LDaCA

Language Data Commons of Australia

A new approach to Machine Actionable Schemas for linked-data, simplifying schema and profile development for metadata systems

Rosanna Smith
Peter (PT) Sefton

ARDC Australian Research Data Commons · NCRIS National Research Infrastructure for Australia · Australian National University · Batchelor Institute · First Languages Australia · QUT · The University of Melbourne · The University of Sydney · The University of Queensland

What is LDaCA

Making nationally significant language data available for use by researchers and communities.

Providing a model for ensuring continued access with appropriate community control.

Showcasing Australia's linguistic diversity
Australia is a massively multilingual society with the world's oldest continuous cultures situated in one of the most linguistically diverse regions in the world

Securing language data
Large collections of language data have been amassed in Australia, but many remain under-utilised or at risk

Computational analysis
Growing demand for text analytics capability and computational methods

2024-2028 Project collaborators



Rose Barrowcliffe
University of Queensland / Macquarie University



Michael Haugh
University of Queensland



Catherine Travis
ANU



Monika Bednarek
University of Sydney



Martin Schweinberger
University of Queensland

First Languages Australia



Jenny Fewster and team
ARDC



Marissa Takahashi
QUT Australian Digital Observatory



Curtis Roman
Batchelor Institute



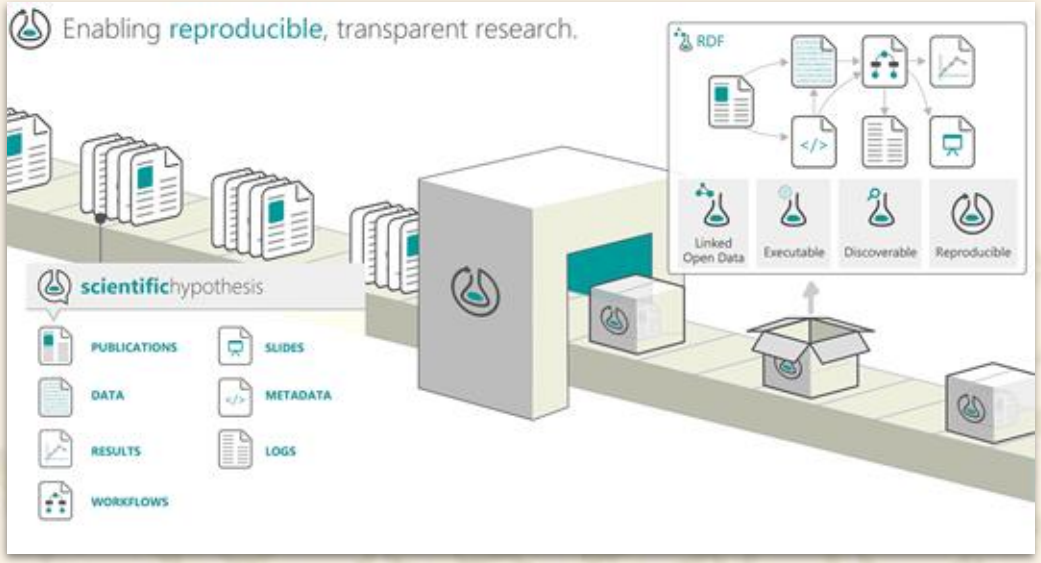
Nick Thieberger
University of Melbourne

LANGUAGE DATA CHALLENGES

Collect & Organise	Language data is rarely organised or described in reusable ways, if it's described at all
Preserve	A lot of language data is at risk of being lost forever
Find	It's difficult to know what language data exists and where to find it
Access	Processes for granting permissions and getting access to data are either absent or aren't easy to understand or apply
Analyse	Ad hoc tools, analysis and annotation methods are used, hampering reproducibility
Guide	Guides for collecting, handling and using data are scattered and hard to find

WHAT IS RO-CRATE?

RO-Crate

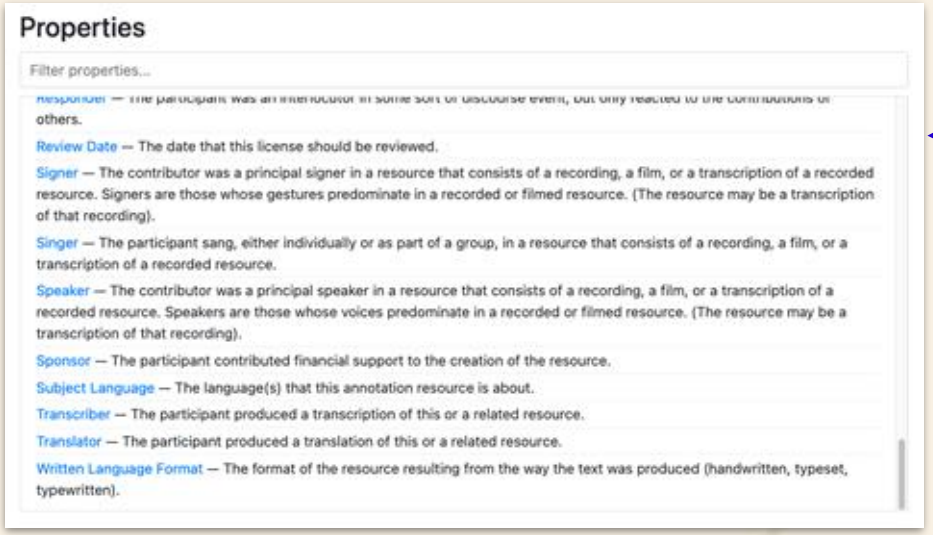


RO-Crate is a spec for packaging and describing linked data.

Linked data is an extensible metadata format that uses URLs so everything is documented.

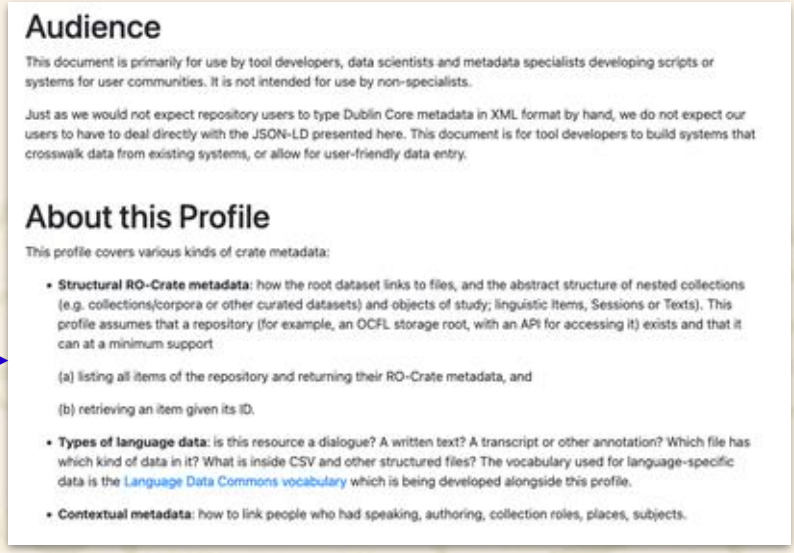
Accessible, usable approach to linked data.

SCHEMAS & PROFILES



Schemas specify a metadata vocabulary of classes, properties and other terms, and may include additional specifications and requirements for particular terms.

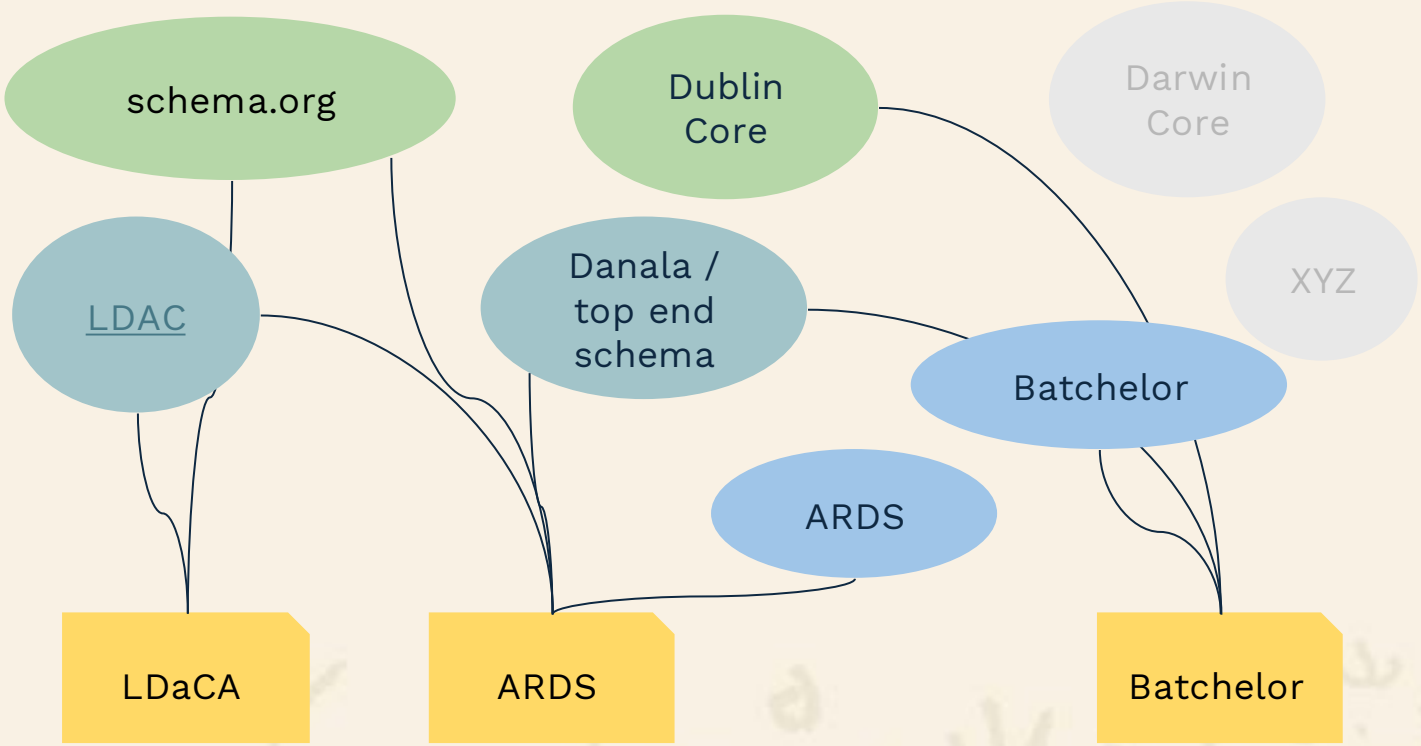
Profiles describe how a schema is applied to a particular domain or use case, and can take metadata from multiple schemas. They are prose documents that provide more guidance on structuring collections, required metadata, access protocols and other additional details.



general

SCHEMAS
groups of
properties

specific



PROFILES
collections of
properties from
different
schemas

Schemas and profiles can be published, publicly or privately, and can have terms and conditions, and licences for how they can be used/shared.

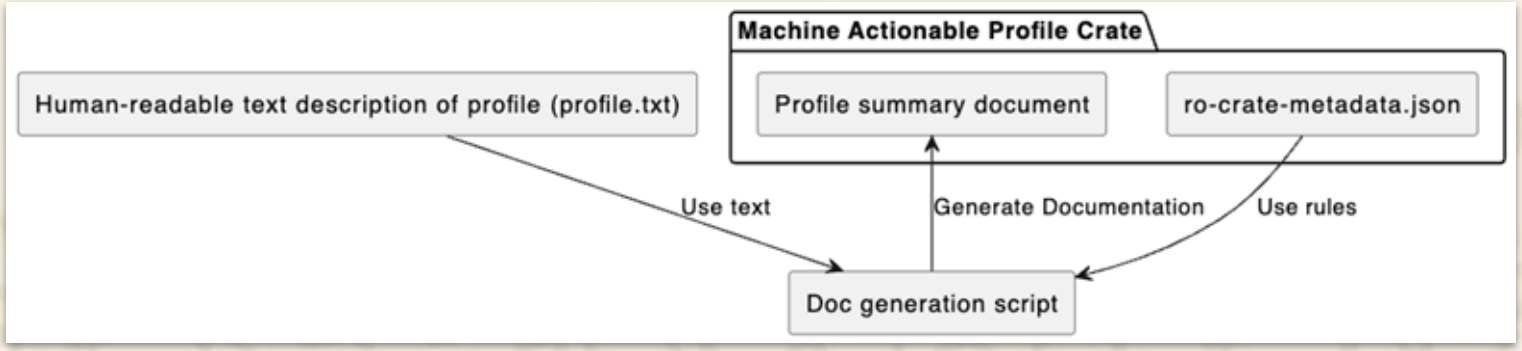
MASP DETAILS

Machine Actionable Schemas and Profiles - generates schema and profile documentation.

github.com/Language-Research-Technology/ro-crate-masp

Applicable to a wide range of metadata description, e.g. LDaCA, ARDS Aboriginal Corporation, Batchelor Institute.

Can be used in combination with textual summary of the Profile to create documentation with an automatically generated summary of what should be in conformant crates.



```
{
  "@id": "#prop_ldac:communicationMode_CreativeWork",
  "@type": "rdf:Property",
  "rdfs:label": "ldac:communicationMode",
  "name": "ldac:communicationMode",
  "prov:specializationOf": {
    "@id": "https://w3id.org/ldac/terms#communicationMode"
  },
  "domainIncludes": {
    "@id": "#class_CreativeWork"
  },
  "rdfs:comment": "The mode (spoken, written, signed etc.) of this resource. There may be r
  "rangeIncludes": {
    "@id": "ldac:CommunicationModeTerms"
  },
  "isPartOf": {
    "@id": "#hasSpecializedSchema"
  }
},
```

	A	B	C	D	
1	@id	@type	rdfs:label	domainIncludes	rdfs
2	#README.id	rdf:Property	@id	"#README_Entity"	REA
3	#prop_accountablePerson_Dataset	rdf:Property	accountablePerson	"#class_Dataset"	The
4	#prop_address_Place	rdf:Property	address	"#class_Place"	The
5	#prop_affiliation_Person	rdf:Property	affiliation	"#class_Person"	The
6	#prop_author_CreativeWork	rdf:Property	author	"#class_CreativeWork"	The
7	#prop_author_Dataset	rdf:Property	author	"#class_Dataset"	The
8	#prop_citation_Dataset	rdf:Property	citation	"#class_Dataset"	Asso
9	#prop_conformsTo_RepositoryCollection	rdf:Property	conformsTo	"#Root_Data_Entity"	A lir
10	#prop_conformsTo_RepositoryObject	rdf:Property	conformsTo	"#class_RepositoryObject"	A lir
11	#prop_contentLocation_RepositoryCollection	rdf:Property	contentLocation	"#class_RepositoryCollectic	The
12	#prop_contentSize_File	rdf:Property	contentSize	"#class_File"	File
13	#prop_creator_RepositoryObject	rdf:Property	creator	"#class_RepositoryObject"	The
14	#prop_creditText_Dataset	rdf:Property	creditText	"#class_Dataset"	A fr
15	#prop_dateCreated_RepositoryCollection	rdf:Property	dateCreated	"#class_RepositoryCollectic	The
16	#prop_dateCreated_RepositoryObject	rdf:Property	dateCreated	"#class_RepositoryObject"	The
17	#prop_datePublished_Dataset	rdf:Property	datePublished	"#Root_Data_Entity"	A da
18	#prop_dct:rightsHolder_Dataset	rdf:Property	dct:rightsHolder	"#class_Dataset"	The
19	#prop_description_Dataset	rdf:Property	description	"#Root_Data_Entity"	An a
20	#prop_description_RepositoryObject	rdf:Property	description	"#class_RepositoryObject"	A de

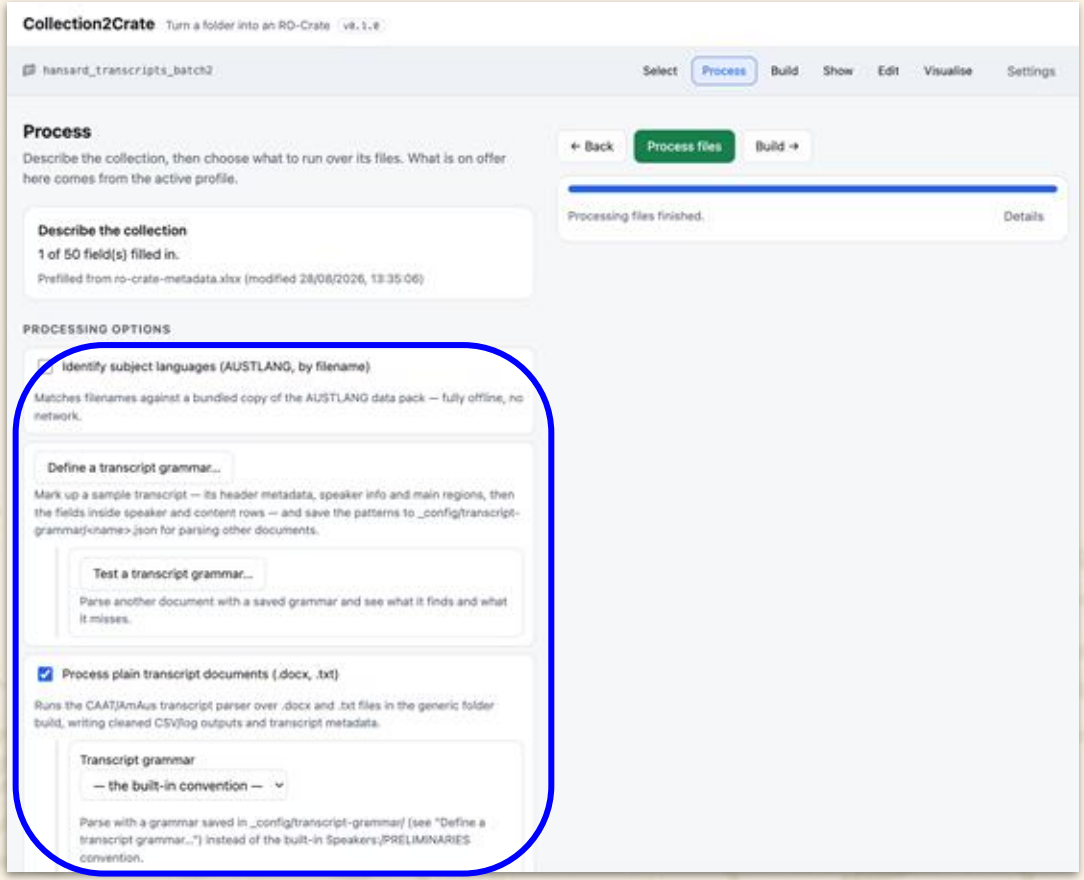
APPLIED MASP: MASP-DRIVEN EDITORS & TOOLS

In development:

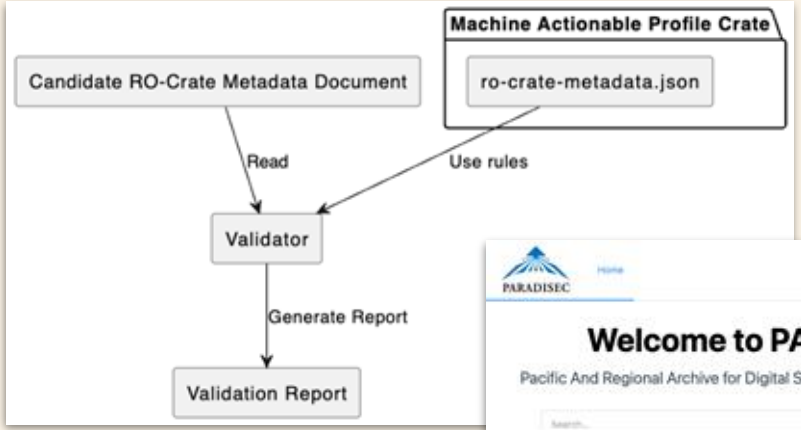
MASP profiles used to configure Collection2Crate (C2C).

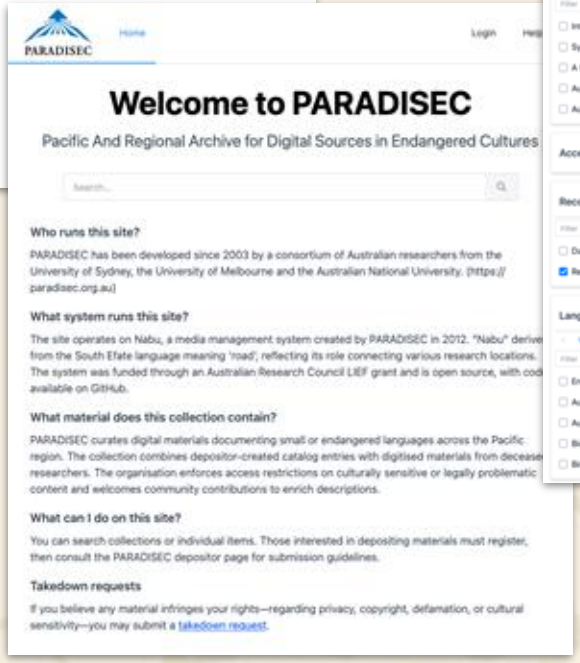
Plugins and options available according to each profile's tool-config.

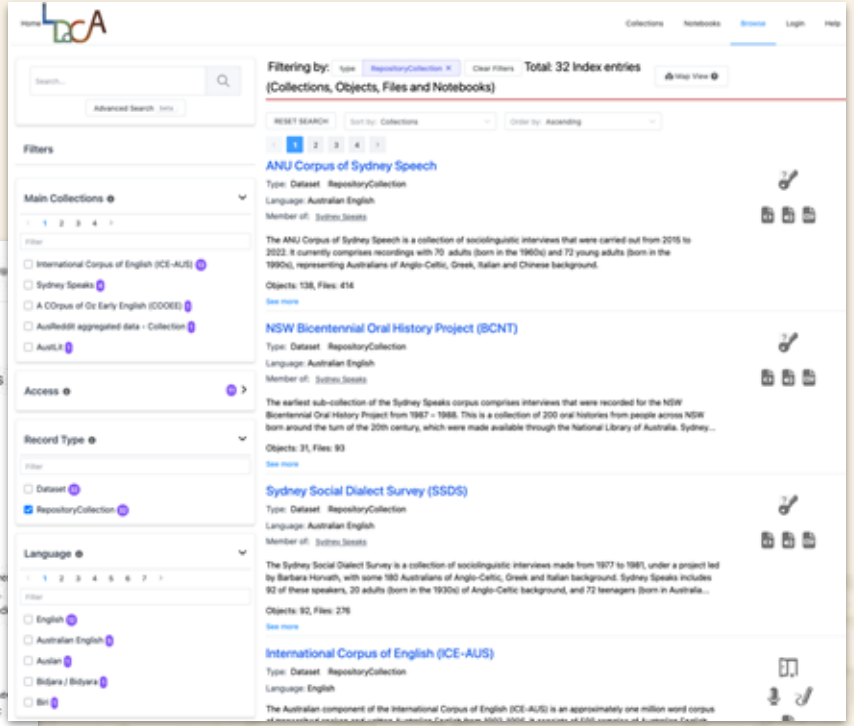
Allows users to select the processing options they want to work with for their collection that are available for the matching profile, e.g. LDAC profile provides AustLang lookup, conversation transcript processing, CHAT (.cha) output generation.



APPLIED MASP: VALIDATING COLLECTIONS







PARADISEC

data.ldaca.edu.au

catalog.paradisec.org.au

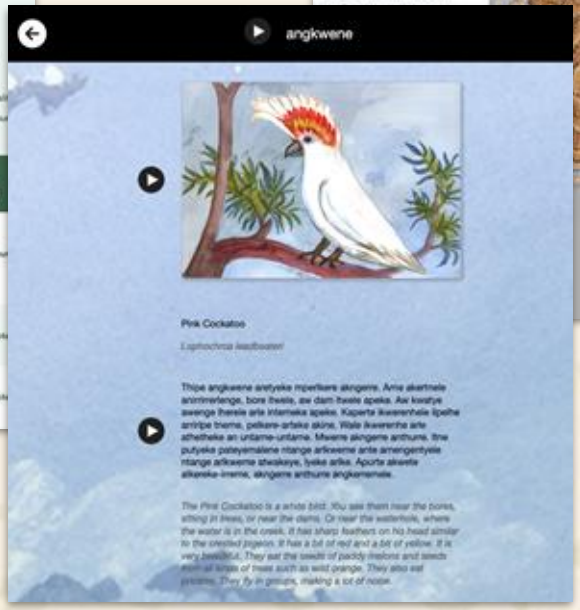
APPLIED MASP: PROFILE-SPECIFIC HTML TEMPLATES



precariousoralhistories.com/project-database



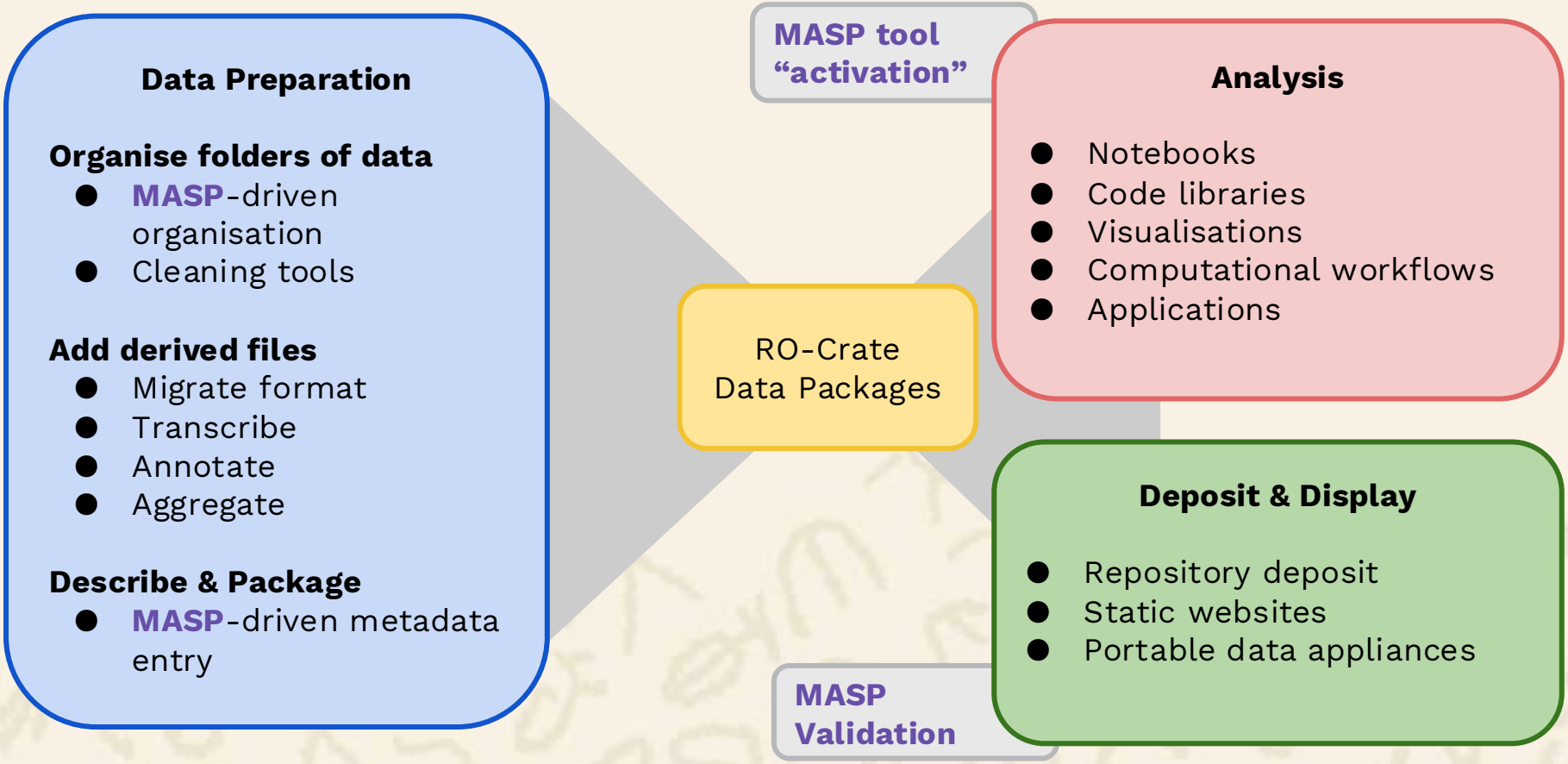
dingolingo.au



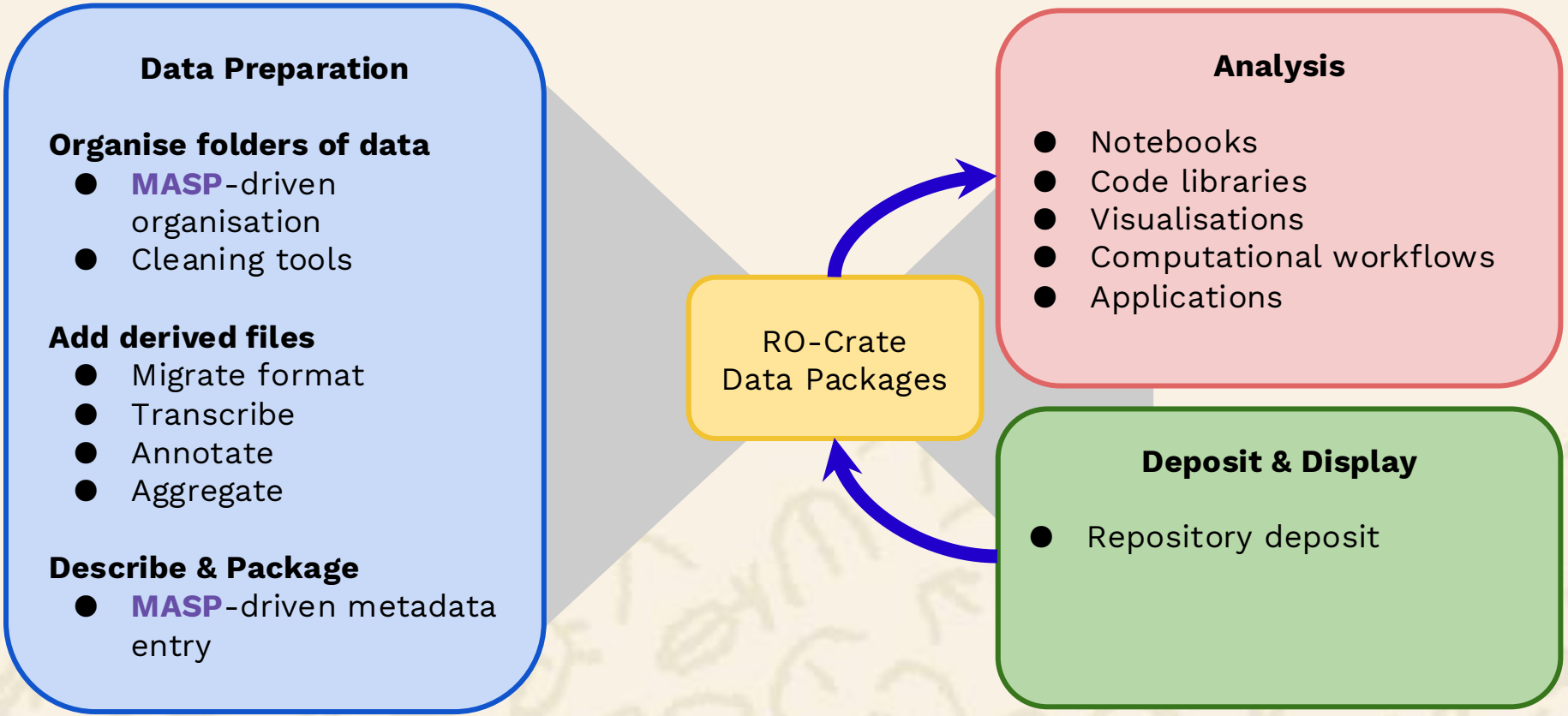
ayeye-thipe-akerte.bird-apps.com

Created with: github.com/crate-works/ro-crate-static-site

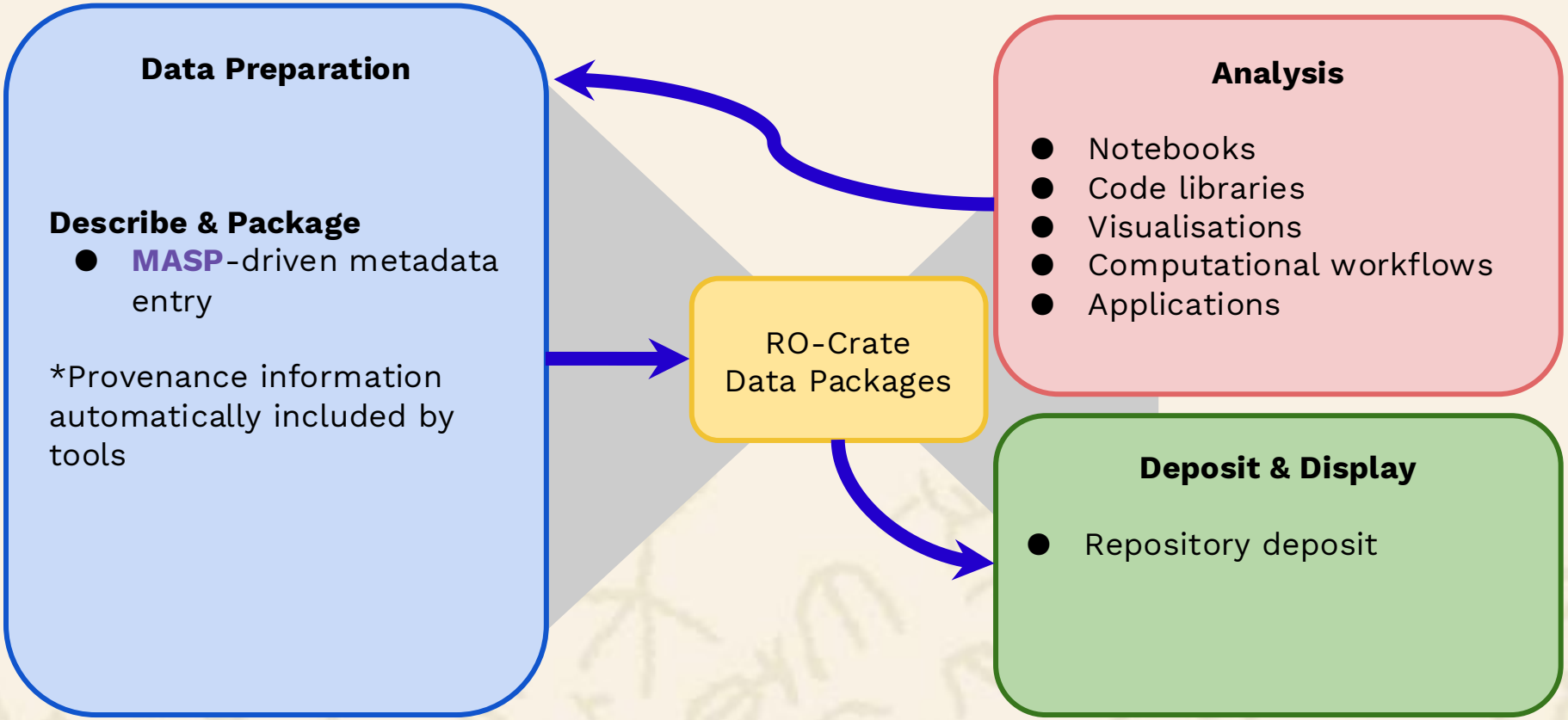
MASP: COMPLETE LINKED DATA ENVIRONMENT



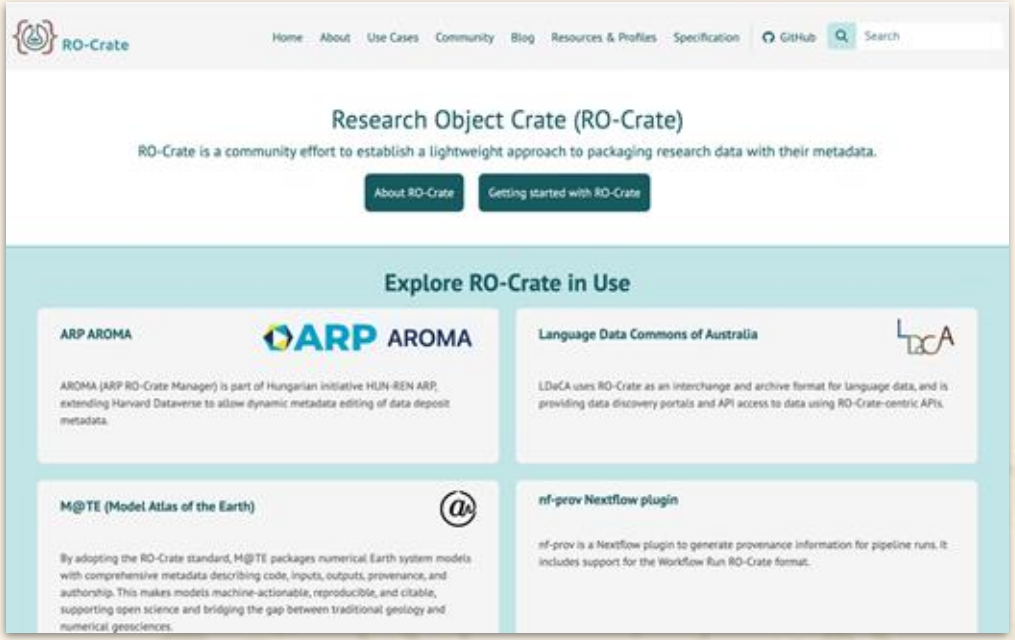
DATA RE-USE OF MASP-VALIDATED PACKAGES



MASP: RESEARCH OUTPUTS PUBLISHED/ARCHIVED



MASP NEXT STEPS



- Establishing MASP under RO-Crate Steering Committee governance
- More tools
  - Example crates to MASP
  - Legacy/existing linked data ontologies to MASP
  - MASP editor app

LDaCA

Language Data Commons of Australia

ldaca.edu.au

ldaca@uq.edu.au

@LDaCA_Project

@ldaca

The **Language Data Commons of Australia (LDaCA)** (DOI: 10.3565/kq2v-9g52) is a co-investment partnership with the **Australian Research Data Commons (ARDC)** through the HASS and Indigenous Research Data Commons. The ARDC is enabled by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).

Stay in the Loop

join our newsletter to receive quarterly updates