21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Complex licence requirements for the Bioregional Assessments Programme managed by provenance

N.J. Car [a] and M.P. Stenson [a]

[a] *CSIRO Land & Water, Dutton Park QLD, Australia*
Email: nicholas.car@csiro.au

Abstract: The Bioregional Assessment Programme (BAP), is a large, multidisciplinary program of work, assessing the impacts of coal and coal seam gas on water resources. It must provide persistent and well managed access to both products (documents such as reports) and the datasets that support development of the products. To fulfil this requirement we have built a project repository that stores documents and datasets, and gives access to them via persistent URIs.

A requirement of the BAP is to make its datasets and products available at the conclusion of the programme. This raises the issue of how to determine data and metadata access based on a datasets' licence conditions. What makes this issue complex for the BAP is that both the licence of a particular dataset and the licences of its ancestors bear on dataset access rights and use.

In addition to designing the BAP project repository, the authors have implemented the presentation of provenance information that describes the lineage of datasets in accordance with the World Wide Web Consortium's PROV data model and standard. This lineage, presented as a graph, is used to lend transparency and some measure of reproducibility to the datasets by revealing their development history. It can be used to find the ancestor datasets for a dataset of interest and hence those ancestors' licences. Thus, along with a dataset of interest's own licence, all of the factors determining dataset access rights and use can be determined automatically.

In this paper we present our licence management methodology. We detail our licence data model which builds on Creative Commons by using a different rights association mechanism that is reliant on dataset ownership metadata stored elsewhere, and provenance graphs for licences derived from other licences. It then associates properties the Creative Commons model sees as licence properties, with other non-licence objects such as organisations, which are managed elsewhere, in other systems that we also briefly introduce. We describe our RESTful licence web service tool used to manage licence objects and how it delivers them using Linked Data principles via a version of Epimorphics' Linked Data API. We then describe how the BAP's project repository associates datasets with licences and how its provenance graph is leveraged to calculate the appropriate access rights for a dataset, based on the dataset's and its ancestors' licences.

Keywords: *Licence, attribution, rights, datasets, provenance, Linked Data*

1. INTRODUCTION

The Bioregional Assessments Programme (BAP) is building, and will persist, a *Repository*[1] of several thousand datasets either sourced from a series of contributing government agencies, private companies and non-government organisations, or derived from work carried out within the programme. These datasets are used to assess the potential impact of coal seam gas and coal mining on water resources within several Australia bioregions. They need to be made as available as possible to the Australian public and parties wishing to conduct future bioregional assessments within the next 10 years.

One aspect of making these datasets available is their licencing. The BAP itself is able to apply a very non-restrictive licence, Creative Commons v3.0 Australia[2], to most of the work done during the life of the programme, however, with datasets being sourced from a range of organisations, many of them are bound by other licences, not all of which are non-restrictive. Some of these licences contain transitive entailments such as the requirement for users of descendent datasets to include some form of attribution in their data citations. Another entailment is the requirement for certain functions to be performed on data before derivations of it may be published, for instance the de-identification of particular threatened species within ecological assets.

In order to preserve the licencing information for *Source* datasets, those given to the BAP by 3rd parties, and for *Derived* datasets, those generated by the BAP, several data and metadata systems have been built and both metadata information models, and registration procedures have been established.

One particularly novel aspect of the work done to preserve licencing information for the BAP's datasets has been the use of a datasets' provenance graphs (originally recorded to fulfill a requirement for dataset generation transparency) to determine the impact of a datasets' licencing on their descendent datasets - those derived from them and others within the life of the BAP.

2. THE BAP LICENCE INFORMATION MODEL

Significant datasets are often licenced and many metadata schema used to describe datasets, such as Dublin Core[3] and the DCAT Ontology[4] contain properties for datasets indicating their licence and/or rights statements. Some widely used schema, such as the ANZLIC Metadata Schema[5], contain fields for dataset use restrictions which often contain licence information.

Initially, the ANZLIC Metadata Schema was mandated for all BAP datasets, however this was found to be insufficient regarding its ability to represent the rights statements and licence information for Derived datasets built on top of Source datasets with different licences. One example is that for a dataset entitled "Traralgon Formation Coal Extent" who's abstract explains "It was derived by the Bioregional Assessment Programme from isopachs supplied by the Victorian Department of State Development, Business and Innovation", the ANZLIC Metadata Schema records the information given in Figure 1.

The rights statement and licence information given in Figure 1 do not explain the interplay between the rights and licence of the Source dataset used in generating this Derived dataset, and the generated dataset itself. Furthermore, the BAP initially mandated that, as per the example in Figure 1, the rights statements of dataset's ancestors must be added to all datasets' rights statements. Figure 1 shows that the rights statement for the Source dataset, `© Department of State Development, Business and Innovation, State of Victoria` has been added to the rights statement of the Derived dataset `© Department of State Development, Business and Innovation, State of Victoria © Commonwealth of Australia (Bioregional Assessment Programme www.bioregionalassessments.gov.au)`. While this dataset has only one ancestor, some Derived datasets have many. Additionally, some datasets are further derived from other Derived datasets which would result in lengthy, compound, and likely unworkable rights statements.

Finally, compound rights statements do accord with typical copyright statement use which is generally of the form "© {ORGANISATION NAME}", as per each statement in Figure 1.

[1] http://data.bioregionalassessments.gov.au
[2] http://data.bioregionalassessments.gov.au/id/licence/5548349a898c0a3c9ae3600a
[3] http://dublincore.org/specifications/
[4] http://www.w3.org/TR/vocab-dcat/
[5] http://www.anzlic.gov.au/resources/metadata

```
<gmd:resourceConstraints>
 <gmd:MD_LegalConstraints>
   <gmd:useLimitation>
     <gco:CharacterString>
       Creative Commons Attribution 3.0 Australia
       © Department of State Development, Business and Innovation,
       State of Victoria © Commonwealth of Australia (Bioregional
       Assessment Programme www.bioregionalassessments.gov.au)
     </gco:CharacterString>
   </gmd:useLimitation>
   <gmd:useConstraints>
     <gmd:MD_RestrictionCode codeList="http://asdd.ga.gov.au/
         asdd/profileinfo/gmxCodelists.xml#MD_RestrictionCode"
         codeListValue="licence">
           licence
     </gmd:MD_RestrictionCode>
   </gmd:useConstraints>
 </gmd:MD_LegalConstraints>
</gmd:resourceConstraints>
```

Figure 1: An XML snippet from the ANZLIC compliant metadata statement for the BAP Derived dataset "Traralgon Formation Coal Extent". The namespaces used are common ANZLIC namespaces.

For the reasons given above, a richer dataset-rights/licence metadata information model was sought by the BAP with the model shown in Figure 2 developed. This model is informed by the Open Data Rights Statement Vocabulary (ODRS-V) (Dodds, 2013) information model with dataset having both a license and a rights property, with ranges of a Licence and a RightsStatement class object respectively (see the 'Schema Diagram' in Dodds (2013).

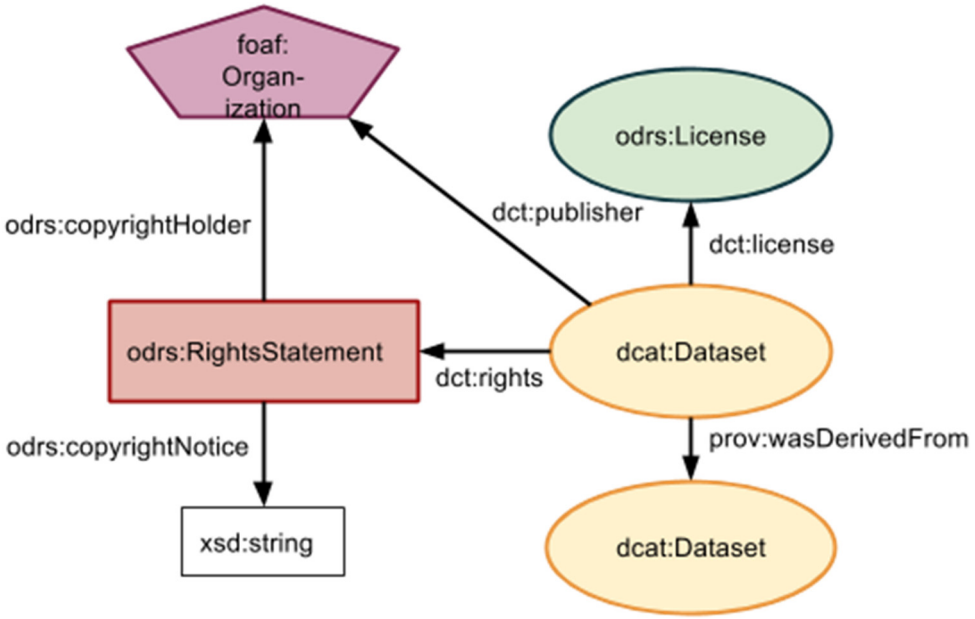

Figure 2: An OWL[6] diagram of the BAP's dataset information model. Only properties relevant to licensing and rights are shown. Namespaces for the RDF/OWL prefixes used are given in Table 1.

[6] Web Ontology Language (OWL) http://www.w3.org/2001/sw/wiki/OWL

The information model shown in Figure 2 is a subset of the BAP's BA Ontology (BA-O)[7] which specifies the complete information model used by the BAP. Figure 2 shows that a BAP dataset has a `dct:license` property pointing to a `odrs:Licence` object as well as a `dct:rights` property pointing to an `odrs:RightsStatement` object. The `dct:publisher` property of the dataset can be inferred (by a reasoning agent) from the link between the dataset and the `odrs:RightsStatement` and the `odrs:RightsStatement` and `org:Organisation`. The text of a copyright notice for a dataset, for example "(c) Geoscience Australia" is contained as a property of a `odrs:RightsStatement` object.

A register of `odrs:RightsStatement` class instances has been created for the BAP that provides a controlled list of copyright notices that can be used for any new dataset. This forces a dataset contributor to ensure that both an appropriate rights holding organisation and appropriate rights statement text are present in the BAP system before rights are able to be assigned to a new dataset. This supports both business rules of the BAP and data integrity by preventing copyright notices without links to further information.

This separation of `odrs:RightsStatement` class instances from `org:Organization` class instances allows the BAP to respond well to governmental change. When an organisation changes its name or remit, as is common for Australian government agencies, rights statements referencing that organisation may be pointed to new replacement organisations if appropriate provenance linking between the old and new organisations is made. An example already seen in the BAP is the case of the Australian Commonwealth's Department of Sustainablility, Environment, Water, Population and Communities (SEWPAC) that no longer exists. Much of its responsibility has been subsumed by the Australian Commonwealth's Department of the Environment (ENV). Rights statements with text such as "(c) SEWPAC 2010" are able to indicate that their current rights holder is ENV while still retaining the original copyright notice text when used.

The full licence information model for the BAP is given in Figure 3. This model is based primarily on the Creative Common licence model contained within the Creative Commons Rights Expression Language (CC-REL) (Abelson, *et. al*., 2008)[8]. The `cc:requires` property of the BAP licence, taken from the ccREL information model, is used to encode licence entailments for datasets. See Section 4 for a full explanation.

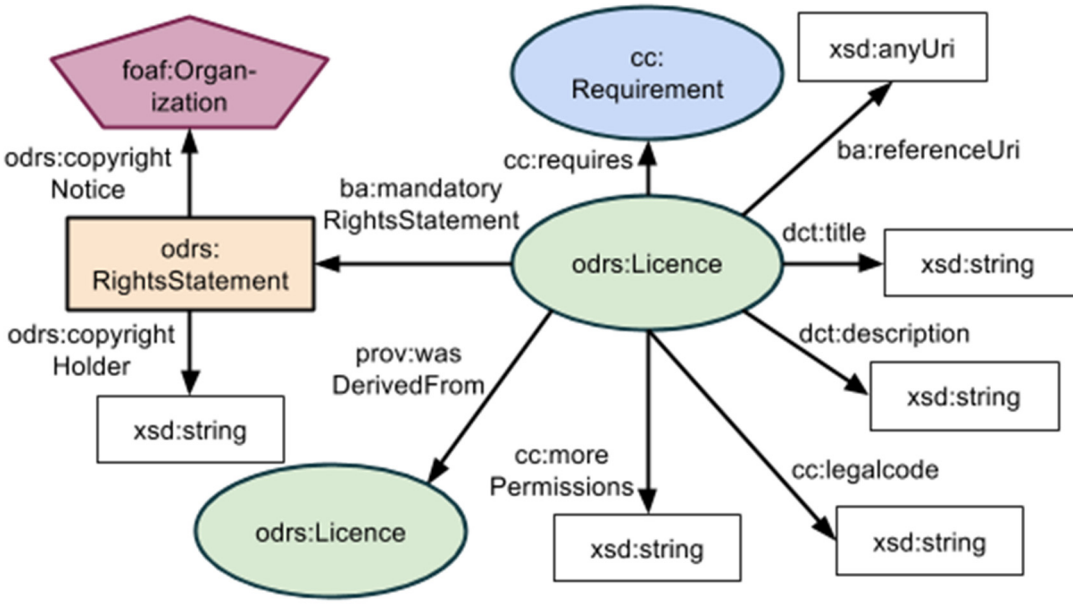

Figure 3: An OWL diagram of the BAP's Licence metadata information model. Namespaces for the RDF/OWL prefixes used are given in Table 1.

The BAP licence information model adds three properties to the standard CC-REL licence object's properties:

1. `prov:wasDerivedFrom` - a standard PROV ontology property to link derived work to originals. This allows for the derivation of licences from other licences;

[7] http://data.bioregionalassessments.gov.au/def/ba
[8] Relevant parts of ccREL for this paper are detailed at http://creativecommons.org/ns

2. `ba:referenceUri` - has range `xsd:anyUri` allowing licenses stored by the BAP to reference their original online presence when that presence does not present as an RDF resource, as is the case for licences represented only by web page text;
3. `ba:mandatoryRightsStatement` - has the range of a particular `odrs:RightsStatement` object. This property allows licences to optionally mandate certain rights statements when used.

3. THE REPOSITORY WEB SERVICES

3.1. Web services locations and types

In order to allow for point-of-truth access to the objects of interest for the BAP Repository, including the use of licences, a series of Web Services were established. The services relevant to licencing are the:

1. **Data Store** – data and basic metadata about datasets (`dcat:Dataset` objects). Also controls access to restricted datasets;
2. **Metadata Catalogue** – complete metadata for datasets (`dcat:CatalogRecord` objects);
3. **People Web Service** – people, groups and organisations (`foaf:Person`, `foaf:Group` & `foaf:Organiszation/org:Organization` objects). Includes authentication services;
4. **Licence Web Service** – licences and rights statements (`odrs:Licence` & `odrs:RightsStatement` objects)

Table 1. OWL namespaces and their prefixes.

Prefix	Namespace
ba	http://data.bioregionalassessments.gov.au/def/ba#
cc	http://creativecommons.org/ns#
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
odrs	http://schema.theodi.org/odrs#
org	http://www.w3.org/ns/org#
prov	http://www.w3.org/ns/prov#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
xsd	http://www.w3.org/2001/XMLSchema#

While the Data Store and the Metadata Catalogue act as common stores and catalogues with web page interfaces for regular use, they also deliver versions of their content in RDF for automated processing. Each Web Service delivers its content via class object type registers specified in the BA Ontology[9]. For example, all organisations can be found at http://data.bioregionalassessments.gov.au/id/organisation/ and the individual organisation "Geosceince Australia" can be found at the URI http://data.bioregionalassessments.gov.au/id/organisation/GA where the standard Linked Data register[10] `id/` denotes a real-world *thing* is being represented and the subregister `organization/` indicates things within it are of type organization – as defined by the BA Ontology. Another example: the licence "BOM, Climate data licence" can be found at the URI http://data.bioregionalassessments.gov.au/id/licence/559d1962898c0a477b44f7ce which uses a database key for the licence ID, rather than an acronym. In addition to the type registers, the top-level register of all *things*, http://data.bioregionalassessments.gov.au/id/, is a register of registers, making type registers easy to find. Also according to Linked Data conventions, the definition space for items within the *data.bioregionalassessmensts.gov.au* domain is `def/`.

All objects of interest to the BAP are thus identified with URIs using the *data.bioregionalassessments. gov.au* domain which was established to allow objects within the multi-agency project to be managed separately from any individual government agency's namespaces. It is expected that this will lend the URIs a level persistence that no individual agency could maintain if their own agency domains were used due to the predilection of Australian government agencies for changing both their web locations, names and structure.

3.2. The common RESTful API

Despite some of the components of the Repository being legacy components, work was done to ensure that objects delivered by any Web Service could be accessed in a similar fashion. A single RESTful[11] API, derived from the Epimorphics ELDA Linked Data API (Epimorphics, 2015) was established allowing for Web Service

[9] http://data.bioregionalassessments.gov.au/def/ba

[10] Best practice for HTTP URIs, as seen by the W3C (http://www.w3.org/TR/ld-bp/#HTTP-URIS) is implemented by the UK Government, see "Creating URIs" (http://data.gov.uk/resources/uris).

[11] https://en.wikipedia.org/wiki/Representational_state_transfer

calls following the patterns in Table 2. The RESTful API described with examples in Table 2 is used to make system-independent calls to collect details about all items of interest to the BAP.

Table 2: Pattern examples of the BAP Repository's RESTful Web Service URIs. All URIs are prefixed with http://data.bioregionalassessments.gov.au. Namespaces for the RDF/OWL prefixes used are given in Table 1.

Web Service call purpose	URI pattern	Query String Arguments
Dataset object's basic metadata, in HTML	/id/dataset/{DATASET_ID}	-
Dataset object's basic metadata, in RDF (turtle serialization)	/id/dataset/{DATASET_ID}	?_format=text/turtle
Organisation object's metadata, in HTML	/id/organisation/{ORG_ID}	-
Organisation object's metadata, in RDF (JSON-LD serialization)	/id/organisation/{ORG_ID}	?_format=application/ld+json
Licence object's legalcode property within the 'cc' (Creative Commons) licence information model view.	/id/licence/{LICENCE_ID}	?_view=cc &_property=legalcode
Summary view of all licences	/id/licence/	?_view=summary
List of all views available for a Licence, in RDF (XML serialization)	/id/licence/{LICENCE_ID}	?_view=alternates &_format=application/rdf+xml
Dataset object's lineage (default information model view)	/id/dataset/{DATASET_ID}	?_property=provenanceAncestry

4. BAP PROVENANCE & LICENCING

4.1. BAP provenance objectives and implementation

An important goal for the BAP's Repository is to assist dataset generation transparency in order to aim for data reproducibility. Dataset lineage has been recorded in the form of standardized provenance graphs according to the PROV data model's ontology, PROV-O (Lebo *et. al.*, 2013). This enables future data users to follow the processing and development of datasets in a system-independent manner. The final web service call shown in Table 2 results in a PROV-O compliant RDF graph for a particular dataset.

In addition to these primary provenance graphs for datasets, additional provenance graphs for other non-dataset class objects in the BAP are maintained for data transparency and semantic integrity, for example for organisations that change over time, as mentioned in Section 2.

4.2. Calculating licencing entailments using dataset provenance graphs

Since the BAP maintains provenance graphs for all datasets, including published products which are a subclass of dataset, one can always discover the ancestors for a particular dataset back to the *Source* datasets received by the BAP. This allows a simplification of the BAP's dataset licence information recording, for the BAP need only record existing licences for *Source* datasets using a standard licence information model and assign licences to *Derived* datasets the programme has produced in order to store all the elements necessary, when joined via a provenance graph, to calculate licence entailments for any dataset.

A subset of three of the many scenario needing licence entailment calculations via provenance are when:

1. A *Source* dataset's licence requires *Derived* datasets to use a certain attribution text when referring to the dataset. Licences such as the "Great Artesian Basin and Laura Basin groundwater recharge areas (GABWRA) Licence"[12] mandate this;
2. Certain sensitive elements within a *Source* dataset must be de-identified in a *Derived* dataset before it can be published. Licences such as "BA Restricted Licence 1"[13] mandate this;
3. A *Source* dataset and derivative works from it may be republished by the BAP however 3rd parties must be made aware that further use or publication of the *Source* or *Derived* datasets must seek permission from the *Source* dataset's rights holder. Licences such as the "Department of Environment – NVIS"[14] mandate this.

[12] http://data.bioregionalassessments.gov.au/id/licence/559b7ac4898c0a477b44f7c8
[13] http://data.bioregionalassessments.gov.au/id/licence/559d18ab898c0a477b44f7cc
[14] http://data.bioregionalassessments.gov.au/id/licence/55949757898c0a477b44f7c1

For all scenarios, a licence will contain a `cc:requires` property with a range of `cc:Requirement`, a specific instance of which will be selected from the BA Ontologies vocabulary of licence requirements[15]. Each licence requirement will contain different specific requirements. In the scenario 1 example, it will indicate a mandatory `odrs:RightsStatement` object containing either an `odrs:attributionText` or an `odrs:copyrightNotice` or both. In the 2 example, the requirement reads "certain elements within this dataset must be de-identified before derived works can be republished, See the dataset's metadata for the list of such elements". Restriction can be automatically determined to potentially be in effect for any *Derived* dataset if a *Source* dataset containing it appears in the *Derived* dataset in question's provenance graph. Determination of the nullification of a requirement is a manual process and will not be possible for some requirements. Where not possible, the requirement is 'sticky' and will affect all descendent *Derived* datasets including products. Where possible, a nullification action can be undertaken and noted by adding an appropriate property to the `prov:Activity`[16] that generates a *Derived* dataset. Once such an action has been flagged, the entailment calculations against *Derived* datasets downstream from that Activity are deemed to no longer be bound by the requirement. The third scenario is included as it is a commonly encountered BA scenario and extends from internal systems to external users. Even though it requires communicating information to 3rd parties is handled using the same logic as scenarios 1 & 2 for Repository interface pages used to deliver information about the dataset are aware of particular `cc:Requirement` classes and alter the information they present to users accordingly.

5. CONCLUSION

The Bioregional Assessments Programme is investigating the potential impact of coal and coal seam gas on some of Australia's water resources, and has a directive to publish all data used to develop information products within the life of the programme. Correct and consistent licencing becomes crucial when datasets are published by a user of the dataset such as the BA Programme, and not its original owner. This becomes even more critical when the datasets being published are derived from multiple source datasets, some potentially with use or publication restrictions, or attributions that must be carried through to each derived child product. Additionally, organisations change over time, and many cease to exist. By separating the capture, storage and maintenance of licences, organisations and datasets, and connecting their point of truth locations through a Linked Data middleware, we are better able to manage evolution of the components, and their relationships, over time.

Provenance graphs maintained for one purpose become points-of-truth for the relationships between things and may serve as a vehicle to calculate facts about things not originally conceived of when the provenance graphs were first captured. The BA Programme has assembled detailed provenance graphs which allow the automation of licence entailment calculation to an extent that should remove the potential for error or inconsistency. This ultimately saves time and effort in handling complex dataset publication, and lowers the risk around incorrect licencing, attribution or publication restrictions.

REFERENCES

Abelson, H., Adida, B., Linksvayer M. and Yergler, N. (2008) ccREL: The Creative Commons Rights Expression Language - Version 1.0. Creative Commons Corporation. Document online at https://wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf. Accessed 2015-07-31.

Dodds, Leigh (2013) Open Data Rights Statement Vocabulary. Web Page, The Open Data Institute. Online at http://schema.theodi.org/odrs/. Accessed, 2015-07-31.

Epimorphics Ltd. (2015). Elda, an implementation of the Linked Data API. Web Page, Epimorphics Ltd. Retrieved from https://github.com/epimorphics/elda. Accessed, 2015-07-31.

Lebo, T., Sahoo, S., & McGuinness, D. (2013). *PROV-O: The PROV Ontology*. W3C web page, Online at http://www.w3.org/TR/prov-o/. Accessed, 2015-07-31.

[15] http://data.bioregionalassessments.gov.au/id/licencerequirement/

[16] `Activity` objects, as per PROV-O, are events that perform work on `Entity` objects such as (datasets.