Lee [1]. Three of these principles are directly applicable to the usage of URIs:

**Use HTTP URIs**

Addressing two principles, *‘use URIs’* and *‘use HTTP URIs’,* governments and their agencies publishing Linked Data ***MUST*** provide HTTP URIs as identifiers for resources, in order to support reuse and data integration/linking on the Web in a Linked Data fashion. HTTP URIs enable URIs to be "looked-up" or "dereferenced", which in turn provides access, via a Web browser, to a representation of the resource identified by these URIs.

**Provide a machine-readable representation of the resource identified by the URI**

In order to enable HTTP URIs to be "dereferenceable", data publishers have to set up the necessary infrastructure (e.g. HTTP servers) to serve representations or descriptions of the resources (e.g. a human-readable HTML representation or a machine-readable RDF/XML representation). For it to be considered Linked data, a publisher ***MUST*** publish the data using RDF (i.e., to define explicitly the meaning of this data) and ***MUST*** publish at least one machine-readable representation (e.g. RDF/XML, JSON-LD, Turtle) via the HTTP URI identifying the resource.

According to these principles this document defines guidance on:

* Linked Dataset URIs,
* Domain structures,
* URI patterns,
* Publishing URIs, and
* URI naming conventions.

## 2. Linked Dataset URIs

For the purpose of this document a *Linked Dataset* published within the data.gov.au domain is defined as a collection of data, each with supporting metadata, published and maintained under the data.gov.au domain, available as RDF, and accessible through dereferenceable HTTP Universal Resource Identifiers (URIs).

HTTP URIs, a component of the World Wide Web, provides a means of uniquely identifying a ‘Thing’ (or ‘Resource’) in this case a *Linked Dataset*. *Linked Datasets* provide the opportunity to share common meaning and common identifiers across the public sector, and to provide comprehensive and reliable identifiers for a collection of ‘Things’ such as the hospitals, schools or roads in a region, climate data for a specific year etc.

A Linked Dataset consists of:

* the URI to identify the set
* metadata to describe its quality characteristics
* a URI that references a list of resources (*Identifier URIs* and *Document URIs*) defined in the Linked Dataset
* references to *Ontology URIs* which define the concept and relationships used within the Linked Dataset

For the *Linked Dataset URI* the following pattern is proposed. The pattern notation used below is based on the “URI Template” specification defined in RFC6570 [7]. In addition square brackets ‘[‘ and ‘]’ are used to introduce optional components and a star, i.e. ‘\*’ following such a bracket component allows arbitrary repetition of the group (zero or more times).

|  |  |
| --- | --- |
| ***Dataset URI*** | The *Dataset URI* ***MUST*** contain the string ‘dataset’, and an appropriate identifier **{datasetid}** describing the nature of the ‘Dataset‘. Optionally, it can also be hierarchically structured with an arbitrary number of path segments that are denoted with the identifier **{module}** below.    **/dataset[/{module}]\*/{datasetid}**    *Top-Level dataset example*  **/dataset/schools**    *Modularised example*  **/dataset/act/schools** |

For expressing the *metadata to describe the quality characteristics* of a dataset the use of DCAT (Data Catalog Vocabulary) [5] is ***RECOMMENDED***, a vocabulary that provides terms and patterns for describing RDF datasets. Consequently, a *Linked Dataset URI* ***SHOULD*** be a member of the class dcat:Dataset.

**Top-Level Dataset and modularised Datasets**

Datasets are typically hierarchically structured, i.e. there exists a superset that consists of multiple parts, for example, a dataset for schools that consist of a dataset for primary schools and secondary schools. Multiple hierarchies may co-exist, for example a dataset for schools may also consist of datasets for public and independent schools. For modelling this hierarchy we propose the use of path segments in the *URI* as described above with the **/dataset[/{module}]\*** pattern. Each module ***SHOULD*** be a member of the dcat:Catalog class. All datasets within a module ***SHOULD*** be referenced with a dcat:dataset property from the *URI* that describes the module (i.e. a member of the dcat:Catalog class). This *Catalogue URI* ***SHOULD*** be dereferenceable at the top-level path segment of the module. For example, for the modularised **/dataset/act/schools** dataset, the dcat:Catalog *URI* is **/dataset/act** that references the schools dataset and all other datasets within this module with the dcat:dataset property.

A top-level module is declared as:

* a dcat:Catalog
* ***SHOULD*** have one or several dataset references with a dcat:dataset property which are themselves members of the dcat:Dataset class
* ***SHOULD*** have one or many publishers defined through the Dublin Core dct:publisher property [6]
* ***SHOULD*** have a license defined, preferably in a common vocabulary such as Dublin Core dct:license

**Dataset ROOT**

The “dataset root” ***MUST*** be identified by a *URI*. Resolving this ***SHOULD*** result in a list of *Dataset URIs* in the domain.

|  |  |
| --- | --- |
| ***Dataset ROOT*** | The *Dataset ROOT* ***MUST*** contain the string ‘dataset’ and ***SHOULD*** list all datasets in the current domain.    **{domain}.data.gov.au/dataset** *[Should dereference a list of all Linked Datasets in the {domain}]* |

**Design principles**

The following principles for Linked Datasets are proposed, derived from existing good practices [1,2,3,4] and revised to meet some specific requirements for the Australian public sector:

**Principle**

|  |  |
| --- | --- |
| Use HTTP so that the *Dataset URI* can be resolved | ***MUST*** |
| Provide at least one machine-readable representation in RDF at the *Dataset URI* | ***MUST*** |
| Provide a human-readable representation in HTML at the *Dataset URI* | ***MUST*** |
| If multiple representations exist, provide a means of discovering specific URIs for each of the available representations | ***SHOULD*** |
| The license for inspection or use of the Linked Dataset shall be provided using a common vocabulary | ***SHOULD*** |
| The metadata for a Linked Dataset should be provided using a common vocabulary and contain the expected longevity and maintenance plans for the *Dataset URI* | ***SHOULD*** |
| The current technical implementation of a data publication system should not be visible in or otherwise affect the URI for a Linked Dataset | ***SHOULD NOT*** |

## 3. Domain structure

All Linked Datasets published under data.gov.au ***MUST*** use a sub-domain of data.gov.au to identify all entities within this Linked Dataset. Data.gov.au currently supports 25 sub-domain names to be used for Linked Datasets that are defined according to the top level of the Australian Governments’ Interactive Functions Thesaurus (AGIFT)[[1]](#footnote-1). The supported sub-domain names are:

business.data.gov.au (BUSINESS SUPPORT AND REGULATION)

communications.data.gov.au (COMMUNICATIONS)

communityservices.data.gov.au (COMMUNITY SERVICES)

culture.data.gov.au (CULTURAL AFFAIRS)

defence.data.gov.au (DEFENCE)

education.data.gov.au (EDUCATION AND TRAINING)

employment.data.gov.au (EMPLOYMENT)

environment.data.gov.au (ENVIRONMENT)

finance.data.gov.au (FINANCE MANAGEMENT)

internationalrelations.data.gov.au (INTERNATIONAL RELATIONS)

governance.data.gov.au (GOVERNANCE)

health.data.gov.au (HEALTH CARE)

immigration.data.gov.au (IMMIGRATION)

indigenous.data.gov.au (INDIGENOUS AFFAIRS)

infrastructure.data.gov.au (CIVIC INFRASTRUCTURE)

justice.data.gov.au (JUSTICE ADMINISTRATION)

maritime.data.gov.au (MARITIME SERVICES)

primaryindustry.data.gov.au (PRIMARY INDUSTRIES)

recreation.data.gov.au (SPORT AND RECREATION)

resources.data.gov.au (NATURAL RESOURCES)

science.data.gov.au (SCIENCE)

security.data.gov.au (SECURITY)

tourism.data.gov.au (TOURISM)

trade.data.gov.au (TRADE)

transport.data.gov.au (TRANSPORT)

The publisher of a Linked Dataset can chose the sub-domain they feel is appropriate for the entities contained within the dataset. If it is unclear which sub-domain is appropriate for a particular Linked dataset the [National Archives search tool](http://agift.naa.gov.au/search.htm) can be used to help identify the government function matching most closely with the domain of the entities contained in the Linked Dataset.

Agencies can request to be the custodian for one of the 25 sub-domains. It is expected that custodians have a robust, secure and highly available hosting environment in place. This is particularly important when the custodian of a sub-domain is also administering redirects for modules under this sub-domain and that are provisioned by other agencies. Where a number of agencies supply content for a single sub-domain (e.g. environment.data.gov.au is likely to be used by multiple agencies publishing Linked Datasets) it may be worth exploring the use of an independent proxy service, which is highly available, and with the single purpose of redirecting traffic to the appropriate infrastructure hosting the dataset.

For choosing the sub-domain name for a Linked Dataset the following principles have been defined which are based on existing good practice and revised to meet the requirements for datasets in the Australian public sector:

|  |  |
| --- | --- |
| **data.gov.au** is the base domain for Linked Datasets that are promoted for re-use | ***MUST*** |
| The government function (e.g. ‘education’, ‘environment’, ‘health’, ‘defence’, ‘location’) according to the top level of the Australian Governments' Interactive Functions Thesaurus (AGIFT) shall be included in the domain name of the URI | ***MUST*** |
| The name of the department or agency currently responsible for a dataset shall not be used in persistent URIs (unless it happens to match the function in AGIFT) | ***MUST NOT*** |
| The sub-domain shall support a direct response (note: this may be implemented as a redirect to department/agency servers from the sub-domain). | ***MUST*** |
| The sub-domain should be maintained in perpetuity | ***SHOULD*** |

## 4. URI patterns

A URI can be used to denote a resource, which is either an Information Resource, or a *Non-Information Resource*. For example, a document, which describes a person including their name, address etc., is an *Information Resource*. The actual person (in the real-world) is a *Non-Information Resource*. Since *Information Resources* (documents, images, datasets etc.) can be served directly, the server returns a representation of the current state of the resource and sends it back to the client, with the HTTP response code 200 OK. *Non-Information Resources* cannot be dereferenced directly, so the server responds by pointing the client to another URI, which is for an information resource which describes the original (non-information) resource, and sets the HTTP response code to 303 See Other.

The following table describes the resource types for which a URI pattern guideline is defined. Section 5 then defines how the server implementing the pattern should resolve these URIs.

|  |  |  |
| --- | --- | --- |
| **Type of Resource** | **Type of URI** | **Definition** |
| A real-world ‘Thing’  ***(Non-Information Resource)*** | ***Identifier URI*** | Identifies some physical or abstract real-world thing that may be referred to in statements.  ***Example of physical real-world ‘Things’:*** a person, a school, a road, a museum etc.  ***Example of abstract real-world ‘Things’:***  a government sector, an ethnic group etc. |
| A regular document  ***(Information Resource)*** | ***Document URI*** | Identifies a document that may be referred to in statements. The *Document URI* should resolve directly to a document on the Web. |
| Definitions of characteristics of a Real-world ‘Thing’  ***(Information Resource)*** | ***Ontology URI*** | Definition of concepts and relations contained within ontologies that define characteristics of a real-world ‘Thing’.  **Example of a definition of a real-world ‘Thing’:** a ‘Person’ class with properties such as a ‘firstname’, ‘lastname’, ‘social security number’, ‘address’, etc.  ‘Ontology URIs’ can be looked up and return a definition or a set of definitions about a real-world thing, e.g. the ‘Person’ class URI will return the set of properties listed above. |

In the following section guidelines for the URI patterns for the different resource types are proposed.

For *Non-Information Resources*, i.e. *Identifier URIs*, there are two URI patterns that are commonly used in the Linked Data Web: *Hash URIs* and *Slash URIs.* Different patterns are proposed for each of these. The HTTP protocol allows resources denoted by *Slash URIs* to be accessed directly, so these are often more effective, but *Hash URIs* are easier to deploy in practice, because only one URI is to be dealt with on the server. Several factors will determine which publishing method is best in a given use case.

Figure 1 provides a decision tree defining tests to best decide between the two publishing methods.

1. The first test is whether the data publisher also has control of the sub-domain in which the dataset is published. If an agency is publishing a Linked Dataset in a sub-domain for which it is not the custodian, *Hash URIs* ***SHOULD*** be used. This recommendation is made to simplify the management of redirects.
2. Similarly, if the custodian of a sub-domain is publishing a Linked Dataset, but does not have full control over its publishing infrastructure (i.e. they cannot set up redirects on their own web servers), *Hash URIs* ***MAY*** be used.
3. If either are possible, i.e. if the Linked Dataset is published by the custodian of the sub-domain and the custodian has full control over its publishing infrastructure, the size of dataset should be taken into consideration for deciding if *Slash URIs* or *Hash URIs* are used. Although there is no definitive threshold, if the dataset only includes up to a couple of hundred entities and is not expected to grow much further in the future, a *Hash URI* pattern ***MAY*** be used for its simplicity. For anything that is considerably bigger and/or is expected to grow significantly in the future, the use of *Slash URIs* is ***RECOMMENDED***.

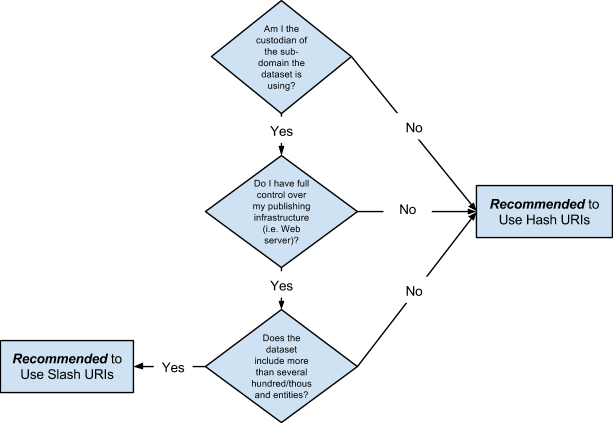


Figure 1: Decision tree for choosing publishing method

***Hash URIs***

URIs can contain a “*fragment identifier”*, an optional special part that is separated from the rest of the URI by a hash symbol (“#”). An example of a “Hash URI” is: **http://education.data.gov.au/resource/schools#2060**. Hash URIs generally identify a secondary resource, subordinate to the main resource, which requires a different processing response. For example, in a web page or linear text document, the fragment identifier typically denotes a position within the page or document, so the browser usually scrolls to that point in the document. In RDF the fragment identifier can be used to point to a subordinate resource. When a *Hash URI* is dereferenced the HTTP protocol requires the fragment part to be stripped off before passing the URI to the server. This means that the server cannot interpret the part after the hash directly. Thus, the presence of a hash in RDF documents can be used to denote other *Non-Information Resources* within the same document without creating ambiguity, as this subordinate URI will never be served directly by the browser. For example, when the URI **http://education.data.gov.au/resource/schools#2060** denoting the school with the identifier 2060 (*Non-Information Resource*) is requested in a browser, the response will be **http://education.data.gov.au/resource/schools**, which is the *Document URI* describing this and potentially other resources.

***Slash URIs***

The second solution for a URI path structure to serve several distinct resource types is the use of so called “Slash URIs” for both, the *Document URI* and the *Identifier URI*, but to use a special HTTP status code, “303 See Other”, to give an indication that a requested *Identifier URI* is not a regular Web document, i.e. *Non-Information Resources*. This style was proposed in W3C's Technical Architecture Group in its httpRange-14 resolution document [8]. For example, if the *Slash URI* **http://education.data.gov.au/id/school/2060** identifying a particular school (*Non-Information Resource*) is requested in a browser, the Web server is configured to answer requests to all these URIs with a “303 See Other” status code and a Location HTTP header that provides the URL of a document (*Information Resource*) that represents the resource, i.e. a redirect from **http://education.data.gov.au/id/school/2060** to **http://education.data.gov.au/doc/school/2060**.

The interested reader is referred to the W3C Note on “Cool URIs for the Semantic Web” [9] for a more in-depth discussion on the difference between the two.

The following table proposes guidelines for the URI patterns for *Slash URIs* and *Hash URIs*.

|  |  |
| --- | --- |
| ***Identifier URI*** | *Slash URI pattern*  The *Identifier URI* ***SHOULD*** contain the token ‘id’, a reference to its concept membership {type} and a local name {name} of the ‘Thing’.  **/id/{type}/{name} → /id/school/2060** *[Canberra Grammar]*  *Hash URI pattern*  For *Hash URIs* the *Identifier URI* ***SHOULD*** contain the token ‘resource’ followed by an appropriate identifier that ***SHOULD*** be the same as the one used for the dataset, i.e. the {datasetid} and a fragment identifier {name} to name the ‘Thing’ locally.  **/resource/{datasetid}#{name} → /resource/schools#2060** *[Canberra Grammar]* |
| ***Document URI*** | *Slash URI pattern*  For *Slash URIs* the URI pattern for a *Document URI* ***SHOULD*** contain the token ‘doc’, a reference to its concept membership {type} and a local name {name} of the ‘Thing’.    **doc/{type}/{name} → doc/school/2060** *[Document about Canberra Grammar]*    *Hash URI pattern*  The *Hash URI* pattern for the *Document URI* ***SHOULD*** contain the token ‘resource’ followed by an appropriate identifier that ***SHOULD*** be the same as the one used for the dataset, i.e. the {datasetid}. The *Document URI* ***MAY*** contain multiple *Identifier URIs* that can be distinguished from the document they are defined in by their fragment identifier.  **/resource/{datasetid}#{name} → /resource/schools#2060** *[Document that contains information about Canberra Grammar (among potentially other resources)]* |
| ***Ontology URI*** | Definition of concepts and relations ***SHOULD*** be denoted by the ‘def’ keyword followed by the ontology name {scheme}, followed by the concept or relationship name {concept}. If the {concept} name is omitted the whole ontology (vocabulary) should be returned. The use of a *Hash URI* pattern *is* ***RECOMMENDED*** for *Ontology URIs* for their simplicity.    *Hash URI pattern*  **/def/{scheme}#{concept} → /def/school#phaseOfEducation** *[The class definition of phaseOfEducation]*  If instances of classes, i.e. the actual entities (*non-information resources*) are modelled as part of the ontology (for example, code lists, finite sets of entities) in a *Hash URI pattern* the URIs used for the entities ***SHOULD*** still follow the *Identifier URI* pattern. |

**Modularising URIs**

Sets of resources may be grouped in modules denoted by URIs that contain an arbitrary number of path segments to indicate a dataset hierarchy as described in Section 2. For each URI type, i.e. for *Identifier URIs*, *Document URIs* and *Ontology URIs* an (optional) [**{module}/]\*** step can indicate the module the particular resource belongs to. If a dataset is part of a module all resources within this dataset ***MUST*** use the same path segment. For example, if no federal identifiers exist for schools, a state may introduce a dataset about schools within a module, i.e. /**dataset/act/schools** where **act** is the **{module}** and **schools** the **{datasetid}**. Schools identified within this module ***MUST*** then use the **act** module, e.g. **/act/id/school/2060** for the Canberra Grammar school (see also in the table below).

Classifications other than a state or the administering authority can form the basis of identifiers of modules. For example, primary and secondary education or public and independent schools could all be distinguished by a separate module, e.g. **/dataset/act/primary/public/schools** and **/dataset/act/primary/independent/catholic/schools**, **/dataset/act/primary/independent/anglican/schools** etc. where **act**, **primary**, **public**, **independent**, **catholic** and **anglican** are all module names, whereas **schools** denotes the datasets in each respective module. However, different type of schools can also be grouped in sub-datasets of a state-wide or even of a federal school dataset resulting, for example, in a path structure like **/dataset/schools/publicSchools** or **/dataset/schools/independentSchools** where **schools** is a federal dataset including all schools in Australia and **publicSchools** and **independentSchools** a sub-dataset including all independent schools in Australia. The decision on how to structure datasets will be use-case specific. However, individual resources ***MAY*** belong to more than one module, and therefore may be identified by more than one URI, each related to a different context.

The following table describes how to modularise the different URI types.

|  |  |
| --- | --- |
| ***Identifier URI*** | *Slash URI pattern*  **/[{module}/]\*/id/{type}/{name} → /act/id/school/2060** *[Canberra Grammar defined in the schools dataset of the act module]*    *Hash URI pattern*  **/[{module}/]\*/resource/{datasetid}#{name}→ /act/resource/schools#2060** *[Canberra Grammar defined in the schools dataset of the act module]* |
| ***Document URI*** | *Slash URI pattern*  **/[{module}/]\*/doc/{type}/{name} → /act/doc/school/2060** *[Document about Canberra Grammar defined in the school dataset of the act module]*    *Hash URI pattern*  **/[{module}/]\*/resource/{datasetid} → /act/resource/schools** *[Document in the act module that contains among other resources information about Canberra Grammar]* |
| ***Ontology URI*** | *Hash URI pattern*  **/[{module}/]\*/def/{scheme}#{concept} → /act/def/school#phaseOfEducation**  *[The class definition of phaseOfEducation in the context of the act module]* |

## 5. Publishing URIs

A URI may identify a resource in a dataset without ever being resolved or dereferenced. However, following the Linked Data principles, a URI ***MUST*** be resolvable using the HTTP protocol [10] making it essentially a URL. As URIs in this document are resolved in a sub-domain of data.gov.au, the URI pattern chosen for a dataset has to be registered with data.gov.au in order to be resolvable as a URL.

**Registering URIs with data.gov.au**

The URI for a dataset, in particular the chosen sub-domain, module and datasetid, have to be registered with data.gov.au. Currently, this process requires one to send an email to [data.gov@finance.gov.au](mailto:data.gov@finance.gov.au), but a Web-enabled management tool is considered for the future.

Depending on the publishing method, different URI path structures have to be registered.

**Hash URIs**

For Hash URIs only the *Dataset URI* has to be registered:

**{domain}**.data.gov.au/dataset**[/{module}]\***/**{datasetid}**

The physical location of the document results from this Dataset URI, i.e.:

**{domain}**.data.gov.au/**[/{module}]\***/resource/**{datasetid}**

**Slash URIs**

Since datasets published in a *Slash URI*pattern ***SHALL NOT*** physically reside on data.gov.au servers, the physical location of the dataset on an agency server has to be registered with data.gov.au. For *Slash URIs* two scenarios have to be distinguished, (1) a custodian of the data.gov.au sub-domain in question is publishing a new Linked Dataset, or (2) an agency that is NOT the custodian of a data.gov.au sub-domain is requesting to publish a new Linked Dataset under this sub-domain that is managed by someone else.

**Custodian of the data.gov.au sub-domain:**

The custodian of a data.gov.au sub-domain only has to register new modules, i.e. the following URIs have to be registered:

**{domain}**.data.gov.au/dataset**[/{module}]\***/

**{domain}**.data.gov.au/**[/{module}/]\***/id/

**{domain}**.data.gov.au/**[/{module}/]\***/doc/

**{domain}**.data.gov.au/**[/{module}/]\***/def/

For each of these URIs the storage location, i.e. the IP address or the hostname, of the data that will be served by these URIs has to be registered with data.gov.au.

Datasets that are published under existing modules or in the top-level URI path, i.e. directly under **{domain}**.data.gov.au/id/, **{domain}**.data.gov.au/doc/ or **{domain}**.data.gov.au/def/ do not need to be registered.

**Agencies that are NOT the custodian of the respective data.gov.au sub-domain:**

Agencies that are not the custodian of the sub-domain can only request a new module within the sub-domain in question, i.e. they cannot publish datasets under the top-level URI path **{domain}**.data.gov.au/id/, **{domain}**.data.gov.au/doc/ or **{domain}**.data.gov.au/def/. The Dataset URI has to be registered as above:

**{domain}**.data.gov.au/dataset**[/{module}]\***/**{datasetid}**

**{domain}**.data.gov.au/**[/{module}/]\***/id/

**{domain}**.data.gov.au/**[/{module}/]\***/doc/

**{domain}**.data.gov.au/**[/{module}/]\***/def/

For each of these URIs the storage location, i.e. the IP address or the hostname, of the data that will be served by these URIs has to be registered with data.gov.au.

If the module name has already been assigned, alternative URI paths will be proposed to the requester.

**Resolving URIs**

*Content negotiation* is a mechanism defined for HTTP that makes it possible to serve different versions of a resource representation at the same URI. Different client applications have different preferences on the data format and language which can be indicated in the HTTP header of the request. For example, a browser usually requests HTML, localized with a natural language such as English or Chinese, while Semantic Web software usually requests RDF. The server then selects the best match, perhaps from its file system, or by generating the desired content on demand, and sends it back to the client.

For dereferencing HTTP URIs there are standard patterns [8] that distinguish between the different resource types. The following paragraphs introduce guidelines for the resolving of the different URI types in the data.gov.au domain.

**Resolving Identifier URIs**

To conform to the Linked data principles [1], a URI for a real-world ‘Thing’ must resolve to a document that contains information about that thing. This principle poses different requirements on the architecture, depending if a *Hash URIs* or *Slash URIs* is used to identify the resource.

* ***Slash Identifier URIs:*** Using the *Slash URI* pattern for *Identifier URIs* requires the setup of a “303 See Other” redirect from **/id/{type}/{id}** to  **/doc/{type}/{id}** to give an indication to the user that the requested resource is a *Non-Information Resource*, and redirecting the user to the document for the ‘Thing’, i.e. the *Information Resource*. Content-negotiation can be used to decide on the specific representation that is returned to the user.
* ***Hash Identifier URIs:*** *Hash URIs* are often used for the simplicity of their implementation on the server, as they work well with file-based systems, and do not require a redirect. The use of the **/resource/{datasetid}#{id}** pattern is proposed for *Hash Identifier URIs*, and the pattern **/resource/{datasetid}** for the *Document URI*. As the fragment part is stripped off by the HTTP protocol, there is no need to rewrite to the *Document URI*, as the *Document URI* is already identified by **/resource/{datasetid}**. In this setting, the *Document URI* is also the ***document storage location***, whereas an *Identifier URI* within this document uses the storage location plus a fragment identifier for identifying its *Non-Information Resource*. For example, if the Identifier URI is **http://education.data.gov.au/resource/schools#2060**, the **#2060** is stripped off and the document is returned, using the *Document URI* **http://education.data.gov.au/resource/schools**. This solution makes the *Hash Identifier URI* pattern very easy to implement while largely maintaining Linked Data principles. It also still allows to retrospectively define *Identifier IDs* of the type **/id/{type}/{id}** or **/id/{type}#{id}** and implementredirects from this *Slash URI* to the original *Document URI* with a *Hash URI* **/resource/{datasetid}** pattern.

**Resolving Document URIs**

A *Document URI* will resolve to the most appropriate representation as defined by the content-type(s) in the ‘Accept’ header of an HTTP request. The behaviour is the same for *Hash URIs* and *Slash URIs*. Where more than one acceptable format is available for the same document containing the same information, content negotiation ***MAY*** be used to decide on the format to be returned.

If the RDF and HTML representations of the resource do not differ in terms of information content the use of the file extension is ***RECOMMENDED*** to distinguish the different representations, e.g. .html, .rdf, .owl, but ***MAY NOT*** explicitly state the representation type in any other part of the URI.

When the RDF and HTML representations of the resource differ substantially, i.e. they are not two versions of the same document, but different documents altogether, a “303 See Other” redirect in combination with a content-negotiation ***SHOULD*** be set up that points to two separate *Document URIs*. In this case the use of a token indicating the file type in the *Document URI is* ***RECOMMENDED***. For example, if the HTML version of **/ doc/school/2060** is fundamentally different to the RDF version a redirect has to point to **/doc/rdf/school/2060.rdf** for the RDF representation and to **/doc/html/school/2060.html** for the HTML representation.

Often it will be required to version a document that is available about a particular real-world Thing. To do this, a ‘date’ token ***SHOULD*** be used for the *Document URIs* to indicate that the information is valid on, or from, a particular date. For example, **/doc/html/2012/school/2060** can be used as a *Document URI* for the school dataset that is current as of 2012.

**Resolving *Ontology URIs***

*Ontology URIs* are a special kind of *Document URI* where the document type is always RDF or OWL. Thus, there is no need for any content negotiation. For classes and properties in a *Hash URI* ontology with the pattern **/def/{scheme}#{concept}**, the hash is automatically stripped off resulting in **/def/{scheme}**,the *Document URI*. If instances of classes (entities) are included in the same file as the classes and properties, the URI scheme of these instances still needs to follow the *Identifier URI* pattern as described in Section 4. Consequently, for ontologies including both, classes/properties and instances and using a *Hash URI* scheme, the file must be stored in two locations, **/def/{scheme}** and **/resource/{type}** in order to allow proper resolving of both URI types.

For *Slash URIs* a redirect to a sub-graph of the RDF graph **/def/{scheme}** describing the **{concept}** or any sub-level of it has to be returned. Implementing this pattern with redirects is more difficult than the *Hash URI* scheme and its use is only ***RECOMMENDED*** for large ontologies.

## 6. URI naming conventions

In the recently published RDF1.1 W3C recommendation [12] URIs were replaced in favour of their internationalized version, the so called Internationalized Resource identifier (IRI) that allows to contain characters from the Universal Character Set (Unicode/ISO 10646), including Chinese or Japanese kanji, Korean, Cyrillic characters, etc. However, considering that government documents in Australia are mostly published in the English language the use of URIs for naming resources in Linked Datasets is ***RECOMMENDED***.

**Character Sets used in URI**

Conforming to the URL Character Encoding using ASCII, a 7-bit character set containing 128 characters, ASCII characters ***SHOULD*** be used, i.e. the numbers from 0-9, the uppercase and lowercase English letters from A to Z, and some special characters, for URIs. Accented letters, diacritical and special language characters ***SHOULD NOT*** be used. Also, the use of spaces in URIs ***SHOULD*** be avoided.

**Naming resources**

English ***SHOULD*** be exclusively used for naming resources, unless the real-world thing is commonly known in English by its native name (e.g. aboriginal name). Everything is sensitive to upper/lower case apart from the domain name. However, using upper/lower case as a differentiating factor in URIs is ***NOT RECOMMENDED***.

|  |  |
| --- | --- |
| **Type of URI** | **Naming convention** |
| ***Dataset URI*** | Lower case ***MUST*** be used for the entire URI path up to the {datasetid} part. No particular recommendations are made for the {datasetid} part, which can use any casing as deemed appropriate for the domain.  Datasets denote a collection of real-world ‘Things’ and thus ***SHOULD*** use the plural for the {datasetid}, e.g. **/dataset/schools**. |
| ***Identifier URI*** | Existing identifiers ***MUST*** always be reused when applicable (even if they are not compliant with the rules on the use of special characters, whitespaces, ... as stated above)  Lower case ***MUST*** be used for the entire URI path up to the {name} part. No particular recommendations are made for the {name} part, which can use any casing as deemed appropriate for the domain.  A singular name ***SHOULD*** always be used for naming one particular physical or abstract real-world thing, except if the word to be used for the thing is only available as plural (e.g. series, species).  The plural ***SHOULD*** always be used for naming a set/list of real-world ‘Things’, e.g. **/id/school/independentSchools** to identify a list of all independent schools in a dataset.    Acronyms ***SHOULD*** all be in upper cases or all in lower cases.    A de-identified scheme ***SHOULD*** be used for persons, i.e. do not include the name of the person in the URI. |
| ***Ontology URI*** | For class names and property names in *Ontology URIs* no formatting guidelines are made. Common practise in ontology engineering is to use either lower or upper camel case, e.g. **/def/phaseOfEducation** or **/def/PhaseOfEducation,** or dashes/underscores as word separators, e.g. **/def/phase\_of\_education**. |

# 7. HOW-TO publish a Linked Dataset on data.gov.au

In this section detailed instruction on how to publish a Linked Dataset on data.gov.au are provided based on the example school dataset that was already used in the previous sections. For the example, let us assume a state government agency governing the educational portfolio in the Australian Capital Territory is to publish the “Locations of all ACT schools” in a Linked Data fashion. Currently, a CSV version of this dataset is available at <http://www.data.gov.au/dataset/location-of-act-schools> published by the Department of Education and Training (ACT).

## Choose Sub-Domain

First, an appropriate sub-domain for the dataset has to be chosen. For the “Location of ACT schools” the **“education”** sub-domain or **“governance”** sub-domain seem appropriate, depending on the level of detail in the dataset and its relation to other datasets. Let us assume, the **“education”** sub-domain is chosen for the example ACT school dataset.

## Choose a path structure (module)

Next, it has to be decided if the dataset is appropriate to be included in the top-level of the respective sub-domain, e.g. **education.data.gov.au/dataset/** or if it is better placed within a module, i.e. an additional path structure that is added to the URI to denote the datasets domain of discourse. The sub-levels of the AGIFT classification of functions in the Australian Government can be consulted to make a decision on the appropriate module. If the dataset is a state dataset, the appropriate state identifier should be used in the module name, i.e. **education.data.gov.au/dataset/act/** in the ACT school example.

## Decide on publishing method

Next, the publishing agent has to decide on the publishing method to use. The decision tree proposed in Figure 1 in Section 4 can be consulted to help deciding whether to use *Hash URIs* or *Slash URIs*. In the case of the ACT school dataset, *Hash URIs* ***MAY*** be used regardless if the publishing agent is the custodian of the education.data.gov.au sub-domain, as the dataset includes only 100 entities and is not expected to grow significantly in the future. For the remainder of this guide, let us assume *Hash URIs* are used.

## Register URI path with data.gov.au

Once the sub-domain and module have been chosen, a name (datasetid) for the dataset is chosen and the URI path has to be registered with data.gov.au. As outlined above, an email has to be sent to [data.gov@finance.gov.au](mailto:data.gov@finance.gov.au) to requests a URI, in our example:

**http://education.data.gov.au/dataset/act/schools**

If the URI is already assigned, an alternative URI will be proposed to the requester. Let us assume the URI is successfully registered for the remainder of this example.

## Develop the dataset

Once the URI path is registered, the Linked Dataset can be developed. As a Hash URI pattern was chosen, entities within this dataset will be identified with URIs such as: **http://education.data.gov.au/act/resource/schools#2060**, denoting Canberra Grammar school.

## Publish the dataset on data.gov.au

When finished the Linked Dataset is uploaded to the data.gov.au servers through the CKAN system. The system will automatically create the metadata that will be accessible at the Dataset URI, i.e. at **http://education.data.gov.au/dataset/act/schools**. The metadata will also include a reference to the storage location, which will be **http://education.data.gov.au/act/resource/schools.rdf**.

## References

[1] Linked Data – Design Issues,<http://www.w3.org/DesignIssues/LinkedData.html>

[2] Designing URI Sets for the UK Public Sector,<http://www.cabinetoffice.gov.uk/resource-library/designing-uri-sets-uk-public-sector>

[3] 223 Best Practices URI Construction,<http://www.w3.org/2011/gld/wiki/223_Best_Practices_URI_Construction>

[4] Defra, UK Linked Data,<http://location.defra.gov.uk/resources/linked-data/>

[5] Data Catalog Vocabulary (DCAT), W3C Working Draft,<http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/>

[6] DCMI Metadata Terms,<http://purl.org/dc/terms/>

[7] [RFC 6570](http://tools.ietf.org/html/rfc6570) - URI Template, Proposed Standard, Internet Engineering Task Force (IETF), March 2012<http://tools.ietf.org/html/rfc6570>

[8] [httpRange-14] Resolved.<http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html>

[9] Cool URIs for the Semantic Web.<http://www.w3.org/TR/cooluris/>

[10] RFC2616 Hypertext Transfer Protocol -- HTTP/1.1, Internet Engineering Task Force (IETF), June 1999<http://www.ietf.org/rfc/rfc2616>

[11] S. Bradner. [Key words for use in RFCs to Indicate Requirement Levels.](http://www.ietf.org/rfc/rfc2119.txt) March 1997. Internet RFC 2119. URL:<http://www.ietf.org/rfc/rfc2119.txt>  
[12] RDF 1.1 Concepts and Abstract Syntax, W3C Proposed Recommendation, 09 January 2014, <http://www.w3.org/TR/rdf11-concepts/>

1. See http://agift.naa.gov.au/ [↑](#footnote-ref-1)