WORKING TITLE

Semantic web-based patterns for research data discovery

Ben Greenwood
Supervisor: Kerry Taylor (CSIRO)

Work-in-Progress Seminar

Research motivation

- well-managed research data should be re-usable
- the Semantic Web and principles of Linked Data correlate well with this

“A question”:

To what extent can we harness the expressiveness of semantic web technologies to extend research data discovery capabilities beyond conventional systems?

A Semantic approach

1. Develop a base RDF model to describe research data

2. Implement mapping mechanisms to harvest discipline-agnostic data into a “data commons”

3. Implement classification/logical reasoning to support data discovery within a linked data environment

Tasks to date

- Identify and familiarise with standards and related technologies
- Research alignment between discovery use cases and benefits from semantic web technologies
- Source (or derive) data to demonstrate hypothesis ***
- Build an experimental framework and tools that can be used to test the capabilities of the proposed technologies

Software Engineering Individual Research Project (COMP4540)

"In the end, you usually create some working software which contributes to the research area, and report the results by writing a research thesis plus a short software engineering process report."

Software outcomes thus far

- Established a controlled instance of a top-of-the-range commercial RDF store and reasoning engine
- Extended an existing RDF library to work within the intended deployment environment
- Built an experiment “workbench” to visualise and manipulate test data
- Built a custom harvesting tool to derive data from existing XML-based systems into the RDF model

Editing Graph: http://purl.org/au-research/data/#first-fleet-and-early-settlement-a-digitised-suite-of-first-hand-accounts-diaries-and-letters-written-by-men-and-women-who-emigrated-to-australia-in-the-first-fleet

TTL input: (regularly used prefixes will automatically be declared)

```
@prefix for: <http://purl.org/au-research/vocabulary/anzsrc-for/2008/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

_:genid1
  a for:for-210303 ;
  rdf:value "Australian History (excl. Aboriginal and Torres Strait Islander History)" .

_:genid2
  a for:for-21 ;
  rdf:value "HISTORY AND ARCHAEOLOGY" .

_:genid3
  a for:for-2103 ;
  rdf:value "HISTORICAL STUDIES " .
```

Update Graph

Graph Representation:

Graph: http://purl.org/au-research/data/#first-fleet-and-early-settlement-a-digitised-suite-of-first-hand-accounts-diaries-and-letters-written-by-men-and-women-who-emigrated-to-australia-in-the-first-fleet

_:genid1
→ rdf:type → for:for-210303
→ rdf:value → "Australian History (excl. Aboriginal and Torres Strait Islander History)"

_:genid2
→ rdf:type → for:for-21
→ rdf:value → "HISTORY AND ARCHAEOLOGY"

_:genid3
→ rdf:type → for:for-2103
→ rdf:value → "HISTORICAL STUDIES "

_:genid4
→ rdf:type → for:for-210301
→ rdf:value → "Aboriginal and Torres Strait Islander History"

http://researchdata.ands.org.au/first-fleet-and-early-settlement-a-digitised-suite-of-first-hand-accounts-diaries-and-letters-written-by-men-and-women-who-emigrated-to-australia-in-the-first-fleet
→ dct:title → "First fleet and early settlement: a digitised suite of first-

Query Result:

?graph_name	?triple_count
rda:the-akchakhan-kala-kazakli-yaktan-kazakly-yatkan-wall-paintings-collection	"33"^^xsd:integer
rda:classic-texts-in-australian-and-international-taxation-law	"13"^^xsd:integer
rda:tin-sheds-gallery-archive	"31"^^xsd:integer
rda:history-of-nursing-and-nursing-education-archive	"58"^^xsd:integer
rda:impact-of-receptor-density-on-calcium-sensing-receptor-casr-mediated-intracellular-calcium-signalling-dataset	"32"^^xsd:integer
rda:first-fleet-and-early-settlement-a-digitised-suite-of-first-hand-accounts-diaries-and-letters-written-by-men-and-women-who-emigrated-to-australia-in-the-first-fleet	"16"^^xsd:integer
rda:national-recording-project-for-indigenous-performance-data-collection	"37"^^xsd:integer
rda:randomised-controlled-trial-dataset-efficacy-and-safety-of-ascorbic-acid-vitamin-c-supplementation-for-children-with-charcot-marie-tooth-disease-type-1a	"39"^^xsd:integer
rda:university-of-sydney-art-collections	"60"^^xsd:integer
rda:antarctic-paleobath-antarctic-paleo-depth-grids	"38"^^xsd:integer

Some challenges so far…

- Identifying a starting point that is current
- Tool and software library maturity
- Disparity between specification and implementation
- Collecting appropriate data to experiment with
- Familiarity with formal notation and experimental research techniques

Some initial outcomes

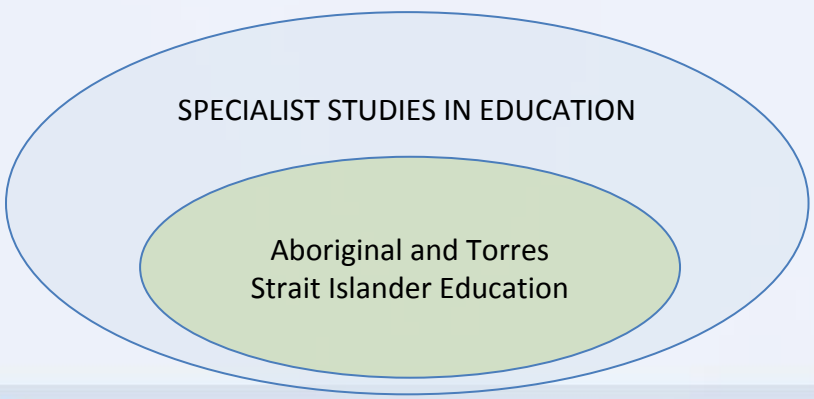
- Applied use cases demonstrate interesting patterns, illustrating:
  - the proposed benefits
  - some limitations in the approach

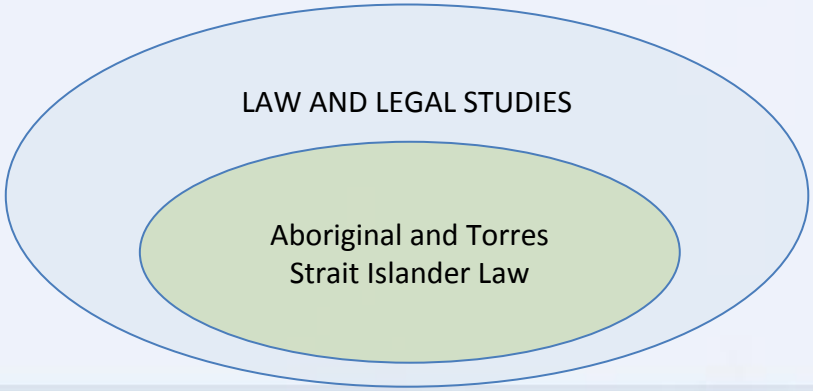
for example…

- Example pattern:

  Using the expressiveness of OWL reasoning to support discovery of interdisciplinary research data:

  Using the hierarchal ANZSRC Field of Research codes to discover relevant data that might otherwise be difficult to identify due to discipline-specific perspectives



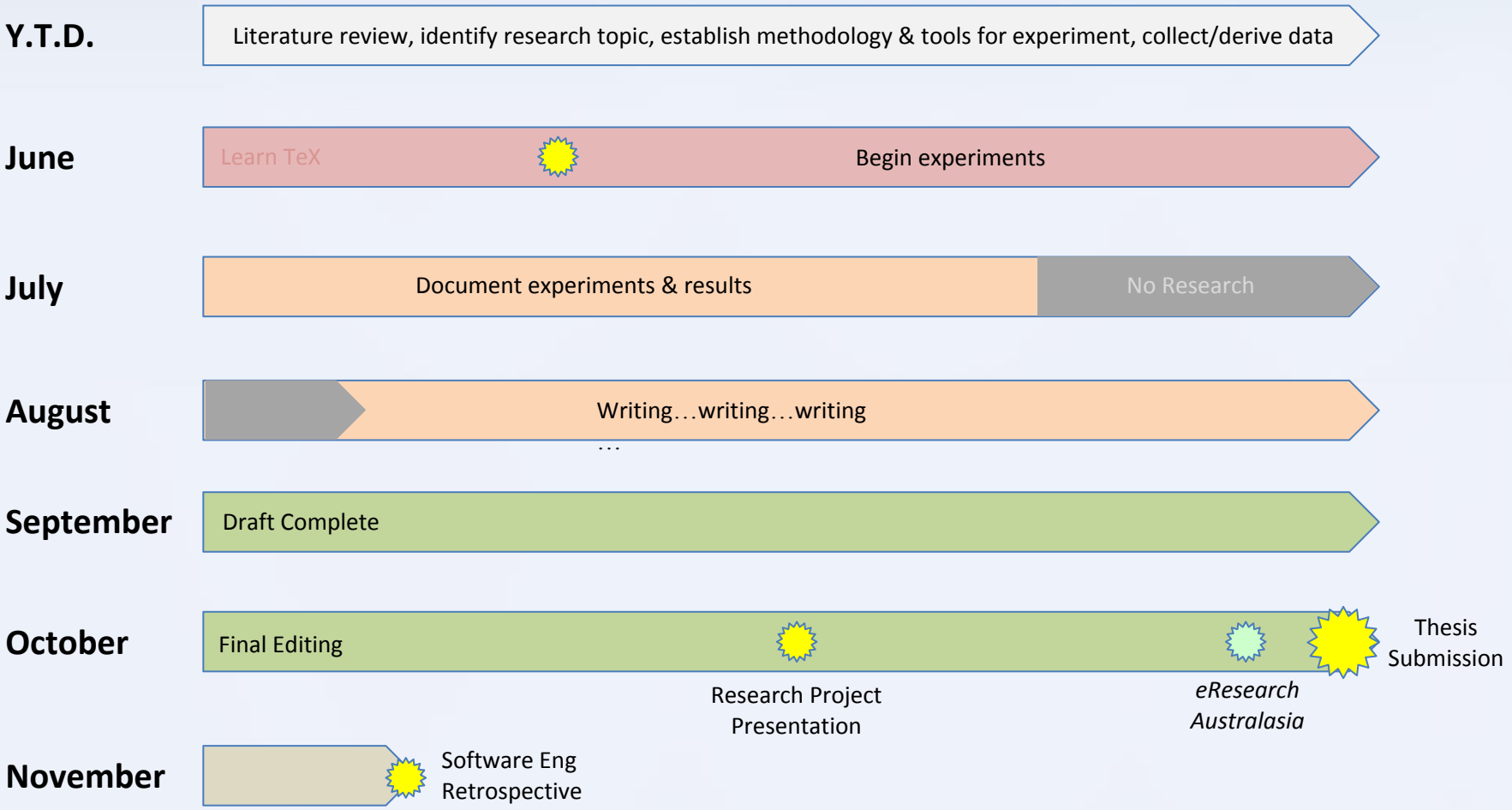


Other areas being experimented with

- Trade-offs between tractability and completeness

- Temporality
- Geo-spatial
- Authority
- Source
- Text Search

Project Schedule

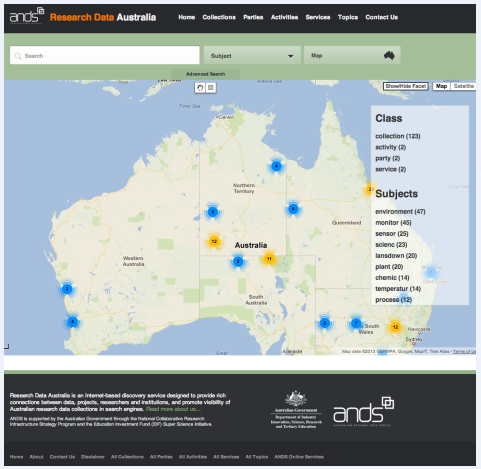


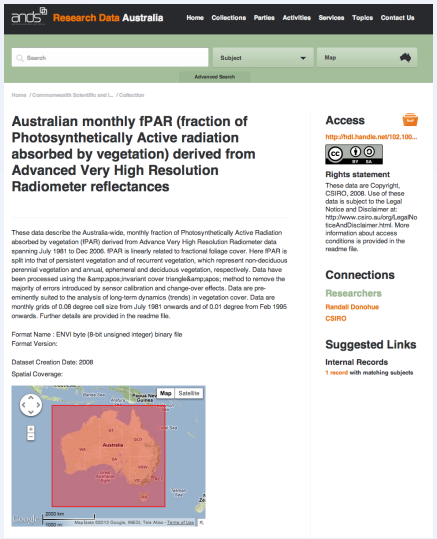
References

- Monash University Library : http://monash.edu/library/researchdata/
- W3C Semantic Web Standard : http://www.w3.org/standards/semanticweb/
- Australian Research Data Commons :http://researchdata.ands.org.au/
- Australian National Data Service :http://www.ands.org.au/

Australian Research Data Commons

A national repository and domain-agnostic discovery portal for “collections” (datasets, catalogues, etc).





Images by Australian National Data Service (CC BY 2.0)

Linked Data & the Semantic Web

TBL's "Linked Data Principles"

1. On the web
2. Machine-readable data
3. Non-proprietary format
4. Machine-useful information (RDF)
5. Linked to other URIs

Sir Tim Berners-Lee

What is Research Data Management?

- The management of data generated by and used for research. (Monash University Library, 2012)

- A systematic approach to ensuring that research data is:
  - discoverable
  - available for re-use and collaboration
  - contributes to improved future research outcomes.

How Linked Data can apply to Research Data Management

- A systematic approach to ensuring that research data is:

    1. discoverable
    2. available for re-use and collaboration
    3. contributes to improved future research outcomes.

A “Linked Data” approach can satisfy #2 and #3! So how do we make it discoverable?

Limitations of the current ARDC

- Based on an [uncommon] XML schema (RIFCS, ISO19115/139)

4.1. List of Schema Elements

accessPolicy
accessRights
activity
address
addressPart
arg
citationInfo
citationMetadata
collection
context
contributor
coverage
dates
date (*dates*)
date (*temporal*)
date (*citationMetadata*)
description (*registryObject*)
description (*relation*)
electronic
endDate
existenceDates
format
fullCitation
identifier (*citationMetadata*)
identifier (*format*)
identifier (*registryObject*)
identifier (*relatedInfo*)
key
licence
location
name
namePart (*name*)
namePart (*contributor*)
notes
originatingSource
party
physical
placePublished
publisher
registryObject
registryObjects
relatedInfo
relatedObject
relation
rights
rightsStatement
service
spatial (*location*)
spatial (*coverage*)
startDate
subject
temporal
text
title (*citationMetadata*)
title (*relatedInfo*)
url (*relatedInfo*)
url (*citationMetadata*)
value
version

4.2. Schema Element Structure

WWTBLS?

…yuck! (☹)

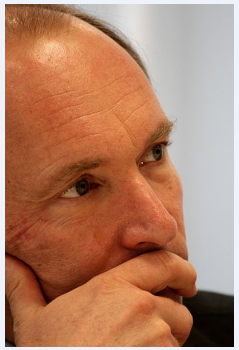


Image by Australian National Data Service (CC BY 2.0)

A Semantic Approach to Research Data Management – Ben Greenwood (2012)

Limitations of the current ARDC

- Forces a distinction between “data” and “metadata”
  - Data is only very loosely “linked”
  - The “web” can very easily become inconsistent again
  - Access to the “metadata only” violates RDM principles #2 & #3
- RIFCS is a lossy transformation from rich discipline-specific descriptions of the data

Limitations of the current ARDC

- RIFCS does not guarantee discoverability, linking:

```
<identifier type="local">ben.g.at.uni-AU</identifier>
```

- RIFCS does not support real Machine-to-Machine understanding of the data.

Semantics are limited in as much as an implementer understands the RIFCS schema elements and vocabulary.

Limitations of the current ARDC

WWTBLS?