WORKING TITLE

A Semantic Approach to Research Data Management

(in building an Australian Research Data Commons)

Ben Greenwood

Supervisor: Kerry Taylor (CSIRO)

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.

What is “semantics” ?

- Semantics concerns the study of meaning
- In the context of the World Wide Web, semantics is a method increasingly used to describe web resources:

> “building a technology stack to support a ‘web of data’”
>
> (W3C Semantic Web Standard, 2013)

We call this “web of data” the **Semantic Web**

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

(Image by Paul Clarke, Wikimedia Commons (CC BY 2.0))

How Linked Data can apply to Research Data Management

Remember my definition?

- A systematic approach to ensuring that research data is:

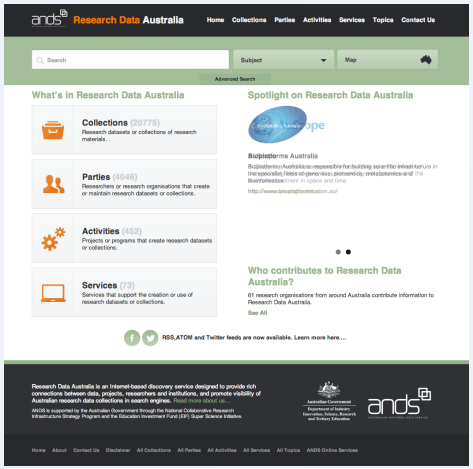
1. discoverable
2. available for re-use and collaboration
3. contributes to improved future research outcomes.

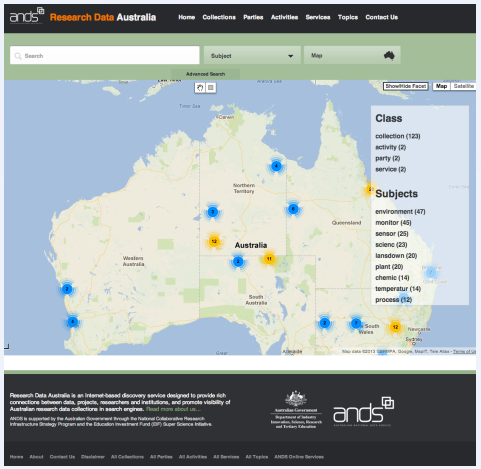
A “Linked Data” approach can satisfy #2 and #3! So how do we make it discoverable?

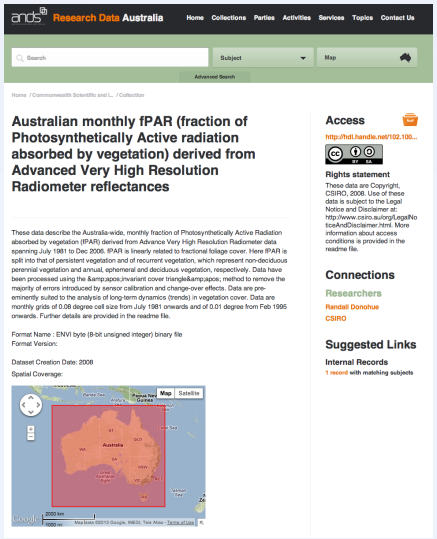
A Semantic Approach to Research Data Management – Ben Greenwood (2013)

Australian Research Data Commons

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







But...

Images by Australian National Data Service (CC BY 2.0)

A Semantic Approach to Research Data Management – Ben Greenwood (2013)

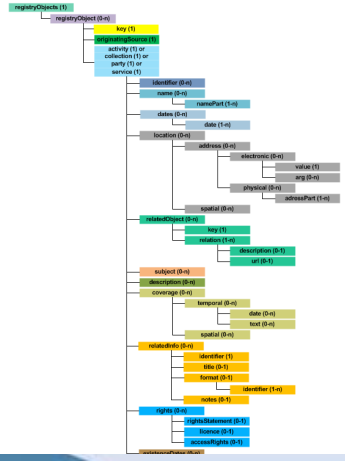
Limitations of the current ARDC

- 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! (☹)

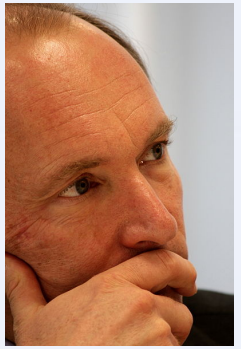


Image by Australian National Data Service (CC BY 2.0)

A Semantic Approach to Research Data Management – Ben Greenwood (2013)

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?

A Semantic approach

1. Develop a base model to describe research data

    - Using modern semantic web standards (RDF, OWL, etc.)
    - Re-using components of existing ontologies (“frankensteining” FOAF, DC, PROV, DVOC, etc.)
    - Consistent with the existing RIFCS schema, but under an “open world assumption”

A Semantic approach

2. Implement the Linked Data Platform to allow the ARDC to discover and collect data

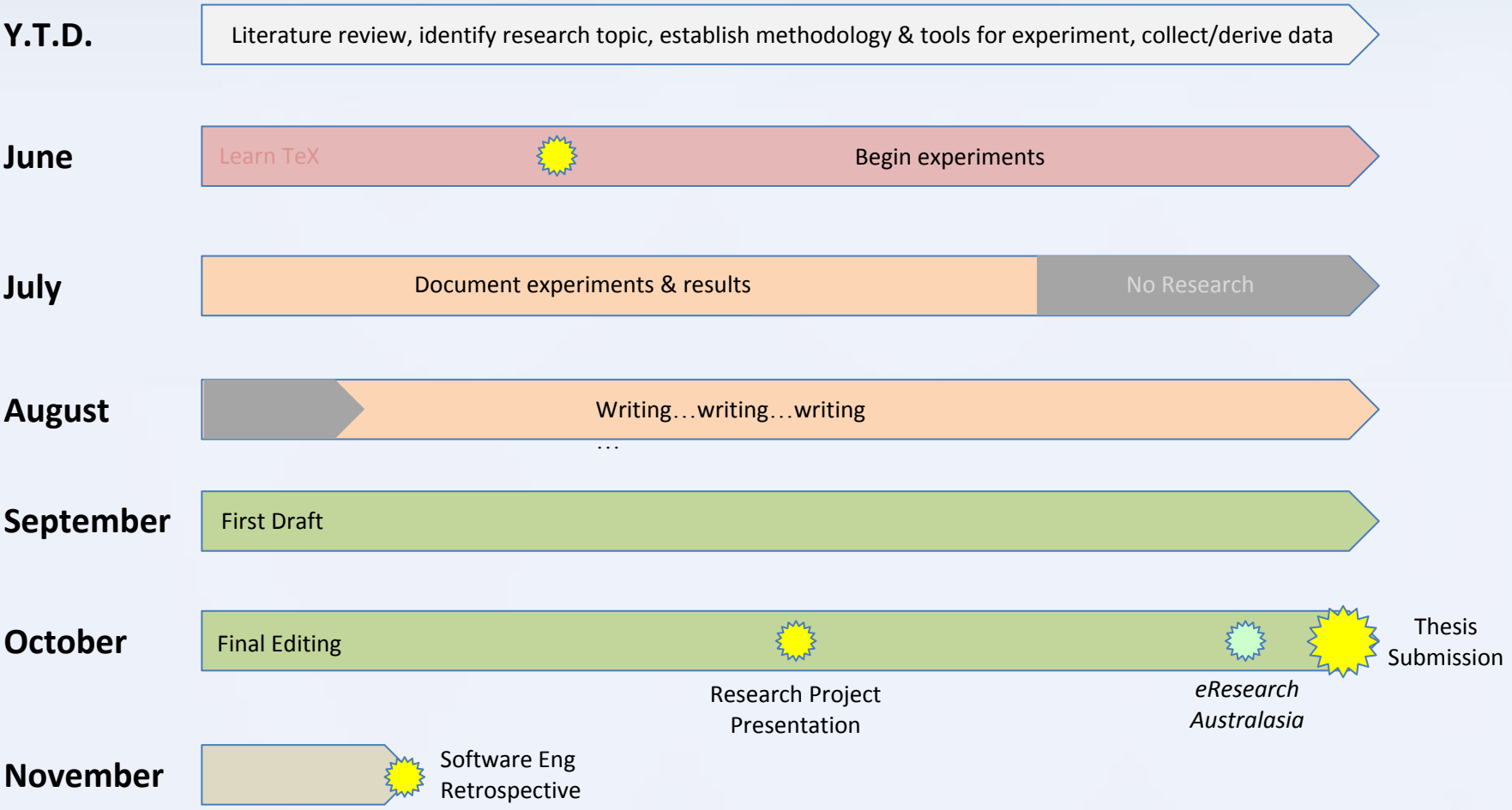
    - Using [more] semantic web technologies (LDP, REST, SPARQL, RDF, OWL reasoning)
    - Implement logical reasoning to determine equivalencies from domain- or discipline-specific ontologies
    - Provide a mechanism for “harvesting” from data sources using the discoverability properties of linked data

A Semantic approach

3. Align the “Semantic ARDC” with existing discovery technologies to facilitate adoption

    - Implement an interface/mechanism whereby existing services can adopt the new capabilities with a low barrier to entry (i.e. RESTful APIs, OAI-PMH)
    - Provide mechanisms to ensure consistency within the ARDC
    - Provide materials and training to assist implementers

Schedule


Y.T.D.
Literature review, identify research topic, establish methodology & tools for experiment, collect/derive data
June
Learn TeX
Begin experiments
July
Document experiments & results
No Research
August
Writing…writing…writing
…
September
First Draft
October
Final Editing
Thesis Submission
Research Project Presentation
eResearch Australasia
November
Software Eng Retrospective

My research project

- Undertake research to tackle steps #1 and #2

  1. Develop a base model to describe Research Data

  2. Implement the LDP (+ reasoning) to allow the ARDC to discover data

  The thesis will centre around an example use case (possibly the BOM Environment Information Initiative) and evaluation this against certain quality measures (the ANDS Data Quality Requirements, TBL's 5 principles, performance, etc).

- Part #3 may be undertaken concurrently within ANDS, but will not form part of the project deliverable (other than for illustrative purposes).

Project Goals

- An prototype and presentation at eResearch Australasia 2013
  (9 June for presentation/workshop or 1 October for poster only)
- Possible submission to International Semantic Web Conference?
  (poster/demo / in-use track – submissions May 17)
- Complete COMP4540 ☺
  (mid- and end-year seminar presentations, a thesis report, SEng retrospective)

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/