Australian Research Data Commons

Guidelines for publishing structured metadata on the Web

Mingfang Wu

Australian Research Data Commons, Australia

7 October 2021
DCMI Virtual

The Australian Research Data Commons is enabled by NCRIS.

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

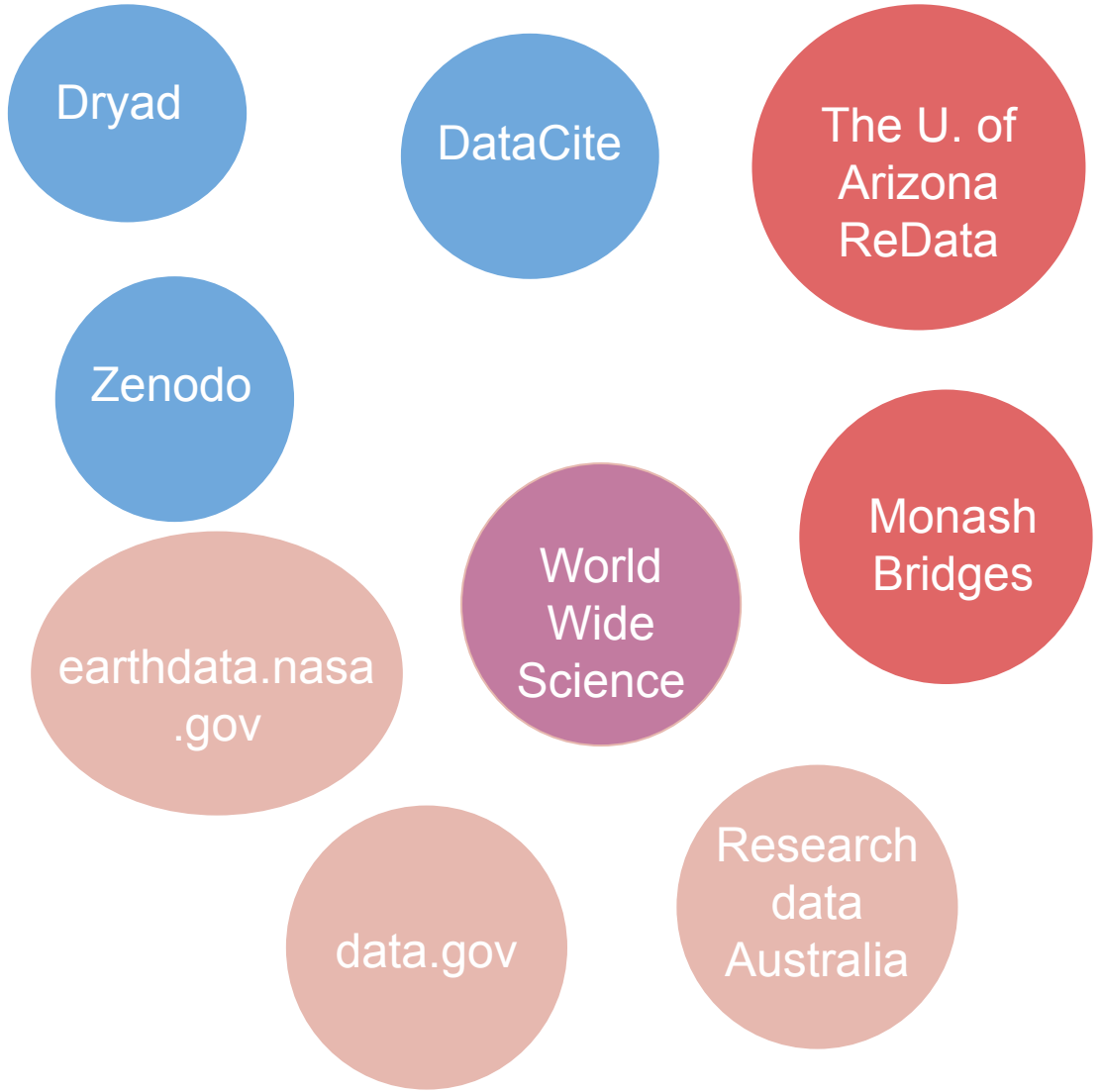
Outlines

- Motivation to publish structured metadata on the Web
- Guidelines for publishing structured metadata on the Web

Research data repository/catalogue

There has been a growing number of data repositories & catalogues for publishing and sharing data

re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years and, by February 2020, the registry had more than 2450 repositories.



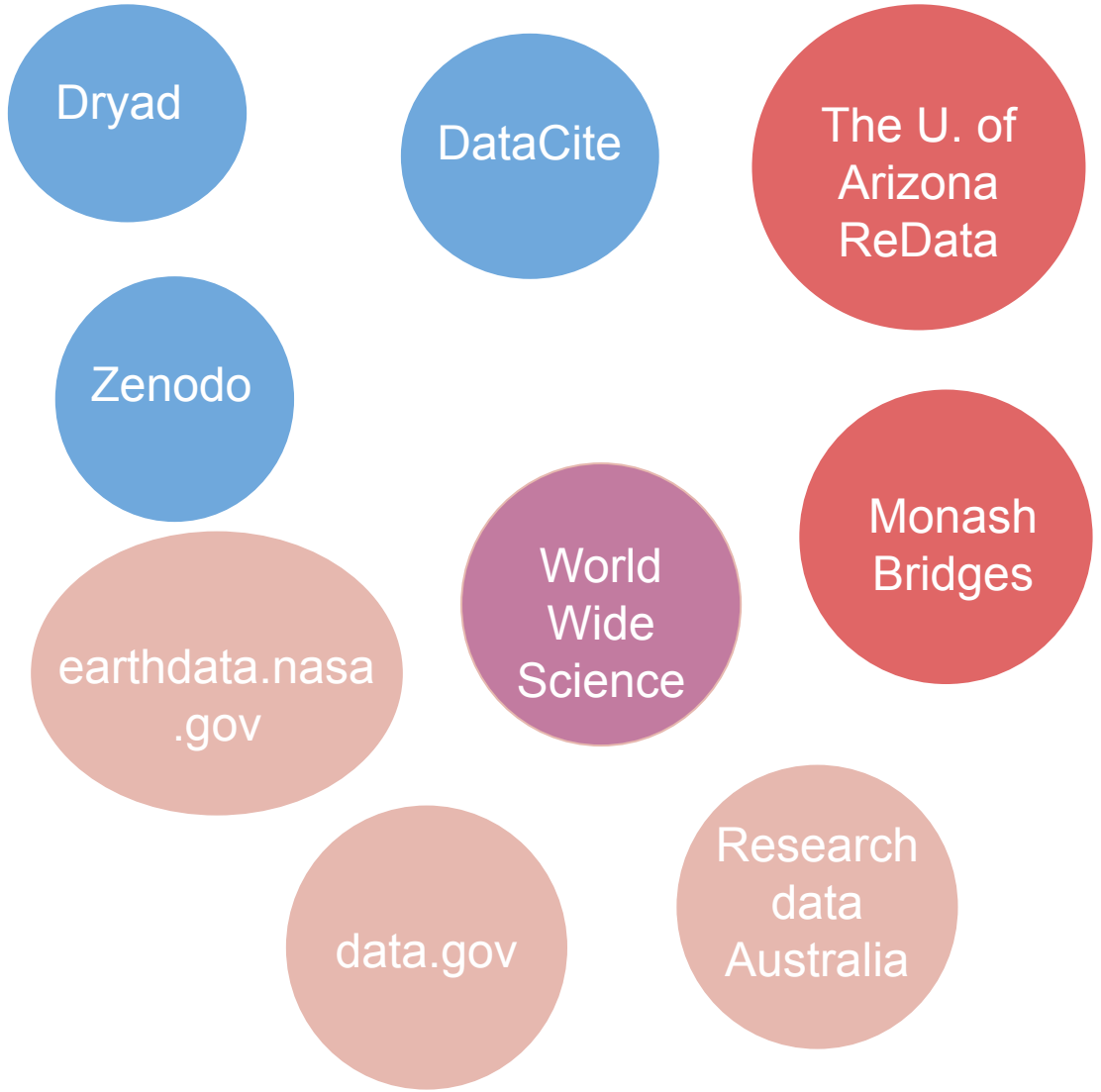
Research data repository/catalogue

Challenge:

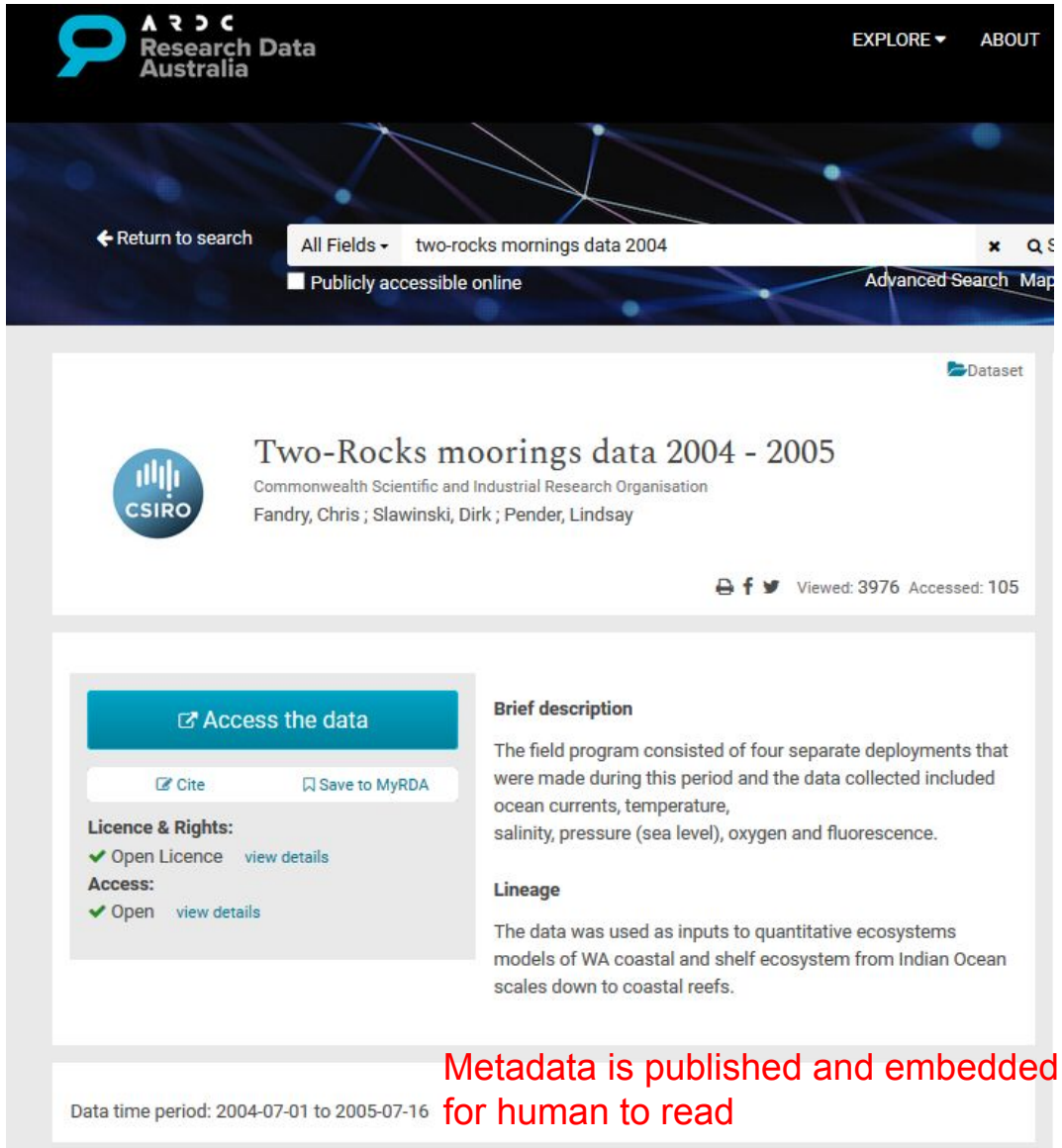
- It becomes harder for repositories exchanging metadata as each repositories may adopt a metadata schema to meet special need of their repository and user community.
- It gets harder for researchers and the public to find relevant data, especially where desired data is dispersed across several repositories.
- Metadata is published at a webpage (landing page) that is designed for human not machine, this limits the development of innovative data discovery applications.

A solution:

- Leverage commercial search engines like Google, Yahoo, Bing etc. to facilitate broader discovery of, and access to, research data.
- Mark up metadata with Schema.org vocabulary and publish marked metadata (structured metadata) in metadata landing page.



What is structured metadata?



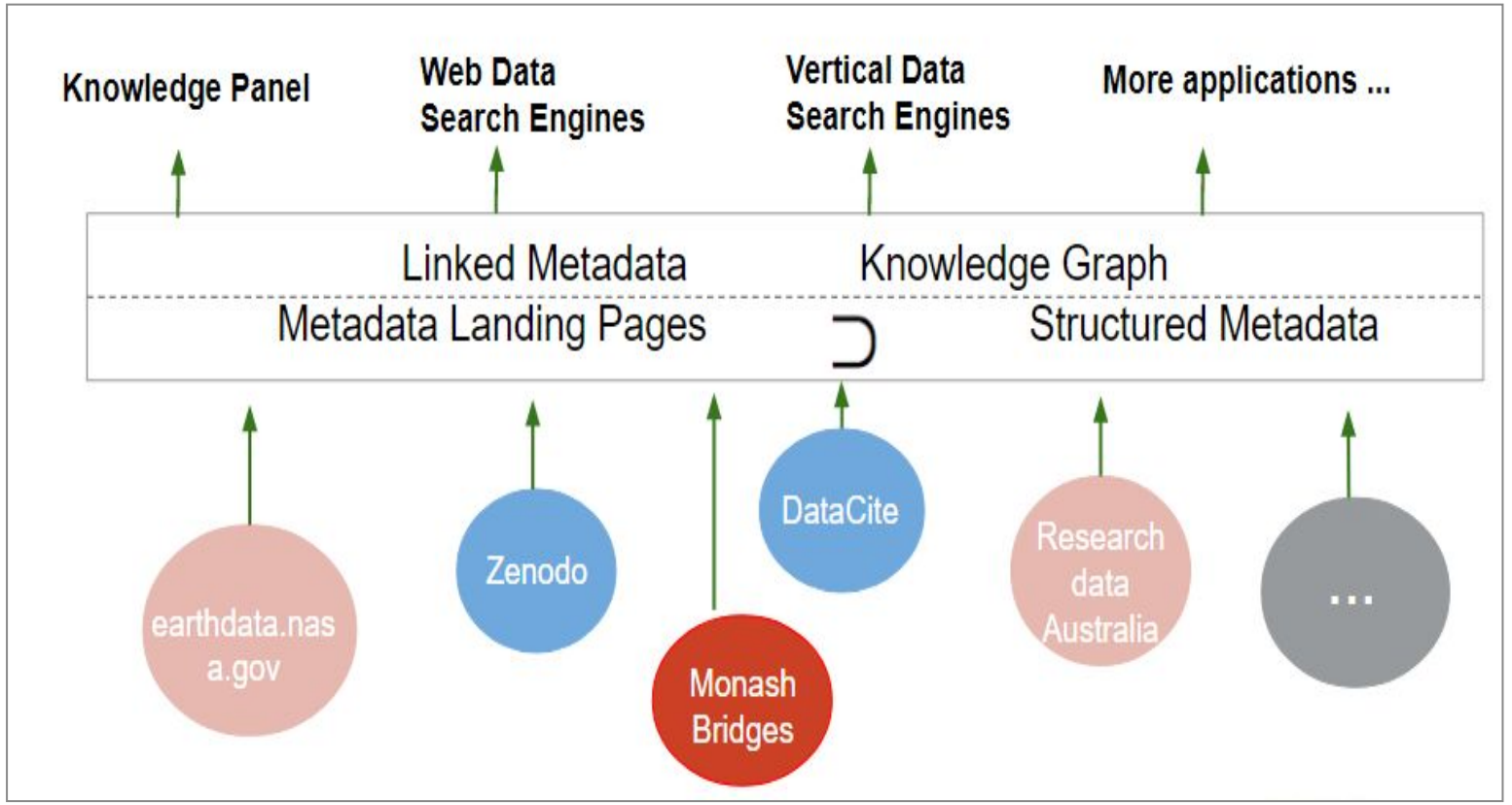
Metadata is published and embedded in a webpage for human to read

Structured (meta)data / Markup

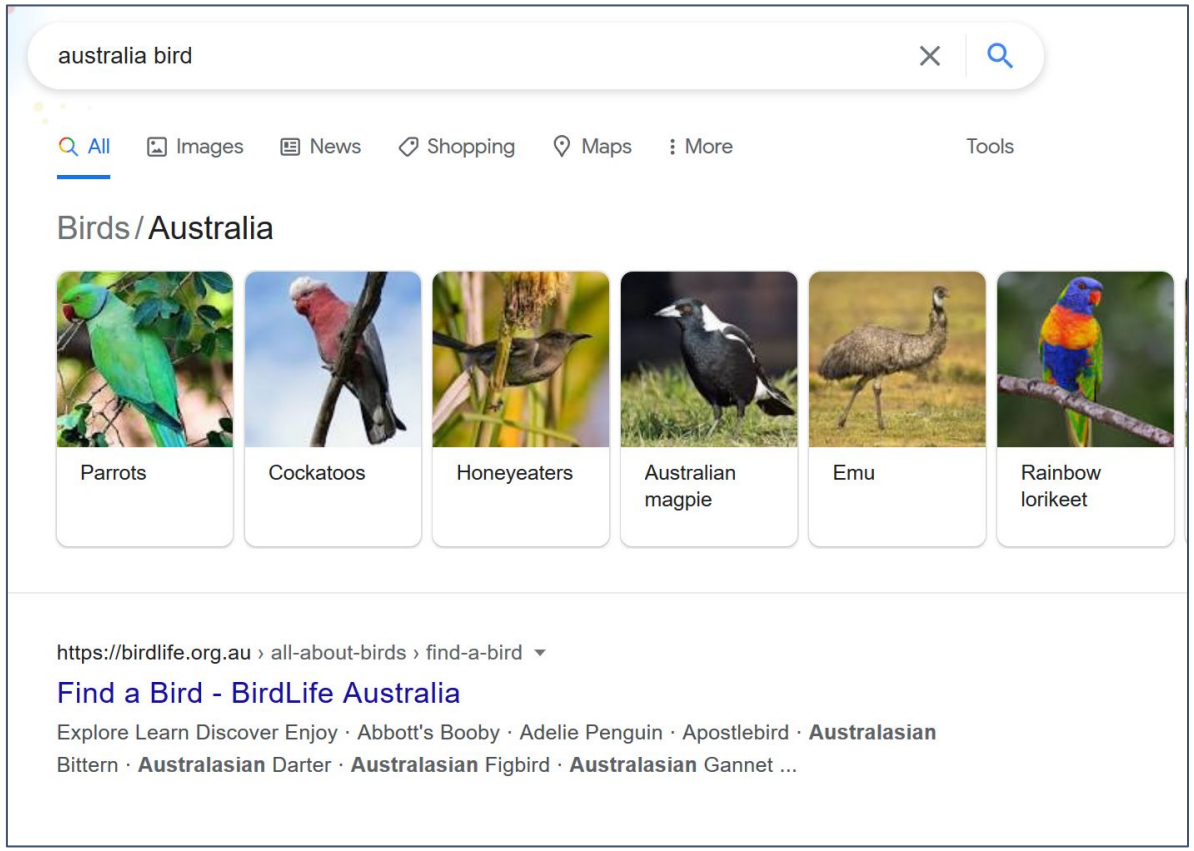
```
<html>
...
<head>
<script type="application/ld+json">
{"@context":"http:\/\/schema.org\/",
"@type":"Dataset","datePublished":"2016-03-22",
"sourceOrganization":{"@type":"Organization","name":"data.gov.au"},
"keywords":" Australia, Cooper subregion, Galilee subregion, Hunter subregion, Namoi subregion
"license":["Birds Australia freely allows non-commercial copying and distribution of the Impor
"description":["## **Abstract** \n\nThis dataset and its metadata statement were supplied to t
"identifier":[{"@type":"PropertyValue","propertyID":"local","value":"bb67a0fd-0aac-433c-992b-b
        {"@type":"PropertyValue","propertyID":"URL","value":"http:\/\/data.gov.au\/datas
"url":"https:\/\/researchdata.ands.org.au\/birds-australia-important-iba-2009\/657761"}
</script>
</head>
...
</html>
```

Benefit

- Enhance interoperability across data catalogues
- Make data more discoverable
- Enable federated search across repositories of a specific domain or related domains relevant to a research need
- Support a spectrum of data search needs from text/keyword search, JSON API to SPARQL queries
- Provide a potentially new method for metadata/content syndication among repositories



Example use of structured data



Example use of structured metadata

rock mooring data

All | Images | Shopping | News | Videos | More | Tools

About 59,100,000 results (0.81 seconds)

Scholarly articles for rock mooring data

The East Cape Current and two eddies: a mechanism … - Chiswell - Cited by 104

… the design of laterally loaded anchor piles in soft **rock** - Erbrich - Cited by 22

https://data.gov.au › dataset › ds-dap-csiro:5268

Two-Rocks moorings data 2004 - 2005 | Datasets | data.gov.au

8 Apr 2013 — Two-**Rocks moorings data** 2004 - 2005 ... were made during this period and the data collected included ocean currents, temperature,

https://www.coris.noaa.gov › activities › resources PDF

Mooring Buoy Planning Guide

by P Rope — Mushroom anchors also, are best suited for use in muddy or silty environments, but not in coarse sand, hard mud or clay, or **rocky** conditions. The mushroom's ...

92 pages

https://www.transport.wa.gov.au › imarine › ml-two-ro...

Two Rocks - Department of Transport

Bunbury, Casuarina Boat Harbour **moorings**. Bunbury, Casuarina Boat Harbour pens. Exmouth boating facility. Fremantle Challenger Harbour. Fremantle Fishing Boat ...

https://www.westmarine.com › WestAdvisor › Construc...

Constructing a Permanent Mooring | West Marine

29 May 2020 — Mushroom anchors work best in a silt or mud bottom, and are not as effective in **rocks** or coarse sand. If a mushroom gets pulled out of the ...

rock mooring data

Last updated | Download format | Usage rights | Topic | Free

95 data sets found

Two-Rocks moorings data 2004 - 2005
researchdata.edu.au
data.csiro.au
Updated Aug 5, 2013

IMOS - ANMN Western Australia (WA) Two Rocks...
researchdata.edu.au
data.wu.ac.at
Updated 2009

Two-Rocks moorings data 2004 - 2005
data.gov.au
Updated Apr 8, 2013

IMOS - ANMN Western Australia (WA) Two Rocks 50...
researchdata.edu.au
researchdata.ands.org.au
Updated Jan 3, 2018

Data from: SRFME: Two-Rocks moorings July 2004 - July...
researchdata.edu.au
Updated Oct 11, 2006

IMOS - ANMN Western Australia (WA) Mooring Sub-...
data.gov.au
html
Updated Jun 19, 2017

Two-Rocks moorings data 2004 - 2005

Explore at researchdata.edu.au | Explore at data.csiro.au

Unique identifier
https://doi.org/10.4225/08/50F624325B0A7

Data set updated Aug 5, 2013

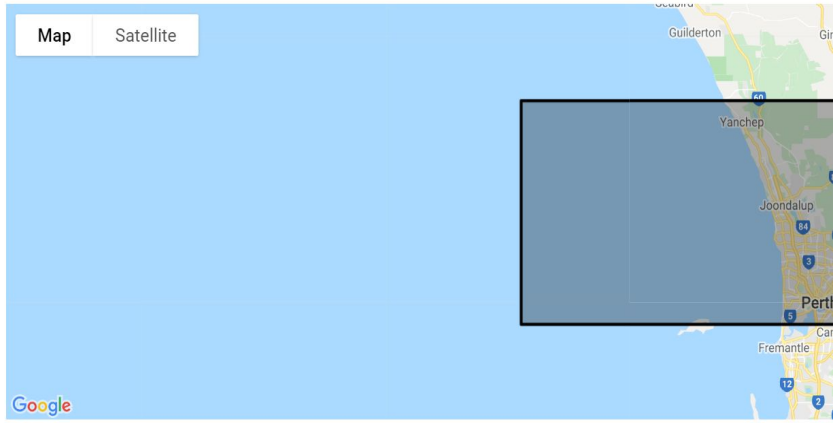
Data set provided by
CSIRO

Authors
Fandry Chris; Slawinski Dirk; Pender Lindsay

Licence
Attribution 3.0 (CC BY 3.0)
Licence information was derived automatically

Time period covered
Jul 1, 2004 - Jul 16, 2005

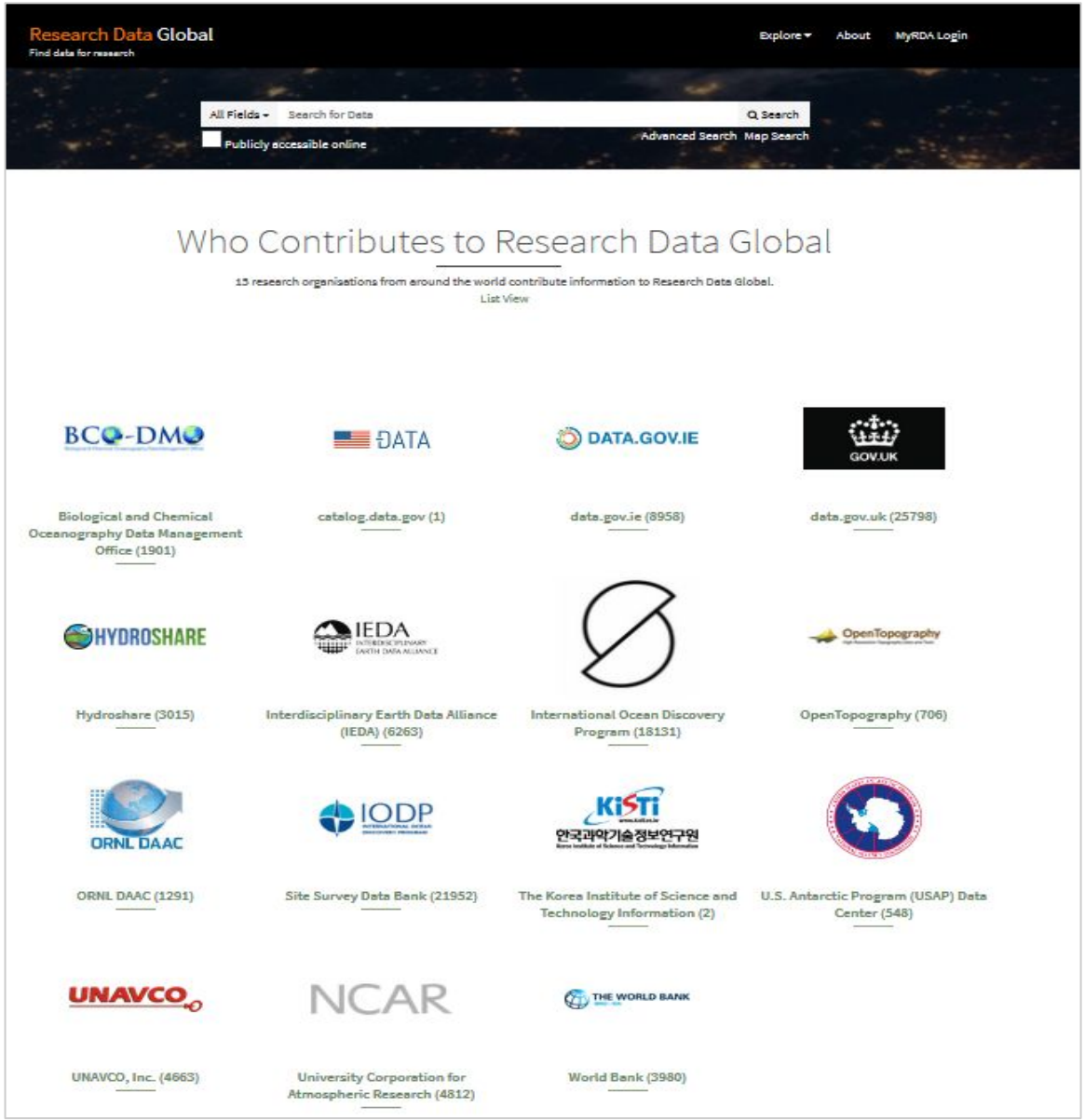
Area covered

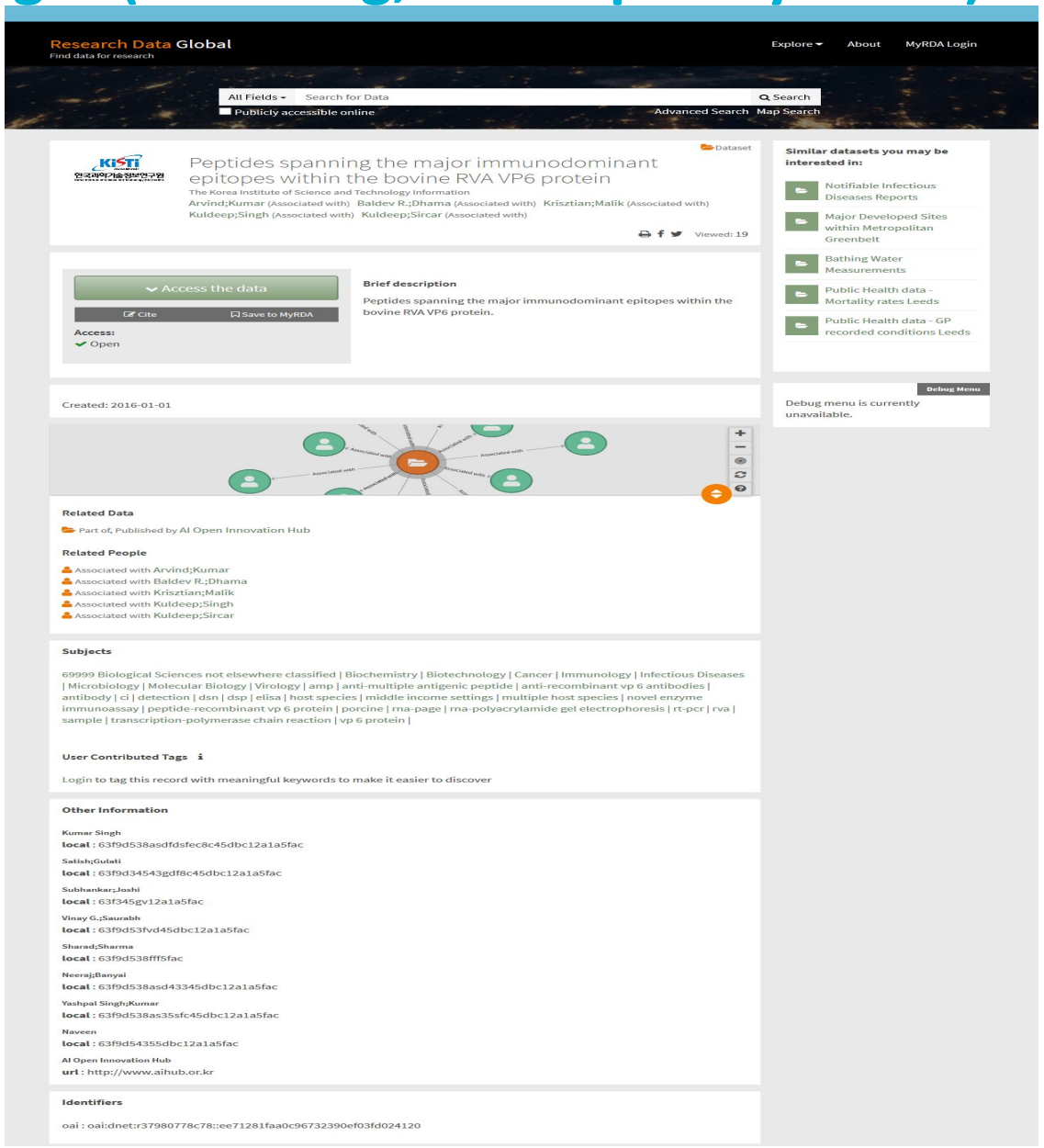


Description

The field program consisted of four separate deployments that were made during this period and the data collected included ocean curr

Research Data Global:
Harvest structured metadata from landing pages (in testing, developed by ARDC)





Introduce the Guidelines for publishing structured metadata on the Web

- Motivation: Adoption of structured metadata within and across domains needs to follow a consistent implementation across data repositories in order to enhance interoperability and discoverability of data, therefore improve accessibility and reusability.

- Research Data Alliance (RDA) Research Metadata Schemas Working Group was formed in 2019, to bring together communities who are working on or desire to published structured metadata about data, software, learning materials and other resources related to research resources. (RDA is an internationally community driven initiative for building the social and technical infrastructure to enable open sharing and re-use of data.)

- The working group have been working on and writing up the “Guidelines for Publishing Structured Data on the Web”.

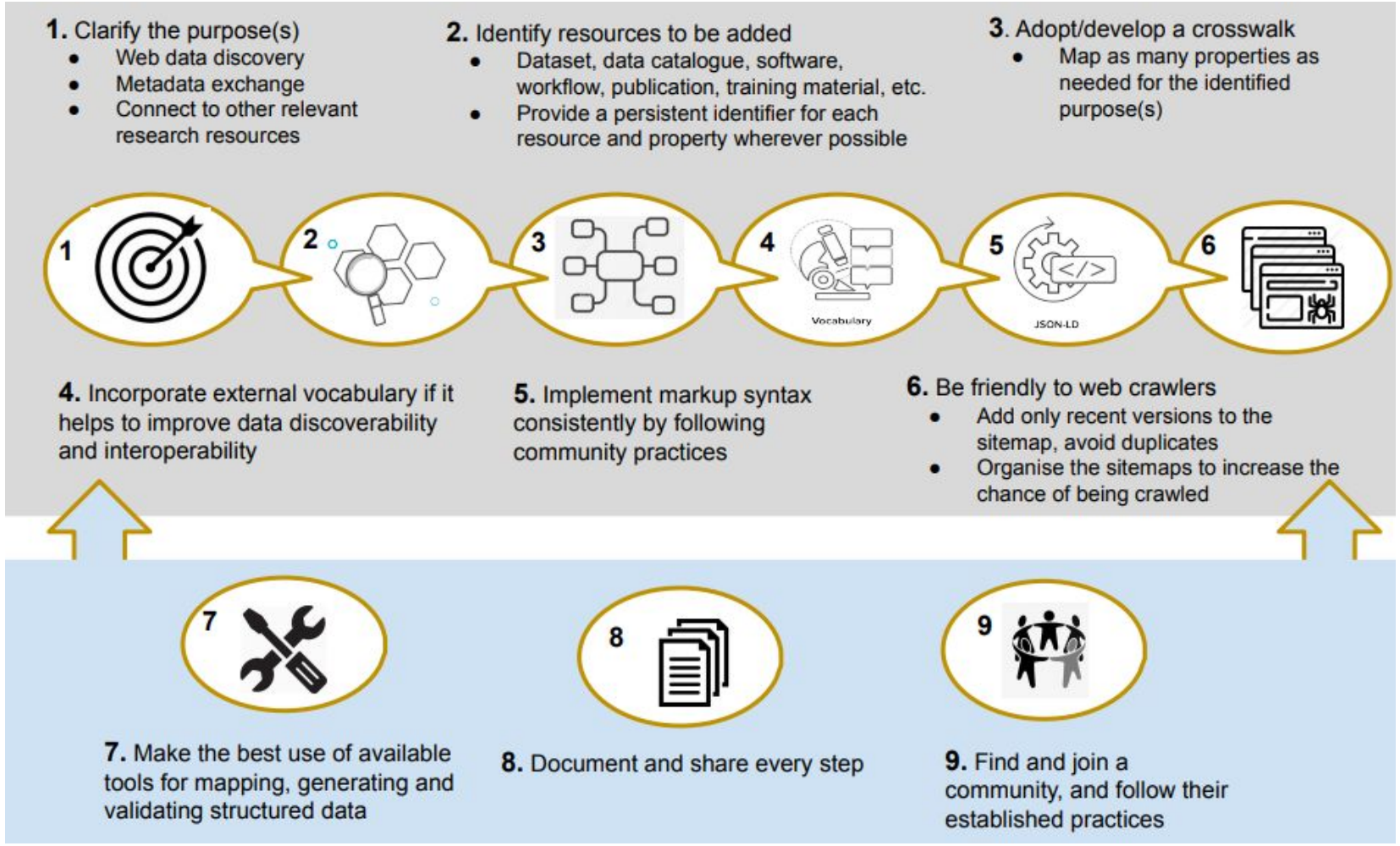
    - Data repository operators can make their data more discoverable over the web

    - Publishers can effectively expose their metadata and benefit from it

    - Aggregators can easily access and use markup provided by consumers

    - Contributing to FAIRness (particularly for metadata)

Nine recommendation for publishing structured data



The guidelines document is available at: https://bit.ly/3wwH6EY

Analysis of crosswalks from 14 metadata schemas to Schema.org (1)

Collected crosswalks to Schema.org from 14 schemas including DC, DCAT, DCAT-AP, DataCite, RIF-CS, Dataverse, DDI, SPASE, ISO-19115:2003, et al.

The crosswalks are available at: https://github.com/rd-alliance/Research-Metadata-Schemas-WG

	Schema.org				DCAT-AP	DCATv2	Datacite M: Mandatory R: Recommended O: Optional	ISO-19115:2003	EOSC/EDMI M: Minimum R: Recommended O: Optional	Dataverse	DATS	RIF-CS
	Parent Type	Property R: Required	Type	Description (text in bracket is from the Google dataset search guide)								
A. From Google dataset search recommendaton	schema:Thing	description (R)	Text	A description of the item.	dct:description(M)	dcterms:description	Description (R)	Resource abstract	description (M)	Description; DescriptionValue (M)	description (R)	collection/description[@type OR collection/description[@type
	schema:Thing	name (R)	Text	A descriptive name of the item (e.g. dataset, software, Organization).	dct:title(M)	dcterms:title	Title (M)	Resource title (M)	name (M)	Title (M)	title (M)	registryObject:collection:nam
	schema:Thing	identifier	PropertyValue or URL	The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	dct:identifier	dcterms:identifier	Identifier (M)	Resource identifier	identifier (M)	Dataset Persistent ID	identifier (R)	collection/citationInfo/citation identifier AND/OR collection/identifier
	schema:Thing	alternateName	Text	An alias for the item (Alternative names that have been used to refer to this dataset, such as aliases or abbreviations.)		rdfs:label skos:altLabel				Alternative title		
	schema:Thing	sameAs	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Wikidata entry, or official website.		owl:sameAs skos:exactMatch	RelatedIdentifer(R) AlernateIdentifier(O)*		sameAs (O)	Alternative URI	alternateIdentifier (O)	
	schema:Thing	url	URL	URL of the item. (Location of a page describing the dataset.)	dcat:landingPage(O)	dcat:landingPage	valueURL	Online resource (function: information")	url (M)		identifier (R)	RDA key= URL of record. "https://researchdata.ands.o ?key=<insert key>"
	schema:CreativeWork	citation	CreativeWork or URL	A citation or reference to another creative work, such as another publication, web page, scholarly article, etc.		dcterms:bibliographicCitation	citation/isBasedOn		citation (O) referenceCitation(O)	Related Publication	primaryPublication/ citations (O)	collection/relatedInfo[@type on"] OR relatedObject:collection whe

Analysis of crosswalks from 14 metadata schemas to Schema.org (2)

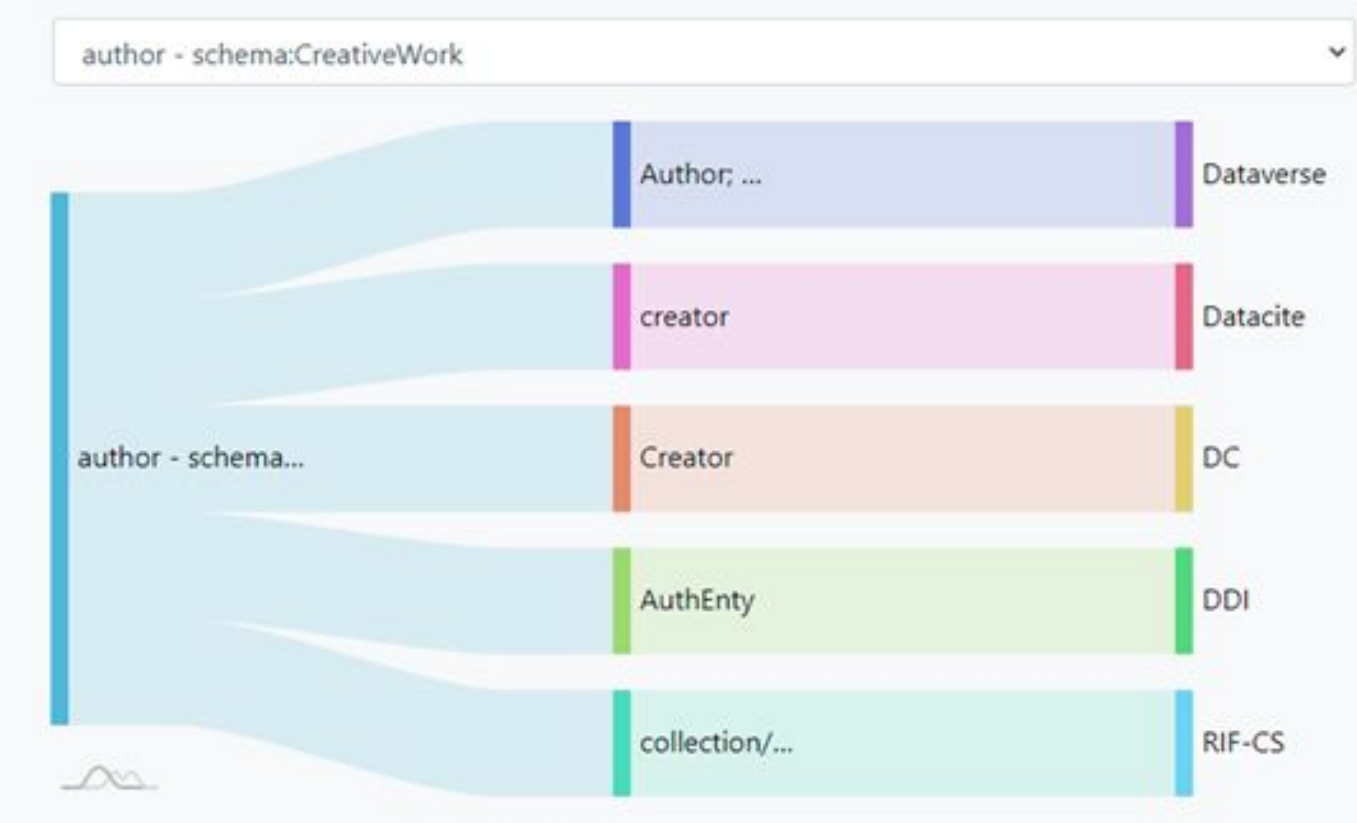
In total, about 384 properties from the 14 crosswalks are mapped to the 41 Schema.org properties.

One DC element “rights” doesn’t have a proper mapping in Schema.org

19 out of 55 dcterms don’t have a proper mapping in Schema.org: accrualMethod, accrualPolicy, audience, conformsTo, dateAccepted, dateSubmitted, educationLevel, extent, **hasVersion**, instructionalMethod, isRequiredBy, **isVersionOf**, issued (published?), medium, **provenance**, requires, rights, tableOfContents, valid

NISO Metadata Type		**Schema.org properties** (The numbers in bracket indicates the number of crosswalks that have a term mapped to the schema.org property)
Descriptive metadata: For finding or understanding a resource		Identifier (14), Name (14), description (14), Creator (14), alternameName (9), datePublished (13), version (13), keywords (12), about or subjectOf (8), inLanguage (10), temporalCoverage (11), spatialCoverage (11), variableMeasured (6), publisher (12), contributor (10), funder (9), producer (8)
Admin. Metadata	**Technical metadata** For decoding and rendering files	encordingFormat (13), contentSize (6), measurementTechnique (6)
	Rights metadata Intellectual property rights attached to content	License (12), copyrightHolder (3), isAccessibleForFree (6)
	Preservation Metadata Long-term management of files	contentUrl (7), URL (14), distribution (9), contactPoint (10), copyrightYear (5), dateCreated (10), dateModified (11), expectedArriveFrom(2), expectedArriveUntil(2), repeatFrequency (5), includeInDataCatalog (8)
Structure metadata Relationships of parts of resources to one another		citation (12), sameAs (8), mentions (3), isBasedOn (6) isPartOf (10), hasPart (10), isRelatedTo (7)

Visualisation of schema crosswalks



Filter Table Data: title

Standard	Term	Schema.org crosswalk	Parent Schema
ISO-19115:2003	Resource title(M)	name	schema:Thing
Dataverse	Title(M)	name	schema:Thing
DCAT	title(M)	name(R)	schema:Thing
DATS	title(M)	name	schema:Thing
Datacite	Title(M)	name	schema:Thing
RIF-CS	registryObject:collection:name (Title as displayed in RDA)	name	schema:Thing
DC	Title - dcterms:title	name	schema:Thing

https://rd-alliance.github.io/Research-Metadata-Schemas-WG/

Developed by by the World Data System- International Technology Office, Canada

Guidelines - with examples to support good practice

Suboptimal example	Good practice example

"creator":"Peter Smith"
 	
Or acceptable example:
 	
"creator":{
"@type":"Person",
"givenName":"Peter",
"familyName":"Smith"
}
	"creator":{
	"@type":"Person",
	"givenName":"Peter",
	"familyName":"Smith",
	"sameAs": "http://orcid.org/0000-0000-0000-0000"
	}

ESIP Schema.org cluster provides a comprehensive list of markup encordings:
https://github.com/ESIPFed/science-on-schema.org

Suboptimal example	Good practice example

"keywords": ["geology", "soil sciences"]	"keywords":[
	{
	"@type":DefinedTerm",
	"url":"http://purl.org/au-research/vocabulary/anzsrc-for/2008/0403",
	"Name":"geology",
	"termCode":"0403",
	"inDefinedTermSet":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept"
	},
	{
	"@type":"DefinedTerm",
	"url":"http://purl.org/au-research/vocabulary/anzsrc-for/2008/0503",
	"name":"Soil Sciences",
	"termCode":"0503",
	"inDefinedTermSet":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept"
	},
	{
	"@type":"DefinedTermSet",
	"url":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept",
	"name":"ANZSRC Field of Research Vocabulary Service (ABS 1297.0)"
	}
	]

Next step ...

- The working group will focus on adoption of the guidelines, as well as revision through feedback from the adoptions.

Acknowledgement

- The support provided by the Research Data Alliance community and structures.

- Contributions from members of the Research Data Alliance Research Metadata Schemas Working Group, especially:
  - Co-chairs of the WG:
  - Co-authors of the guidelines: Leyla Jael Castro, Adam Shepherd, Sarala Wimalaratne

    Nick Juty, Julia Collins, Ruth Duerr, Chantel Ridsdale, Adam Shepherd, Chantelle Verhey, Leyla Jael Castro

Join the RDA Research Metadata Schemas Working Group (and/or other Communities of Practice) for following on the community discussion and contributing to the guidelines

- You can join the group if you are an RDA member:
  - Log in to the RDA Web site with your RDA userid/password, go to the Group page, and press the Join group button on the right (near the top of the page).
    https://www.rd-alliance.org/groups/research-metadata-schemas-wg
- If you are not an RDA member, you can join RDA here: https://rd-alliance.org/user/register (it's free!)

Australian Research Data Commons

Thank you

Contact:
mingfang.wu@ardc.edu.au

www.ardc.edu.au

australian-research-data-commons

ARDC_AU

The Australian Research Data Commons is enabled by NCRIS.

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative