Describe, Manage and Discover Research Software

Sue Cook (CSIRO)
Jens Klump (CSIRO)
Paola Petrelli (CLEX)
Margie Smith (GA)
Geoff Squire (CSIRO)
Lesley Wyborn (NCI)
Mingfang Wu (ARDC)
Outline

- Introduction, landscape of software citation and publish, changes in Research Data Australia (RDA) for promoting software (Mingfang Wu, ARDC)
- New requirement from publishers and funders for software citation (Lesley Wyborn, NCI)
- Lightning talks:
  - Software Citation and GA: Motivations, outcomes and future direction (Margie Smith, GA)
  - Software in the CSIRO DAP: Description (Sue Cook, CSIRO)
  - CLEX software publishing workflow (Paola Petrelli, CLEX)
  - Describing software for Virtual Laboratories (Geoff Squire, CSIRO)
- Q/A, group discussion and feedback (Jens Klump, CSIRO)
Why do we care

- Software is pervasive in research
  - >90% of researchers acknowledge software is important for their own research
  - ~70% say their research would not be possible without it.
  - Of 40 papers examined in Nature Jan-March 2016, 32 contain 211 mentions of distinct pieces of software, for an average of 6.5 mentions per paper

12 scientific software challenges

**Open Research and Scholarly Communication**
- Intellectual property
- Publication and peer review
- Software dissemination, catalogues, search, and review

**Sustainable Software**
- Training and education
- Software engineering
- Portability
- Multidisciplinary science
- Reproducibility
- Reusability

**Sustainable community**
- Incentives, citation/credit models, and metrics
- Career paths
- Software communities and sociology
- Sustainability and funding models

Daniel S. Katz: *Software in Research: Underappreciated and underrewarded*. Keynote speech from 2017 eRA.
The FAIR Data Principles
Findable, Accessible, Interoperable, Reusable

The FORCE11 Software Citation Principles
Importance, Credit and attribution, Unique identification, Persistence, Accessibility, Specificity

The OSS Recommendations
Make source code publicly accessible from day one
Make software easy to discover by providing software metadata via a popular community registry
Adopt a license and comply with the license of third-party dependencies
Define clear and transparent contribution, governance and communication processes

**Who and What (Internationally)**

### Open Research and Scholarly Communication
- FORCE11 Software Citation Implementation WG
- RDA Research Software Source Code IG
- Nature software submission guidelines (2018)
- Journal of Open Source Software
- Elsevier – Why publish a software
  SoftwareX, Science of Computer Program, Neurocomputing

### Sustainable software
- US Research Software Sustainability Institute (URSSI)
- UK Software Sustainability Institute
- Working Towards Sustainable Software For Science (WSSSPE)

### Sustainable community
- Research Software Engineer Association

**Support from disciplines and organisations**
- ESIP: Software Guidelines
- AGU: Enabling FAIR Data Project
- Astrophysics: AAS Journals, Astrophysics Source Code Library
  - ...

Australian activities supporting research software

- Research Data Australia (catalogue)
- Australian Research Software IG
- RSE Association – Australian Chapter
- ARDC Skills and Training Program
- Uni., Gov. agencies, NCRIS facilities, etc. are treating software as research output

**Force11 software citation implementation group**
- RD-A software source code interest group
- DataCite

**AGU (Enabling FAIR Data Project)**
- ESIP Software and Services Citations cluster

**Working Towards Sustainable Software For Science (WSSSPE)**
- Research Software Engineer Association
Supporting catalogue/repository

General repository
- DataCite
- Zenodo
- Code Ocean
- Code.gov
- Figshare

Domain specific software repository
- Astrophysics Source Code Library (http://ascl.net/)
- OMICStools (https://omictools.com/)
- Bio.tools (https://bio.tools/)
- Bioconductor (https://www.bioconductor.org/)

Software code archive:
Software Heritage
Software metadata/ontology/vocabulary

CodeMeta Project

- DataCite
- Zenodo
- Dublin Core
- R Package Description
- Trove Software Map
- Perl Module Description (CPAN:Meta)
- Debian Package
- Python Disutils (pyPI)
- GitHub
- NodeJS
- Software Discover Index
- OntoSoft
- Software Ontology
- Figshare
- Research Data Australia (RIF-CS)
What do we (ARDC) do


- Recommended software citation format

  Creator (PublicationYear): Title. Version No. Publisher. [resourceTypeGeneral]. Identifier.
What do we (ARDC) do

- Amended Research Data Australia (RDA) registry schema (RIF-CS) for describing software as a distinct resource type
What do we (ARDC) do

- 212 registered software records from RDA (was 173 in Nov. 2017)
  - Commonwealth Scientific and Industrial Research Organisation 87
  - Geoscience Australia 70
  - Australian Ocean Data Network 34
  - Monash University 15
  - The University of Adelaide 4
  - ARC Centre of Excellence for Climate System Science 1
  - National Archives of Australia 1

108 of them have minted DOI
Enhance software discoverability in RDA

Clearly label software object
Enhance software discoverability in RDA

A new software link from this Contributor page
Enhance software discoverability in RDA

A new filter

A new “Explore” page for software
New requirements from Publishers (and the ARC/NHMRC/Universities Australia)

Lesley Wyborn
National Computational Infrastructure
The Drivers for Change

- Fifty years ago, most data that underpinned a publication could be represented in typeset tables and methods could be described in text.
  - Most calculations were done using slide rules and log tables

- With the advent of the digital age and the computerisation of instruments, volumes of data collected became too large to process manually and publish as tables: computer code became integral to modern scientific research.

- Increasingly data and software became included as a supplement in the research paper, accessible by contacting the journal, or else ‘by contacting the author’.
The Problem

- The inability to access primary data, samples, and software limits the ability to test the veracity and reproducibility of any publication.
- They do not guarantee accessibility and persistence of input research artefacts (data, software and samples in particular).

https://thelifeididntchoose.com/2018/08/14/life-is-absolutely-not-fair/
How do we fix it?

1. In 2017, a grant from the Laura and Arnold Foundation was awarded to the American Geophysical Union (AGU) and other partners to significantly improve the interconnection of data, software and samples in the literature in the Earth and environmental sciences, based around the FAIR guiding principles.

2. A coalition of Earth and environmental science publishers, disciplinary data repositories, and supporting organizations joined forces to work together to a commitment statement on FAIR publishing.

3. AuScope, ARDC and the NCI were all partners in the project and various members participated in stakeholder meetings and made contributions to final outcomes.
The Commitment Statement

This states that publication of scholarly articles in the Earth environmental science communities is conditional upon the concurrent availability of underpinning data and software. These should, to the greatest extent possible, be shared, open, and stored in community-approved FAIR-aligned repositories.

This has been signed by publishers, repositories, professional societies, institutions, research data infrastructures and individuals (including AuScope, NCI, ARDC)

What does this mean for the Earth and environmental sciences?

For The Publishers?
Publishers are now working towards following consistent policies for sharing and citing data, samples and software and will move from having these as supplements to using trusted repositories for publishing supporting research artefacts.

For Repositories?
Repositories will need to move towards be able to provide persistent identifiers, rich metadata, and related services for the data, software and samples they hold.

For Researchers?
Researchers will need to know how to consistently share, document, and reference data, samples and software and use globally persistent identifiers to uniquely identify their research outputs.
ARDC has developed guidelines for citing software based on international recommendations of FORCE 11 software citation principals, DataCite, CodeMeta, and others.

Software Citation and Geoscience Australia
Motivations, outcomes and future direction

Margie Smith
Data Policy and Informatics
Main motivations

1. Government Policy and Legislation
   a) Digital Continuity 2020
   b) The Archives Act 1983

2. Geoscience Australia Data Strategy

3. Geoscience Australia’s Science Principles

4. Geoscience Australia’s Strategic Priorities

5. …
Government policy considerations

Government Data / digital transformation agenda through the DC2020


and the Archives Act 1983 requiring provenance of method as described in the Records Disposal Authority.
Data management planning considerations

**Upstream**
- GA scientist
- Contractor
- Commercial operator
- Other gov’t organisations

**Input channels**
- Sensor data stream
- Web data stream
- Email attachment
- External disk drive
- Physical sample
- Physical copy

**Science Areas work spaces**
- Database
- National Computational Infrastructure
- Corporate Data Store
- Metadata catalogue
- Laboratory / software data
- Cloud services (e.g. AWS)

**Outputs**
- Web services
- Raw/processed data
- Metadata
- Scientific publications
- News and mailing lists
- Reports
- Workflows and code

**Downstream**
- Other gov’t organisations
- Geoscience Australia
- Scientific institutions
- Other geographic organisations

**Data management planning considerations**

**Science Areas work spaces**

**Outputs**
- Web services
- Raw/processed data
- Metadata
- Scientific publications
- News and mailing lists
- Reports
- Workflows and code

**Downstream**
- Other gov’t organisations
- Geoscience Australia
- Scientific institutions
- Other geographic organisations

**Data management planning considerations**
Embed best practice data management

Persistently identify all objects to enable provenance and cataloguing
Outcomes – tracking provenance through the standard
Current eCat search is not granular
Outcomes – tracking provenance through the standard

- IBTRrACS 70008 provides link to external p.o.t.
- TCRM GitHub https://github.com/GeoscienceAustralia/tcrm/releases/tag/v2.0.2
- TCRM 77484 Model eCat record
- Product Management Plan internal CMS link D2017-130697

Tropical Cyclone Risk Model Stochastic Event Catalogue 82033
Machine discoverable (?) but not people friendly
Moving towards correct citation for advice

Citation cases 1. (proposed workflow tool N. Car 2017-05-11)

• {AUTHORS} ({YEAR}) {TITLE}. {TOOL_TYPE} {REPO_BRANCH}[0,1]. {PUBLISHER}. {DOI | URI}. {ACCESSED_DATE}[0,1]

What we hope to have in eCat for provenance against advice generated:

Future direction – discovery and linkage improvement

Tropical Cyclone Risk Model Stochastic Event Catalogue

The TCRM Stochastic Event Catalogue contains artificially generated tropical cyclone tracks and wind fields representing 10000 years of tropical cyclone activity. The catalogue is stored by year, with a track file and wind field file. The wind field file contains the maximum wind speed from all events occurring in the corresponding track file (i.e., it represents annual maximum wind speeds).

About this resource

Scope Code: dataset

Categories

- Climatology, meteorology, atmosphere
- Data Package
- DC2020
- Published_internal

Australian and New Zealand Standard Research Classification (ANZSRC)

- Natural Hazards

Legal constraints: Creative Commons Attribution 4.0 International Licence

Author: Arthur, W.C.

Contact for the resource

- Custodian: CSEMID
- Owner: Commonwealth of Australia (Geoscience Australia)

Citation

If you wish to cite this record as you would a publication, please use the following format:

Future direction – discovery and linkage improvement

Slowly improving compliance

Status of GA's GitHub Repo's metadata

Totals

- Total: 195
- Passed: 36
- Failed: 159

Previously

- Total: 155
- Passed: 2
- Failed: 153

Repos failing tests

| DefinitelyTyped | passed | README does not contain a subsection titled 'Contacts'
|----------------|--------|------------------------------------------------------------------
| repo_must_contain_readme |        |                                                                  |
| readme_must_start_with_title |       |                                                                  |
| repo_must_have_license_file |      |                                                                  |
| readme_must_contain_license_section |     |                                                                  |
| readme_must_contain_contacts_section |   | README does not contain any GA email addresses for contact people |

| GeodePy | passed | README does not contain a subsection titled 'Contacts'
|--------|--------|------------------------------------------------------------------
| repo_must_contain_readme |        |                                                                  |
| readme_must_start_with_title |       |                                                                  |
| repo_must_have_license_file |      |                                                                  |
| readme_must_contain_license_section |     |                                                                  |
| readme_must_contain_contacts_section |   | README does not contain any GA email addresses for contact people |
Thank you

See THE BOSS ask Bob a question about the data.

See Bob squirm.

See Bob search and collect data.

Bob didn't write any metadata, you see!

Now, see Bob feeling queasy.

data@ga.gov.au
Software in the CSIRO DAP:
Description

Sue Cook | Data Librarian
17 October 2018
Search CSIRO collections

Search By Location

Featured Collections

Data from the ASKAP latitude 50 Fast Radio Burst (FRB) sample

The collection accompanies the paper "The dispersion-brightness relation for fast radio bursts from a watch-fiel survey" It contains 3 directories: full_data/ ASKAP CRAFT search mode data...

Detecting Social Roles in Twitter

Social roles are one particular demographic characteristic, which includes work, non-spatial, community and familial roles. We create a new annotated dataset for the task of detecting social...

Silver Nanoparticle Data Set

This is a set of silver nanoparticles final configurations, for use in data driven studies. These structures have been optimized (fully relaxed) using Density Functional Tight Binding...
Search Results

Found: 90 results Display: 10|25|50 results

Privacy Preserving Linkage Software
A set of software tools for privacy preserving entity linkage. * anonlink: A library for carrying out the low level hash comparisons required server side * entity-service: Our linkage server implemen... more
Confidential Computing - - Published 05 Oct 2018

AusFarm Decision Support Software
AusFarm modelling tool built using the Common Modelling Protocol. One CSIRO Rural Decision Support - Software development - Published 13 Sep 2018

PorosityPlus
The PorosityPlus code can be used to calculate the surface area, volume and pore size distribution (PSD) of particle networks. These particles can be multiscale ranging from atoms, to nanoparticles to... more
MMM Research & Applications - MMM Software - Published 04 Sep 2018

cuda-fixnum
https://data.csiro.au/collections/
GrainScan - Software for analysis of grain images

Alex Whan, Matt Böcker, Leanne Bischof

This collection is to accompany the publication of the paper "GrainScan: A low cost, fast method for grain size and colour measurements." It contains the software version that is referred to in that publication.

20181008090840 Test Data Collection
Suresh Palaniyandi
20181008090835 Test Data Collection
Suresh Palaniyandi

Software (140)
Data (1339)
Service (60)
GrainScan - Software for analysis of grain images

**About this Collection**

**Collection Title:** GrainScan - Software for analysis of grain images

**Collection Description:** This collection is to accompany the publication of the paper "GrainScan: A low cost, fast method for grain size and colour measurements". It contains the software version that is referred to in that publication.

**Field of Research:** Plant Biology not elsewhere classified

**DOI:** [http://doi.org/10.4225/08/S36302C43FC28](http://doi.org/10.4225/08/S36302C43FC28)

**Contact:** CSIRO Enquiries

**Keywords:** Grain; cereal; image analysis; seed size; software

**Related Materials:** Collection: Repository containing maintained versions of GrainScan.


**Collection:** Whan A., Cavanagh C: Scanned wheat grain images, 10.4225/08/S2P9AP7J62532.

**Supporting Files:** GrainScanSupplement.docx

**Licence:** CSIRO Binary Software Licence

**Organisations:** CSIRO (Australia)

**Attribution Statement:** Whan, Alex; Bolger, Matt; Bischof, Leanne (2014): GrainScan - Software for analysis of grain images. v2. CSIRO. Software Collection. [http://doi.org/10.4225/08/S36302C43FC28](http://doi.org/10.4225/08/S36302C43FC28)

**Rights Statement:** All Rights (including copyright) CSIRO Australia 2014.

**Access:** The metadata and files are available to the public.
Workspace: Scientific Workflow Platform

About this collection

- CSIRO, Matt Bolger, Paul Cleary, Lachlan Hetherington, Chris Rucinski, David Thomas, Damien Watkins

Collection description

Workspace is a powerful software platform designed to address two specific user scenarios: 1) Scientists who want to create and share scientific workflows in one coherent, simple environment where much of the "heavy lifting" has already been developed and proven over a number of years 2) Developers who want to make their software available as commercial products, plugins or components that can be freely mixed with capabilities from collaborators.

Access

The metadata and files (if any) are available to the public.

Related links

- Publication: Workspace: A Platform for Delivering Scientific Applications
- Publication: Workspace: scientific workflow platform
- Website: Workspace website at CSIRO

Supporting Files

workspace.license

About this project

- CSIRO, Cleary, Paul, Hetherington, Lachlan, Bolger, Matt, Rucinski, Chris, Sankaranarayanan, Nilupama; Thomas, David; Watkins, Damien; Zhang, Zhi; Subramaniam, Rajesh; Nguyen, Dang Oanh; McNally, Matt (2018). Workspace: Scientific Workflow Platform, v1.4. CSIRO. Software Collection. https://doi.org/10.25919/5b3c1df633cd3
Workspace: Scientific Workflow Platform

Software

Environment requirements
Windows/Linux/Mac

Language (programming)
C++

Operating system
Windows/Linux/Mac

Version
3.4.0
DAP common metadata: CSMD-CCLRC Core Scientific Metadata Model
More about this Collection

If you are entering software or data that requires specific research area fields, select extra descriptors to further search capabilities:

**Metadata Schema:**
- Software

**Environment Requirements:**
- ANZLIC
- Darwin Core
- Marine Community Profile

**Language (Programming):**
- VO Resource

**Operating System:**
- Software

**Version:**
- Sensor

**Software Documentation:** *Upload the Software Documentation using Supporting Attachments.*
More about this Collection

If you are entering software or data that requires specific research area fields, select extra descriptors to further search capabilities:

Metadata Schema: Software

*Environment
Requirements:

*Language
(Programming):

*Operating System:

*Version:

## Software Schema Usage

<table>
<thead>
<tr>
<th>Metadata access</th>
<th>Used Software Schema?</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Public</td>
<td>64</td>
<td>26</td>
</tr>
<tr>
<td>CSIRO Only</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Specific Users</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>68</strong></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>
Schema.org tags

```script
type='application/ld+json'>
{
  "@context": "http://schema.org",
  "@type": "Dataset",
  "name": "Workspace: Scientific Workflow Platform",
  "description": "Workspace is a powerful software platform designed to address two specific user scenarios: \n1) Scientists who want to create and share scientific workflows in an coherent, simple environment where much of the \n2) Developers who want to make their software available as commercial products, plugins or components that can be freely mixed with capabilities from collaborators\n",
  "datePublished": "2018",
  "keywords": "scientific workflow platform",
  "license": "https://wiki.csiro.au/display/dmsdoc/CSIRO+Binary+Software+License+Agreement",
  "citation": "CSIRO; Cleary, Paul; Hetherton, Lachlan; Bolger, Matt; Rucinski, Chris; Sanjaramanayanan, Nirupama; Thomas, David; Watkins, Damien; Zhang, Zikai; Subramanian, Rajesh; Nguyen, Dang Quan; McNally, Matt (2018): Workspace: Scientific Workflow Platform. v14. CSIRO. Software Collection. 10.25919/5b3c1dc633cd3",
  "publisher": "CSIRO",
  "isAccessibleForFree": true,
  "author": [{"@type": "Person", "name": "Paul Cleary"},{"@type": "Person", "name": "Lachlan Hetherton"},{"@type": "Person", "name": "Matt Bolger"},{"@type": "Person", "name": "Chris Rucinski"},{"@type": "Person", "name": "Nirupama Sanjaramanayanan"},{"@type": "Person", "name": "David Thomas"},{"@type": "Person", "name": "Damien Watkins"},{"@type": "Person", "name": "Zikai Zhang"},{"@type": "Person", "name": "Rajesh Subramanian"},{"@type": "Person", "name": "Dang Quan Nguyen"},{"@type": "Person", "name": "Matt McNally"}],
  "funder": [{"@type": "Organization", "name": "CSIRO" }],
  "identifier": "DOI: 10.25919/5b3c1dc633cd3",
  "URL": "https://doi.org/10.25919/5b3c1dc633cd3"
}
</script>
Next

Rest of 2018

• Tech debt and consolidation
  • Could address some of our gaps - eg collection types in DataCite and schema.org

2019/20

• New UI for depositor pages
• Greatly enhanced API for deposit

Vision

• Deposit API and CodeMeta could be mapped to pull in software from code repositories.
Make your research software citable

Did you know that you can make your research software citable?

More and more journals are adopting the FORCE 11 Software Citation principles and encouraging researchers to make the software that was used in their research available. The easiest way for people to find out about the availability of your research software is to cite it in your references.

The way to do this in CSIRO is to publish your software in the CSIRO Data Access Portal (DAP). This will create a snapshot of the version that you used and give you an attribution statement, including a DOI, which can be used as your citation. If you want to update your software later, you can simply update the existing record with the new release and the DAP will give you a new DOI and keep both versions preserved. See how other researchers have already created software records.

There is a Software Release Process that needs to be completed. This will help you to select the right licence for your software. Putting a record on the DAP will generate an approval process so you know your CSIRO compliance issues are taken care of.

See Using the DAP for software or contact researchdatasupport@csiro.au for more information.
Thank you

Thanks to:
Dom Hogan for the statistics and
Research Data Support team for feedback
The Centre of Excellence for Climate Extremes (CLEX) is a major initiative funded by the Australian Research Council. The Centre is an international research consortium of five Australian universities and a network of outstanding national and international partner organizations.
What we’re trying to achieve

Encourage our community to share their codes

Provide a source of relevant and reliable code for our community

Supply a place to publish software in case our researchers want or have to
Starting point

We are not an institution we rely on others for services (mostly)

We need to act quickly: less than 6 yrs left

We have a “data source” with RDA

We work at NCI and so publish data with their services

We manage a github organization: https://github.com/coecms

We manage a DMP web tool based on the UK DCC roadmap/dmponline
Which software

From github:

- Code produced by our team: manage data/model and analysis
- Code produced by student and researchers: for analysis, often used by the all research group, occasionally by a wider community

Lost somewhere:

- Model related combination of code and data, as: configurations, alterations of a model scheme, tutorials.
Proposed workflow

Communities created and curated by Zenodo users

- climate extremes

My communities
- CLEX: Australian Centre of excellence for Climate Extremes

Zenodo to publish and assign DOI

Publishing also on RDA using existing data source

CleX Roadmap data plan tool to create metadata and/or keep track of records.
Proposed workflow

Version control
Most code already on github or bitbucket.
Model configurations? Probably provide some form of template to help collecting them.

Add metadata
Collect information from repository and save as codemeta.json and zenodo.json files:
- harvest metadata from repository (python)
- template on CleX Roadmap
Upload json files to repository

Publish zenodo & RDA
Admin review records and publish to zenodo CleX community: python code using zenodo api.
Export record to RDA
If metadata harvested directly from repository, then harvest record from zenodo to CleX Roadmap.
Describing software for Virtual Labs

Geoffrey Squire

www.data61.csiro.au
Describing software for use

- Publishing software is easy
  - e.g. via GitHub, PyPI, download
  - Findable
  - Accessible
- But users want to Just Use It!

The goal is to make published software more useable by enabling automation.
Information Model

Applications run **Solutions** that use **Toolboxes** to solve Problems

- Machine-readable descriptions
- Metadata for searching, understanding, citation and provenance
- Sufficient to provision and run software automatically
Toolbox

Describes a software environment that can run published software

• Links to the published software
• How to instantiate the environment
• Dependencies
  - python
  - puppet
  - toolbox
• Implementation
  - cloud image
  - HPC
  - puppet module
  - execute instructions
Solution

*Describes a workflow that uses a Toolbox to solve a specific Problem*

- Link to published artifact (e.g. python script)
- Link to the Problem it solves
- Dependencies
  - Toolbox (usually)
- How to implement the Solution
  - Specification of inputs and outputs
    - Name
    - Description
    - Type
    - Constraints
Making it all useful

- A Solution Centre is a catalogue of Toolboxes and Solutions
- Developers (or others!) can publish descriptions of their software
- URIs for reference, citation and provenance
- Client apps can discover Solutions plus info to use them
- WIP on CodeMeta, RIF-CS and DOIs

https://sssc-vgl.geoanalytics.csiro.au
Making use of it all

• Virtual Laboratories (VLs) are clients
  - Data sets and services from registries
  - Solutions and toolboxes from the Solution Centre
• Users find relevant and **useable** data and solutions
• VL automates the job:
  - Generate UI to configure the Solution (parameters and inputs)
  - Assemble software environment
  - Wrangle input data
  - Execute and monitor the job
  - Store outputs and notify user when complete
  - Provide a provenance record

https://vgl.auscope.org
THANK YOU

Data61/Unit Name
Geoffrey Squire
Software Engineer

t  +61 2 6216 7064
e  geoffrey.squire@data61.csiro.au

www.data61.csiro.au
Q/A, Discussion
Thank you ...

Research software interest group: