Submission to the **Productivity Commission**

Data Availability and Use

Executive summary

The Australian Government Linked Data Working Group is pleased to make this submission to the Productivity Commission's Inquiry into "Data Availability and Use". Specifically, this submission addresses the questions posed in the inquiry around the "impediments that may unnecessarily restrict the availability and linking of data", the "role of third party intermediaries to assist consumers in making use of their data" and the "options for, and benefits and costs of, standardising the collection, sharing and release of public and private sector data."

Australian Government policy has recognised that publishing data using open data formats poses significant opportunities for enhanced innovation, efficiency and productivity across the economy. As of July 2016, 8.3k datasets have been published on *data.gov.au*, of which about two-thirds publish the data in some structured data format, such as CSV, JSON, XML or RDF. However, only a small fraction of the datasets are expressed using a rich data model that allows the authors of the dataset to link data to other sources where appropriate and to enable the users of the data the potential to draw connections between data published by different agencies that are not apparent from within one data source. The role that such so called *Linked Data* can play in transforming the discovery of data, efficiency in use of data and the ability to build novel services using the data is outlined in this submission.

Australian Government Linked Data Working Group (AGLDWG)

This submission is prepared by members of the Australian Government Linked Data Working Group *(AGLDWG)*. Established in August 2012, the AGLDWG has been drafting policy and technical guidance on the implementation of Linked Data for the Australian Government, in particular, the group aims to:

- Establish technical guidance publishing public sector information using Linked Data as a delivery technology
- Determine governance rules and processes for the effective management of Australian Government Linked Data
- Promote Linked Data across the Australian Government
- Engender the development of Linked Data infrastructure

Specifically, the group maintains and develops guidelines aimed at helping government stakeholders to define and manage URIs for 'Linked Datasets' and the resources described within. The group also reviews ontologies in the context of other initiatives such as the W3C Data On the Web Best Practices *(http://www.w3c.org/TR/dwbp-ucr/)* and where necessary develops purpose-built ontologies for 'Linked Datasets' published by the Australian Government. The group also maintains, and is expanding, infrastructure to deliver persistent URIs for Government based datasets using the **data.gov.au/** pattern.

Further information on the AGLDWG – including the current membership – can be found at *(http://linked.data.gov.au/)*.

Linked Data

Before outlining some of the initiatives that are undertaken by several federal agencies using Linked Data, we give a definition of the term 'Linked Data', distinguishing it from a term it is frequently confused with, 'Data Linkage'.

The term '*Linked Data*' refers to a set of practices for publishing and linking structured data. Data is '*Linked Data*' when it is linked to other data and if, in turn, it can be linked to/from other data. *Linked Data* is also data that is published in a machine-readable way, meaning all terms used are explicitly defined both in format and meaning. The four core best practices that define Linked Data formally are:

1. **Use URIs to name things:** A Uniform Resource Identifier (URI) is a compact sequence of characters that can be used to identify an abstract or physical resource. URIs can be used for all kind of things, but the important characteristic of a URI is, that it is globally unique. The example URI, *http://lab.environment.data.gov.au/id/acorn-system/Adelaide_023090_023000*, identifies a measurement station in Adelaide that is operated by the Bureau of Meteorology. Identifying and naming real-world and abstract 'THINGS' with URIs is extremely powerful, because, it *does not require any up-front agreement* as you are the owner of that URI and others can *reuse the URI* and make further/other statements about that 'THING'. Government agencies already keep lists of identifiers for 'THINGs' they are responsible for *(e.g. people, places, companies, products, roads, hospitals, schools, events, natural resources, etc.)* and are best placed to define authoritative URIs for these 'THINGS'.

2. **Put these URIs on the Web:** Once a URI is assigned to a 'THING', information should be put up at the URI on the Web *(i.e. it should be given a URL)* that describes the 'THING' on the Web. If you type in the URI in your browser, you should get some information about the 'THING'.

3. **Provide useful information at that URL:** Following the Linked Data principles, the information provided at that URL should use a knowledge representation language such as RDF or OWL. The essential characteristic of these languages is that you can make statements in the form of *subject - predicate - object* expressions, for example: statements such as *http://dbpedia.org/resource/Canberra - http://dbpedia.org/property/capital - http://dbpedia.org/resource/Australia*, which represents that Canberra is the capital of Australia. Providing useful information also entails that meaning to 'THINGS' should be assigned through ontologies (vocabularies), i.e. a set of concepts (shared objects) and relationships between those concepts (shared predicates), in this example, *http://dbpedia.org/property/capital* is a property defined by DBpedia, a semantic version of Wikipedia, that indicates a '*capital of*' relationship.

4. **Include links to other things:** Identifying things with URIs works best when everyone uses the same URI to refer to the same 'THING'. For the Government that means in most cases, that it should define a URI for a 'THING', but if there exists already an established URI *(defined by a third-party such as DBpedia)* for that 'THING' it should link to a URI describing the same 'THING'.

Data Linkage

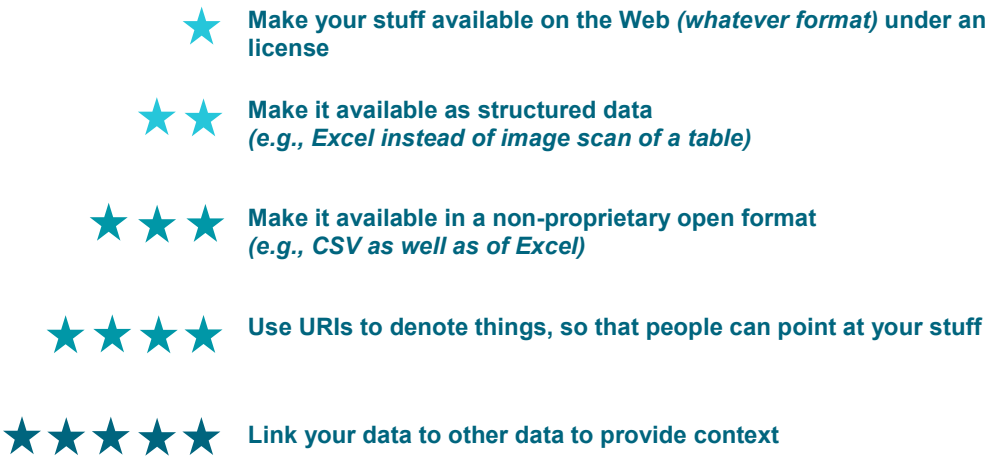
'Data Linkage', also known as 'statistical data integration' differs from 'Linked Data' in that it is concerned with matching entities in two or more datasets. This matching is intended to occur when the entities in the different datasets represent the same actually-existing entity. For example, matching records that represent the same drivers' license holder in two datasets. The emphasis on records indicates that it is a technique that is typically applied to traditionally-structured datasets. Data Linkage is used to combine datasets in order to obtain insights that span the scope of the linked datasets in order to address specific information needs. As such, the Linked Data principles outlined above will aid in the task of Data Linkage, but not all Linked Data initiatives will require the introduction of the strong equivalence relationships that are needed in Data Linkage. In the Commonwealth Public Sector data linkage projects are performed by trusted integrating authorities, such as the Australian Bureau of Statistics, in order to manage the risk of improper disclosure.

A definition of data linkage can be found in "A Guide for Data Integration Projects Involving Commonwealth Data for Statistical and Research Purposes" *(https://statistical-data-integration.govspace.gov.au/)* and is quoted below:

> Statistical data integration involves combining data from different administrative and/or survey sources, at the unit level (i.e. for an individual person or organisation) or micro level (e.g. information for a small geographic area), to produce new datasets for statistical and research purposes. This approach leverages more information from the combination of individual datasets than is available from the individual datasets separately.
>
> *In this guide, data integration refers to the full range of management and governance practices around the process, including project approval, data transfer, linking and merging the data and dissemination.*
>
> ***Linking** (also referred to as 'data linkage' or 'data matching') is that part of the process that involves creating links between records from different sources based on common features in those sources.*
>
> *For records that can be linked, **data merging** is the process of combining individual records (or information in those records) into an integrated dataset specific to the purpose of the analysis. It is recommended that the integrated dataset be de-identified, unless the use of identified data is required and approved for the purpose of the project. (from* https://statistical-data-integration.govspace.gov.au/about-3/what-is-statistical-data-integration/*)*

Linked Open Data vs. Linked (Closed) Data

When Sir Tim Berners-Lee, inventor of the Web, suggested to define what *Linked Open Data* is, he proposed the following five star rating system *(see https://www.w3.org/DesignIssues/LinkedData.html):*



In this five-star rating system, Linked Data has to be open *(1st star)* and use an open format *(3rd star)*, whereas the linking of data is in stars four and five. "Open data" here means that data is available for anyone to use, reuse and redistribute *(see the http://opendefinition.org for more details).* If you assume that the 4th and 5th star depend on fulfilling all the previous requirements, Linked Data has to be openly available in an open format. Often, open data systems such as most of the datasets published on *data.gov.au*, provide the first three stars *(e.g. a RESTful API providing data in JSON format)*, but do not necessarily comply to the fourth and fifth star. However, *not all Linked Data has to be open*, and in the case of many datasets managed by the Government and enterprises for that matter, the benefit stems from conforming to the four Linked Data best practices *(described above)*, while publishing the URLs behind some proprietary firewall for internal use only. In the remainder of this report we refer to such Linked Data as *Linked Closed Data*.

Linked Data and the Australian Government

Linked Data and its associated technologies and methods have already been used by Australian Government agencies. Some examples follow.

Linked Open Data in the Australian Government

Sharing data *openly* on the Web allows Government agencies around the world to increase transparency, deliver more efficient public services and encourage greater public and commercial use and re-use of government information. In order to reap the most benefits from sharing the data, Governments should publish the data in its raw *(original)* form, build a catalog *(e.g. http://data.gov.au)* and make the data human- and machine readable. Linked Data is the accepted best practice to make a dataset available in a human- and machine readable form. Linked Data, as defined by the best practices above, relies on the data to be represented in RDF/OWL. A process commonly known as triplification is typically required to represent the source raw data in terms of RDF classes and properties. This process requires an RDF vocabulary *(ontology)* to be used as the base for generating the RDF data. The construction of such *(open)* vocabularies is extremely important, because the more one is reusing existing vocabularies *(ontologies)*, the easier it will be to interlink the results to other existing datasets. This approach greatly improves the ability of third parties to use the information provided by Governments in ways not previously envisioned, such as the creation of mashups, i.e., the merging of data from different data sources, in order to produce comparative views of the combined information *(see https://www.w3.org/TR/gov-data/)*. One of the objectives of the Australian Government Linked Data Working group is to promote the development and reuse of vocabularies *(ontologies)* within the Australian Government. The following sections outline example *Linked Open Data* implementations by Australian Federal Government agencies that have developed purpose built ontologies, while at the same time reusing existing internationally recognised ontologies. This approach not only allows the reuse of data in a local, Australian context, but allows the data to be compared and interlinked with other international open government and non-government data.

Australian Governments' Interactive Functions Thesaurus (AGIFT)

The Australian Governments' Interactive Functions Thesaurus (AGIFT) is a three-level hierarchical thesaurus that describes the business functions carried out across Commonwealth, state and local governments in Australia. AGIFT contributes to the discovery of government information and services that are delivered online by:

- providing standard terms for government agencies for use in the 'Function' element of the AGLS metadata element set
- helping users, who are unsure of terms or information and service responsibilities of the various levels of government, to search government entry points
- providing a framework for government agencies to develop a more detailed agency-based functional thesaurus for their own classification needs.

The 2010 Report of the Government 2.0 Taskforce recommended that agencies should deploy metadata standards and whole of government taxonomies such as AGIFT for the sharing and reuse of public sector information. The Government agreed with this recommendation and stated:

> *To enable and assist the discovery, sharing and reuse of PSI [Public Sector Information], agencies should deploy endorsed metadata standards such as the Australian Government Locator Service Metadata Standard (AS 5044) together with whole of government taxonomies such as the Australian Governments' Interactive Functions Thesaurus (AGIFT) as outlined in the Australian Government's Information Interoperability Framework. Whenever not being able to meet such standards would appreciably delay the release of PSI, agencies should release non-compliant data until such time as they are able to comply with the standards* http://www.finance.gov.au/publications/govresponse20report/.

AGIFT was subsequently developed as an online thesaurus to operate more efficiently as a resource for online information. Each individual function and term has a page describing its scope and application.

In late 2014, AGLDWG approached the National Archives to discuss the possibility of AGIFT to be published through a linked data platform. The Archives subsequently contracted a member of ALDWG to develop AGIFT as a SKOS model, to enable it to be published as Linked Data. This work included linking to the Classification of the Functions of Government *(COFOG)* thesaurus defined by the UN. This work is now complete and the Archives is currently making arrangements to host AGIFT as a linked data service on its website. A testbed implementation is already available in the CSIRO registry *(as discussed below)* at: *http://registry.it.csiro.au/def/agldwg/agift*

AGIFT as a linked data service enables computer processable meanings of Australian Government functions to be available on the Web and for use by Australian datasets such as those found on *data.gov.au*. Publishing AGIFT as a linked data service will help standardise the description of Australian Government datasets and assists with interoperability of government datasets. AGIFT as a linked data service will assist government datasets to maintain links between different sources using explicit AGIFT terms. Publishing AGIFT as a linked data service will also assist the application of AGIFT terms to datasets available on *data.gov.au*. Up until now the application of AGIFT terms has been a manual process. AGIFT as a linked data service will assist the automation of this process.

ACORN-SAT as Linked Data

In 2012 the Australian Bureau of Meteorology published a dataset, ACORN-SAT, containing the homogenised daily temperature observations of 112 locations throughout Australia for the last 100 years. The dataset employs the latest analysis techniques and takes advantage of newly digitised observational data to monitor climate variability and change in Australia. The observations in ACORN-SAT were initially published only as comma separated values, whereas the metadata was published in a PDF report. In 2013 the BoM together with the CSIRO converted the metadata and the observation data into RDF and published the result as Linked Open Data, accessible online via a pilot government linked data service built on a Linked Data API at http://lab.environment.data.gov.au/.

The tabular time series data of the original ACORN-SAT data was mapped to RDF data based on ontologies that were available in the public domain and ontologies that were specifically built for the domain knowledge needed in this context. External ontologies that were used included the Semantic Sensor Network ontology (http://purl.oclc.org/NET/ssnx/ssn) to deliver the publicly available metadata about the BOM weather stations and their deployment history as linked data. The RDF Data Cube Vocabulary (http://www.w3.org/TR/vocab-data-cube/) was used to publish the tabular time series data and structure it into slices to support multiple views and query endpoints. Ontologies that have been developed specifically for this use case included the definition of adjusted aggregate variables and associated parameters for the ACORN-SAT homogenised observation data *(see* http://lab.environment.data.gov.au/def/acorn/sat), the BOM weather stations *(see* http://lab.environment.data.gov.au/def/acorn/system) and the BOM Rainfall districts.

For the URI schema, the URI guidelines issued for the publication of public sector data in the UK have been used, as the work predated the URI guidelines published by the AGLDWG. However, they are largely complimentary, and data.gov.au was used as the root domain for URI sets that are promoted for re-use within the Australian Government by the AGLDWG and the domain prefix "environment" was used to split the governance of these URI sets into sectors matching the competencies of agencies owning shareable data. The URI scheme also supports *Concept* identifiers with a URI starting with "def", based on a word capturing the essence of the real-world “THING" that the set names *(e.g.* http://lab.environment.data.gov.au/def/station/Station) and *Individual* identifiers with a URI starting with "id", based on a code used to identify an individual instance of a concept, where possible based on existing ID schemes. For example, http://lab.environment.data.gov.au/id/station/014015 reuses the code defined by BoM for a Station located near Darwin.

The ACORN-SAT Linked Sensor Data service that was developed in this project uses the ELDA open source implementation of a Linked Data API base on which additional mashup services have also been developed. For example, a mashup where the 112 sensor locations of the ACORN-SAT dataset were embedded in a Google map to let a user explore the yearly, monthly and daily *(min, max, mean)* temperature for a chosen location on the map in a Google Area Chart has been provided at http://lab.environment.data.gov.au/mashup/drilldown. A simple query interface where a user can select a date range for a chosen location via a dropdown box has been developed too, available

at: *http://lab.environment.data.gov.au/mashup*. Based on the user input a SPARQL query is constructed and the resulting JSON document is used to plot the min and max temperature and the rainfall data.

The Linked Open Data version of the ACORN-SAT dataset allows an easy reuse of the data in other datasets such as long term time series data published by the government like census or biodiversity data, which then will become easy to integrate together. There are also opportunities to link the ACORN-SAT dataset to other sparse data sets such as climate data like cyclone tracks. Using Linked Data this could be done without extra duplication of the published observation data.

Arguably the most useful interlinking of long-term climate data is with current weather observations, to provide the user of current observation data with meaningful context. In 2014 the CSIRO used the ACORN-SAT Data Cube to support event detection on live data feeds from a soil moisture wireless sensor network deployed on a farm near Armidale in New South Wales, Australia. In that case, no explicit linking vocabulary was used, but the ontology for the private on-farm weather stations includes the relevant BoM rainfall districts. The rainfall district may be used in federated SPARQL queries (the query language of RDF) to join the local data with ACORN-SAT. In 2015 the BoM and CSIRO have enabled the comparison of current observations published by the Bureau of Meteorology with ACORN-SAT observations made in the past 100 years. This is done using a harvesting and mapping approach, whereby up to date weather observations are regularly imported from the Bureau's "Latest Weather Observations" service *(e.g. http://www.bom.gov.au/fwo/IDN60903/IDN60903.94925.json*) and published using the same Linked Data API as ACORN-SAT at *http://lab.environment.data.gov.au/weather/*. The JSON source observations are provided as an ordered, descending in time, array of observations of meteorological phenomena, including air temperature, rainfall, atmospheric pressure, wind speed and wind direction. Weather observation data are retrieved every 30 minutes, to align with the Bureau's schedule of updates for most sites. The weather observations ontology for the harvested weather observations is also based on the SSN ontology and extends it with concepts for describing specific types of weather observations, such as observations of ambient temperature, wind direction and precipitation. Weather stations are assigned a URI based on their World Meteorological Organization *(WMO)* identifier *(e.g. http://lab.environment.data.gov.au/weather/id/station/94926* identifies the station at Canberra Airport). They are also described using their BoM identifier, which can be used to find corresponding ACORN-SAT stations. For weather stations that are also ACORN-SAT stations, that correspondence is explicitly captured using an *owl:sameAs* relationship, defining an equivalence between the two URIs. Where ACORN-SAT data are available for a station, the weather observation's description of the station includes a link to the corresponding ACORN-SAT time series.

Controlled vocabularies as Linked Data

Term lists occur ubiquitously in all aspects of government and business, as well as most other formalized endeavours such as science and technology. Term lists should be published openly as they are most effective when the same terms are shared across multiple applications.

ABS and NAA maintain many term lists that apply across the community, and specialized agencies like BoM and GA maintain lists of technical terms, primarily for internal use. Many of these are made available externally, but even when published on the Web, most are provided as a Web page or PDF document or spreadsheet, with all terms included in a single file or document. Thus, in order to make reference to a single term, two pieces of information are needed *(i)* the web address for the term list, and *(ii)* the term name. With little standardization of the internal structure of such a document, it can only be interpreted by a human, and does not support machine-machine interactions. Furthermore, for large vocabularies *(e.g. >10,000 terms is not uncommon in technical applications)* the burden of handling all the data at once is very onerous.

Following the principles of Linked Data, every term in a vocabulary can have a unique web identifier *(URI).* Then the website or service that hosts such a vocabulary can respond to a request to deliver only a single term as requested. Furthermore, if the term list has internal structure (e.g. thesaurus-type relationships) then these can be used in a query interface. Finally, this kind of management and presentation allows for change in vocabularies to be managed more comprehensibly over their lifetime of use.

CSIRO has established a number of vocabulary services, and also developed technology that underlies the *Research Vocabularies Australia* service - hosted by the Australian National Data Service. In particular, at *registry.it.csiro.au* CSIRO hosts vocabularies covering

- a number of agriculture vocabularies
- science keywords
- the geologic timescale
- vocabularies supporting soil and environmental science

CSIRO also hosts a number of environmental vocabularies on behalf of *data.gov.au* at *environment.data.gov.au/def*, including:

- water quality and environmental parameters, including lists of chemicals and units-of-measure
- the bioregional assessments glossary

These are used in a number of services, including the *eReefs portal*, by *Sense-T*, and are increasingly used in scientific applications, primarily within CSIRO.

These services are based on standard technologies *(the Simple Knowledge Organization System or SKOS; the Linked Data Platform)* and provide both high-level APIs, low-level query interfaces *(SPARQL)* as well as a Web user interface. CSIRO's deployments are also consistent with international norms and standards, and we are active contributors to innovations on this topic through both general and discipline-specific initiatives.

Linked Closed Data in the Australian Government

The use of Linked Data is not confined exclusively to 'open' *(publicly discoverable and accessible)* data-related activities. There has been extensive use of Linked Data methods and technologies in sectors that do not generally make their data available to the public - for business or personal safety reasons. As mentioned in the example *4.1.3 Controlled vocabularies as Linked Data* the Linked Data approach also offers the opportunity to manage the frameworks that are used to give meaning and interpretation to data resources more efficiently. These definitional documents, for example the Standard Classification of Occupation, are used by Government and others to organise, classify, and understand data and are of great importance in supporting the discovery and effective use of data resources.

Australian Bureau of Statistics - Linked Employers and Employees

The Australian Bureau of Statistics (ABS) is currently experimenting with Linked Data prototypes to determine how these methods and technologies can be used as part of the ABS Transformation Programme. The advantages offered in terms of discovery and ease of integration which are obtained from the use of the Linked Data approach for publicly available data resources also apply to internal or authorised external use. An example of this work is the prototype Graphically Linked Information Discovery Environment *(GLIDE)*.

The prototype GLIDE is an integrated platform for exploratory and explanatory analysis of linked cross-sectional and longitudinal data derived from survey responses, administrative records and emerging Big Data sources (e.g. sensors, commercial transactions, user online activity). It provides a proof-of-concept implementation of the technical components needed to represent and store information in the form of an entity-relationship network, to model the semantics of statistical concepts, and to retrieve, manipulate and visualise entity-level and aggregate data. (see ISWC 2015 SemStats; Connectedness and Meaning: New Analytical Directions for Official Statistics.)

In addition to this application to statistical data integration and economic analysis Linked Data approaches to the management of statistical conceptual frameworks is being investigated.

This work serves as an example of how the methods, patterns and technological components typically used for open uses of data can also benefit the use of closed data which might always be subject to restrictions on access use for privacy or national security reasons. The ability to use the same 'stack' of components and standards - and the same on-staff experts - for open or closed data activities has been identified as a highly attractive feature of this approach given that traditionally these concerns have had independently-maintained 'silos' within an organisation's enterprise architecture.

The Department of Human Services Service Delivery Ontology

The Department of Human Services is demonstrating the use of a Service Delivery ontology to describe the data that is developed at each stage in a government service delivery lifecycle. A model has been developed to describe the information system components: humans, systems and processes, at different stages. For example, the legislation and matters arising in the Administrative Arrangement Order are built in the ontology to show the relationship between the legislation, ministers and departments. As machinery of government changes occur, the demonstration ontology provides a way to assess the impact of the changes. If the demonstration is successful, then the current Linked Closed Data arrangement with limited access could be considered as a Linked Open Data candidate.

Conclusion

There has been increased interest by Australian Federal government agencies in the use Linked Data, either internally or openly, since the first pilot project, the Linked Data version of ACORN-SAT, has been made available on *data.gov.au* in 2012.

The lessons learned from this project and the open challenges that arose in relation to identifying and naming government resources led to the establishment of the Australian Government Linked Data Working Group (*http://linked.data.gov.au*). The working group has since been developing technical guidance *(see https://github.com/AGLDWG/TR/wiki/URI-Guidelines-for-publishing-linked-datasets-on-data.gov.au-v0.1*), ontologies and best practice on the use of Linked Data by the Australian Government. These artifacts and guidelines have subsequently been applied in several Linked Data projects that have been described in this report.

However, more has to be done, in particular, in regards to establishing a government wide 'vocabulary repository' and a 'linked data management platform'. A 'vocabulary repository' is required so to ease the reuse of domain ontologies that have been developed by a Government agency in a specific context, but that have wider potential use. For example, the BoM and CSIRO have developed, within the ACORN-SAT project, an ontology for describing weather stations and a catalog of weather stations that are and have been deployed in Australia. The station ontology itself may be of interest to other scientific agencies when they publish data about their measurement stations. A 'vocabulary repository' where everyone can search terms like "weather station" would help in the discovery of this existing ontology, either by other government agencies or the general public. The catalog of stations should also be easily discoverable so that other datasets can refer to the unique identifiers (URIs) that have been defined for each weather station that is under the custodianship of the BoM. *Data.gov.au* is a first step to such a "Linked data management platform", but a management of the proposed URI schemes goes beyond the capabilities of the current *data.gov.au* platform.