Existing efforts and practices related to Software Source Code in academia

Software Source Code IG

Morane Gruenpeter

Inria, Software Heritage (France)

research data sharing without barriers
rd-alliance.org

10th November 2020 - RDA 16th Virtual Plenary Meeting

Housekeeping

- [Collaborative notes](https://tinyurl.com/y2kunpf5) https://tinyurl.com/y2kunpf5
- [These slides](https://tinyurl.com/yyargmeu) https://tinyurl.com/yyargmeu

- Meeting etiquette
  - Add your name to the participants list
  - Add your questions in the chat
  - Raise your hand if you wish to speak
  - **Please be aware that the session is being recorded and will be made publicly available**

Agenda

Introduction : Interest Group goals and past activities **Ice-breaker**: Why are you interested in software source code?	10
Software source Code IDentification (SCID) WG output: Use cases and identifiers schemes for software source code identification	15
FAIR4 Research Software WG: Introduction and Invite to join the discussion	5
FORCE11 Software Citation Implementation Working Group update ● including the ongoing task forces (CodeMeta, journals, repositories...)	10
Overview of other ongoing efforts related to software	10
Group activity: Collecting existing practices	30
Next steps for the SSC Interest Group	10

RDA
RESEARCH DATA ALLIANCE

The Software Source Code IG

Co-chairs:

- Neil Chue Hong
- Julia Collins
- Roberto Di Cosmo
- Mingfang Wu

VP16 coordinator:

- Morane Gruenpeter

https://www.rd-alliance.org/groups/software-source-code-ig

Objectives:

A forum for discussing **research software** inside RDA

- issues on management, sharing, discovery, archival and provenance of software source code.
- It will pay special attention to source code that generates research data and plays an important role in scientific publications.

Chronology

BOF RDA P9, Barcelona April 2017
motivations => 60 participants

RDA P10, Montreal September 2017
motivations, survey of ontologies,
metadata use cases

RDA P11, Berlin March 2018   started
the idea for a dedicated identification
WG

RDA P13, Philadelphia April 2019
FAIR for Software Source Code and
launch of the **SCID WG**

FORCE2019, Edinburgh October 2019
full day hackathon on **Research
Software**

RDA VP15, Australia March 2020
Open discussion about the creation of a
new group, the **FAIR4RS WG** (which
was launched in June 2020)

RDA VP16, Costa Rica November 2020
Existing efforts and practices in
Academia

Software at RDA and in academia

- Related groups:
  - RDA, ReSA and FORCE11 FAIR for Research Software Working Group (FAIR4RS WG)
    - Welcome to join the work defining FAIR principles for research software
  - RDA & FORCE11 Software Source Code Identification WG(SCID IG)
    - Output published in September 2020
  - FORCE11 Software Citation Implementation Working Group (SCIWG)
    - Ongoing WG about software citation
- Related software sessions during VP16:
  - Software Source Code IG meeting[1]
  - FAIR software roadmap BOF meeting[2]
  - FAIR 4 Research Software (FAIR4RS)[3]
  - Computational Notebooks BOF meeting[4]
  - CURE-FAIR WG meeting[5]
 - Research Data Management in Engineering: Data Provenance and Research Software in Engineering IG meeting[6]

Research Software in Open Science

Why software source code?


```
52
53   # THE MASTER IGNITION ROUTINE IS DESIGNED FOR USE BY THE FOLLOWING LEM PROGRAMS:  P12, P40, P42, P61, P63.
54   # IT PERFORMS ALL FUNCTIONS IMMEDIATELY ASSOCIATED WITH APS OR DPS IGNITION:  IN PARTICULAR, EVERYTHING LYING
55   # BETWEEN THE PRE-IGNITION TIME CHECK -- ARE WE WITHIN 45 SECONDS OF TIG? -- AND TIG + 26 SECONDS, WHEN DPS
56   # PROGRAMS THROTTLE UP.
57   #
58   # VARIATIONS AMONG PROGRAMS ARE ACCOMODATED BY MEANS OF TABLES CONTAINING CONSTANTS (FOR AVEGEXIT, FOR
59   # WAITLIST, FOR PINBALL) AND TCF INSTRUCTIONS.  USERS PLACE THE ADRES OF THE HEAD OF THE APPROPRIATE TABLE
60   # (OF P61TABLE FOR P61LM, FOR EXAMPLE) IN ERASABLE REGISTER `WHICH' (E4).  THE IGNITION ROUTINE THEN INDEXES BY
61   # WHICH TO OBTAIN OR EXECUTE THE PROPER TABLE ENTRY.  THE IGNITION ROUTINE IS INITIATED BY A TCF BURNBABY,
62   # THROUGH BANKJUMP IF NECESSARY.  THERE IS NO RETURN.
63   #
64   # THE MASTER IGNITION ROUTINE WAS CONCEIVED AND EXECUTED, AND (NOTA BENE) IS MAINTAINED BY ADLER AND EYLES.
65   #
66   #                HONI SOIT QUI MAL Y PENSE
67   #
68   #       ********************************************
69   #            TABLES FOR THE IGNITION ROUTINE
70   #       ********************************************
71   #
72   #                NOLI SE TANGERE
73
74   P12TABLE     VN     0674       # (0)
75                TCF    ULLGNOT    # (1)
76                TCF    COMFAIL3   # (2)
77                TCF    GOCUTOFF   # (3)
78                TCF    TASKOVER   # (4)
79                TCF    P12SPOT    # (5)
80                DEC    0          # (6)   NO ULLAGE
81                EBANK= WHICH
82                2CADR  SERVEXIT   # (7)
83
84                TCF    DISPCHNG   # (11)
85                TCF    WAITABIT   # (12)
86                TCF    P12IGN     # (13)
87
88   P40TABLE     VN     0640       # (0)
89                TCF    ULLGNOT    # (1)
90                TCF    COMFAIL4   # (2)
91                TCF    GOPOST     # (3)
92                TCF    TASKOVER   # (4)
93                TCF    P40SPOT    # (5)
```

[Go to the code!](#)

"Source code provides a view into the mind of the designer."
Len Shustek, Computer History Museum
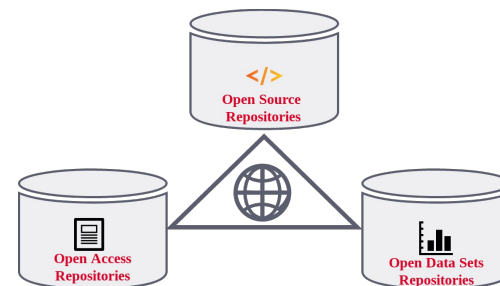
Three pillars of Open Science, Software Heritage CC-By 4.0 2019

Ice-breaker question

Why are you interested in Software Source Code?

https://tinyurl.com/y2kunpf5

research data sharing without barriers
rd-alliance.org

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

Joint RDA & FORCE11 effort

Software Source Code Identification Working Group
Output: Use cases and identifier schemes for persistent software source code identification

research data sharing without barriers
rd-alliance.org

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

9

Software Source Code Identification Working Group

Co-chairs
- Roberto Di Cosmo
- Martin Fenner
- Daniel S. Katz

Secretariat Liaison
- Stefanie Kethers

RDA page

Repository

Output **DOI:10.15497/RDA00053**

Output Card

Authors of the SCID WG output (alphabetical order by name)

- Alice Allen, Astronomy Source Code Library & U. Maryland, USA
- Anita Bandrowski - University of California San Diego, USA
- Peter Chan - Stanford University Libraries, California, USA
- Roberto Di Cosmo - Software Heritage, Inria and University of Paris, France
- Martin Fenner - DataCite, Germany
- Leyla Garcia - ZB MED Information Centre for Life Sciences
- Morane Gruenpeter - Inria, Software Heritage, France
- Catherine M Jones - UKRI STFC, UK
- Daniel S. Katz - University of Illinois at Urbana-Champaign, USA
- John Kunze - California Digital Library, University of California, USA
- Moritz Schubotz - swMATH, FIZ Karlsruhe, Germany
- Ilian T. Todorov - UKRI STFC Daresbury Laboratory, UK
- And the participants of the SCID WG (listed in Appendix B)

Editor: Morane Gruenpeter - Inria, Software Heritage, France

research data sharing without barriers
rd-alliance.org

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

The SCID WG Goal

1. **capture and analyze** the software identification state-of-the-art in the scholarly ecosystem
2. Output structure
   - Definitions
     - Actors in the scholarly ecosystem
     - What do we want to identify or the granularity of software?
     - What is at stake
   - Use cases
     - Classified into one of the following actions: archiving, referencing,describing, citing
   - Identifiers schemas
     - Intrinsic identifiers
     - Extrinsic identifiers
   - Summary of findings

Identification target - what do we want to identify?

Software concept / project / collection
 Description in registry, a homepage or any other form of metadata record
- Project versions (for example Python2 and Python3)
- Modules
- Sub-modules

Software artifact
- Executable (download link)
- Software source code
    - Dynamic artifact - current development code (on collaborative development platform)
    - Archived copy
        - Snapshot (all branches, all dev history)
        - Release / Package
        - Commit- a specific point in development history
        - Directory
        - File
        - Algorithm

Software context
- Complementary artifacts - Software artifacts that are external to the source code
    - the software environment, tutorial (Jupyter notebook), Data (input/output data), etc.
- Articles
- Documentation

Pyramid levels (top to bottom):
Project — GL1
Project versions — GL2
Modules — GL3
Sub-Modules — GL4
Snapshots — GL5
Releases — GL6
Commits — GL7
Directories — GL8
Files — GL9
Code fragments — GL10

GL= Granularity Level

research data sharing without barriers
rd-alliance.org

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE
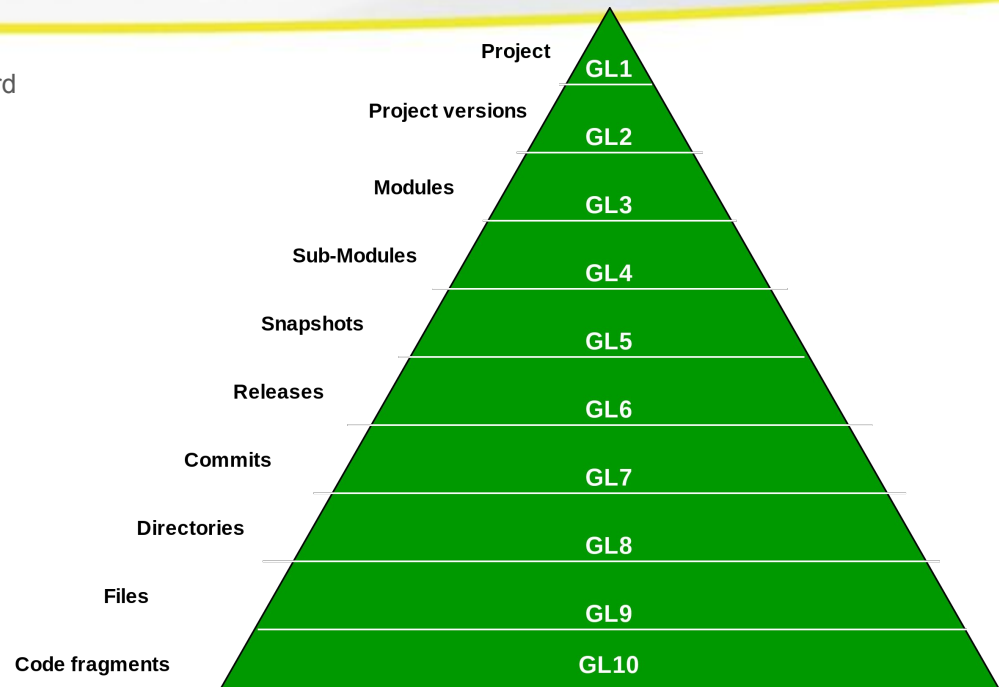
The use cases collection (a small excerpt)

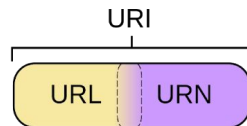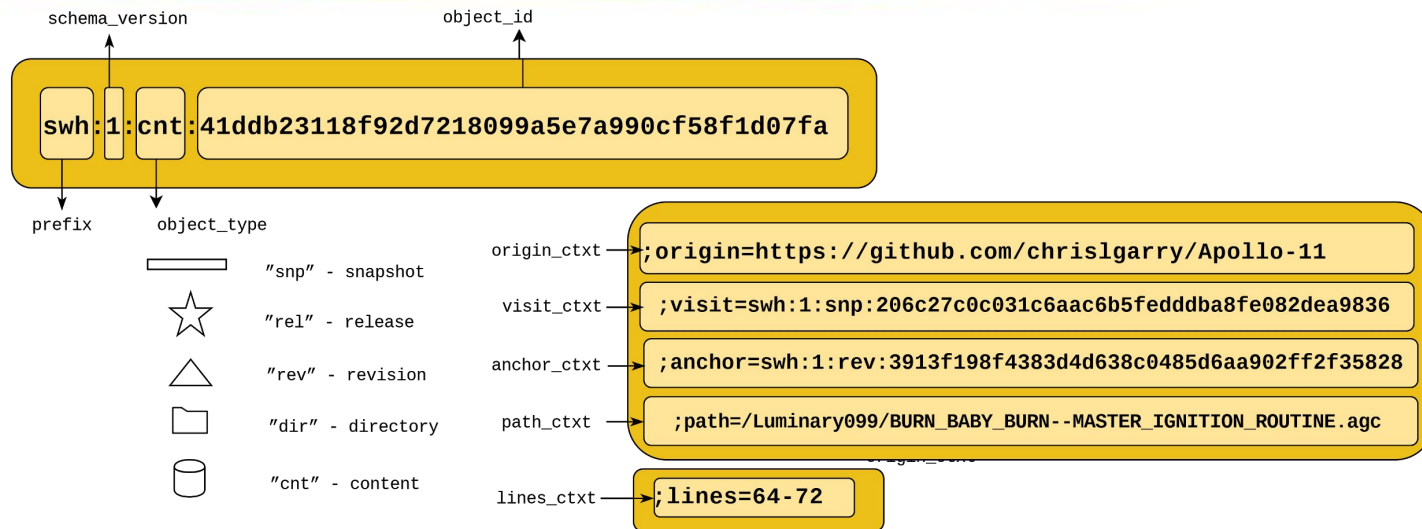| Actor	Use case description	Action	Identification target
Archive	Identify all the software artifacts I hold	Archiving, referencing	Release and smaller artifacts
Citation manager	Curate the software citation entries	Credit	Project, release
Curator / librarian / digital archivist	Catalog and browse the development history of legacy software source code for preservation purposes (The Apollo mission source code is a good scenario on how making code available on GitHub isn't enough for persistence purposes)	Archiving	Project, release and smaller artifacts depending on the reference
Publisher	Create/retrieve identifiers quickly for use in the paper for all software including commercial packages.	Referencing, describing	Any item (all granularity levels)
Registry	Identify and curate the software entries I hold	Archiving, referencing, describing, credit	Project
Researcher as a software user (RSU)	Access and use SSC no longer available on a collaborative platform	Archiving	Snapshot, release, revision, directory

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

Identifiers schemes

HAL-ID

Digital Object Identifier

URI

Wiki Item identifier (Qxxx)

ASCL.net
Astrophysics Source Code Library

Software Identification

ARK
Archival Resource Key

RRID

Handle
Handle System identifiers

git

swMATH
an information service for mathematical software

SWHID
Software Heritage identifiers

research data sharing without barriers
rd-alliance.org

FORCE11
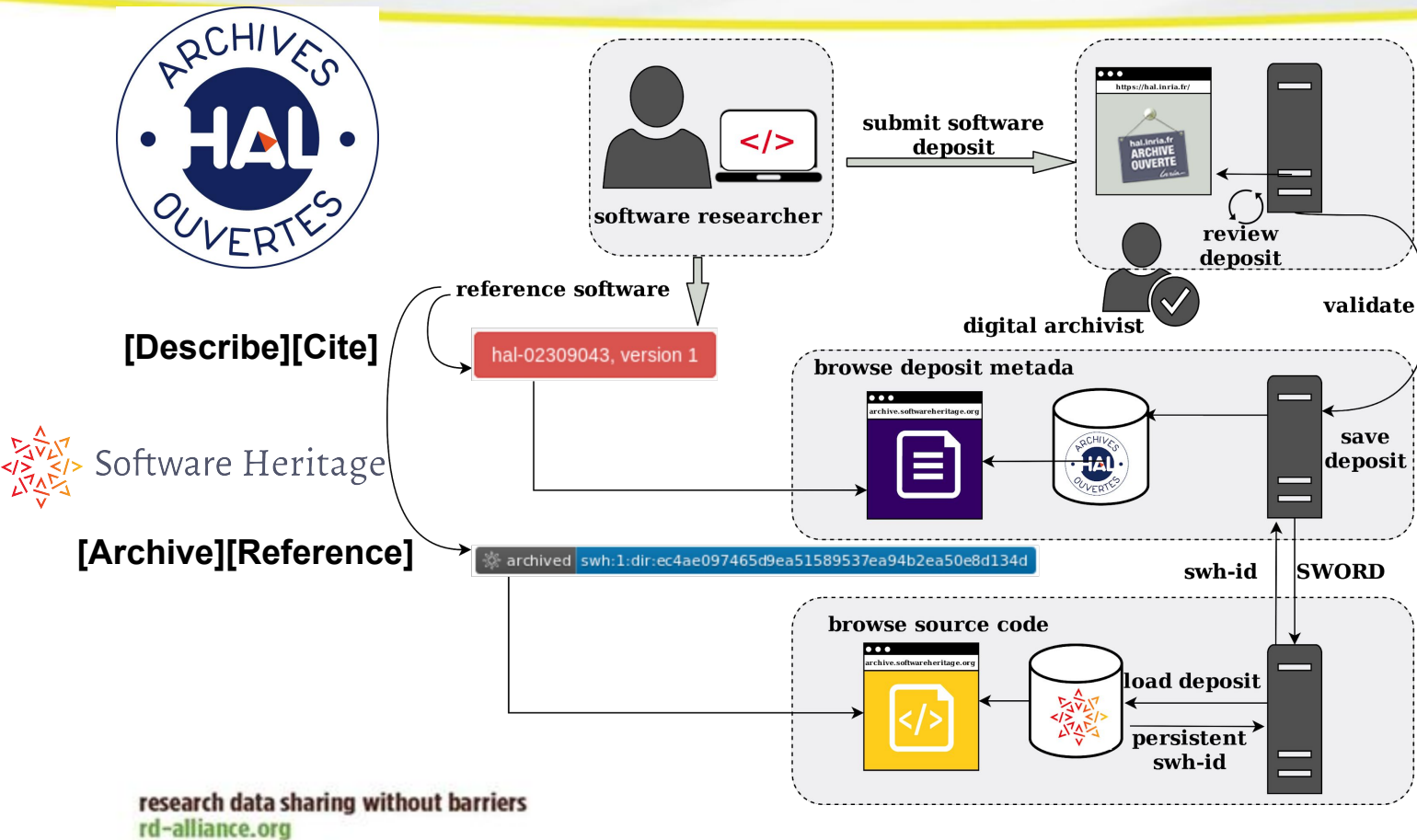The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

Intrinsic identifier: the Software Heritage ID (SWHID)

- **Intrinsic**: compute a unique **digital fingerprint**

- **decentralised**: do not need a registry, only agreement on a standard

- **cryptographically strong** identifiers


```
schema_version                      object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix      object_type
```

object_type:
- ▭ "snp" - snapshot
- ☆ "rel" - release
- △ "rev" - revision
- 🗁 "dir" - directory
- 🗄 "cnt" - content

- origin_ctxt → ;origin=https://github.com/chrislgarry/Apollo-11
- visit_ctxt   → ;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
- anchor_ctxt → ;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
- path_ctxt   → ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
- lines_ctxt   → ;lines=64-72

Extrinsic identifier: the **HAL ID**

Summary

Granularity level (GL)	ID target	Extrinsic identifiers									Intrinsic identifiers	
		ASCL	ARK	DOI	HAL	URL	RRID	SwMath	Wikidata		Hash	SWHID
									entity	property		
GL1	project	X	X	X	X	X	X	X	X			
GL2	project version		X						X			
GL3	module		X						X			
GL4	repository		X			X			X			
GL5	repository snapshot		X						X			X
GL6	release		X	X					X		X	X
GL7	commit		X						X		X	X
GL8	directory		X	X	X*				X		X	X
GL9	file		X	X							X	X
GL10	Code fragment		X									X

SCID WG next steps

- The working group has now **completed** its work in its current form
- Maintenance of the SCID output transfers to the **SSC IG**

"The next step would be to produce a set of recommendations based on these findings."

RDA, ReSA & FORCE11

FAIR for research software WG (FAIR4RS WG)

The FAIR4RS WG

Co-chairs:

- Paula A. Martinez
- Michelle Barker
- Daniel S. Katz
- Leyla Garcia
- Neil Chue Hong

Steering committee members:

- All co-chairs
- Fotis Psomopoulos
- Morane Gruenpeter
- Jen Harrow
- Carlos Martinez

Objectives:

- Coordinating of a range of existing community-led discussions on:
    - How to define and effectively apply FAIR principles to research software,
    - How to achieve adoption of these principles.

Deliverables:

- FAIR principles for research software (6 months)
- Guidelines on how to apply the FAIR (12 months)
- Implementation guidelines and adoption examples (18 months)

<ReSA>
Research Software Alliance

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

FORCE11

Software Citation Implementation WG (SCI WG)

FORCE11
The Future of Research Communications and e-Scholarship

FORCE11 SCI WG

FORCE11 Software Citation Implementation Working Group
(co-chairs: N. Chue Hong, M. Fenner, D. S. Katz)

Following-on from FORCE11 Software Citation Working Group and the Software Citation Principles it developed

Objective: Produce concrete guidelines for software citation, and implement them within the scholarly research community (software developers, repositories and registries, journals and conference and publishers, indexers, institutions)

A community with monthly calls to discuss challenges and progress in implementing software citation, with task forces for

- **Guidance** - developing documents for developers, authors, and reviewer
- **Journals** - coordinating editors and publishers to simplify and implement guidance
- **Repositories** - developing best practices document for handling software
- **CodeMeta** - standardizing metadata for software, moving towards merging into schema.org

research data sharing without barriers
rd-alliance.org

FORCE11

The Future of Research Communications and e-Scholarship

FORCE11 SCI WG - Guidance Task Force

GitHub repository

(Lead: N. Chue Hong)

Objective:
Develop guidance for different stakeholders to help implement software citation, principally authors of research articles seeking to cite software correctly and developers of software looking to make their software easier to cite

Milestones and activity:

- Developed and published Software Citation Checklist for Authors and Software Citation Checklist for Developers
- Developed Software Citation Primer which was used as basis for paper by the Journals Task Force
- Guidance Task Force on hiatus while the Journals Task Force provides feedback on guidance

FORCE11

The Future of Research Communications and e-Scholarship

FORCE11 SCI WG - Journals Task Force

(Lead: Daniel S. Katz)

Objective:
Work with organizations that publish journals, proceedings, monographs to improve how software is cited in their works and the scholarly processing ecosystem

Milestones and activity:

- Published paper (in peer-review):
- Working on comms plan
- Communities and institutions will produce versions of the document with software examples and citation styles that are appropriate for their intended audience
- Next steps:
  - Work on what happens after article is submitted – how citations are processed and indexed – to ensure they are correctly registered and tracked
  - You: tell your communities about this, encourage publishers to support it, encourage authors & reviewers to follow it

Daniel S. Katz [1], Neil P. Chue Hong [2], Tim Clark[3], August Muench [4], Shelley Stall [5], Daina Bouquin[6], Matthew Cannon [7], Scott Edmunds[8], Telli Faez[9], Patricia Feeney[10], Martin Fenner[11], Michael Friedman [12], Gerry Grenier [13], Melissa Harrison [14], Joerg Heber[15], Adam Leary[16], Catriona MacCallum [17], Hollydawn Murray[18], Erika Pastrana[19], Katherine Perry [20], Douglas Schuster[21], Martina Stockhause [22], Jake Yeston[23]

[1] University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2] EPCC, University of Edinburgh, Edinburgh, UK
[3] University of Virginia, Charlottesville, VA, USA
[4] American Astronomical Society, Washington, DC, USA
[5] American Geophysical Union, Washington, DC, USA
[6] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
[7] Taylor & Francis Group, Oxford, UK
[8] GigaScience Press, BGI Hong Kong, Hong Kong, Hong Kong
[9] Elsevier, Amsterdam, The Netherlands
[10] Crossref, Lynnfield, MA, USA
[11] DataCite, Hannover, Germany
[12] American Meteorological Society, Boston, MA, USA
[13] Publishing Technology, IEEE, Piscataway, NJ, USA
[14] Production, eLife, Cambridge, UK
[15] PLOS, San Francisco, CA, USA
[16] Oxford University Press, Oxford, UK
[17] Open Science, Hindawi, London, UK
[18] F1000Research, London, UK
[19] Springer Nature, New York, NY, USA
[20] Product Management, Wiley, Boston, MA, USA
[21] National Center for Atmospheric Research, Boulder, CO, USA
[22] German Climate Computing Center (DKRZ), Hamburg, Germany
[23] AAAS, Washington, DC, USA

◎FORCE11
The Future of Research Communications and e-Scholarship

24

FORCE11 SCI WG - Repositories Task Force

(Lead: Alice Allen)

Objective:

Bring together representatives of Research Software Registries and Repositories to discuss and improve practices

Milestones and activity:

- Drafted Quick Start Guides for Research Software Registries and Repositories Best Practices document
    - Aiming to release by end of 2020
- Working on guidance document for users/community best practice
- Working on code that takes information from a Python setup.cfg file and generates a codemeta.json file
- Held workshop for discipline-specific software registries and repositories in Nov 2019
    - Produced drafts of documents such as "A good-enough workflow for software citation"
- Task force was somewhat stalled during COVID-19, but is restarting

research data sharing without barriers
rd-alliance.org

FORCE11
The Future of Research Communications and e-Scholarship

FORCE11 SCI WG - CodeMeta Task Force

link

(Leads: M. Fenner and M. Gruenpeter)

Objective: express all codemeta properties
using schema.org

- ○ Step 1: internal TF discussion
- ○ Step 2: open issues in the CodeMeta repository
- ○ Step 3: validate & integrate proposals in the next release (v3)
- ○ Step 4: prepare formal proposal to schema.org

research data sharing without barriers
rd-alliance.org

CodeMeta initiative

- A subset of schema.org
- An academic community discussing software metadata
- A crosswalk table - mapping the metadata landscape
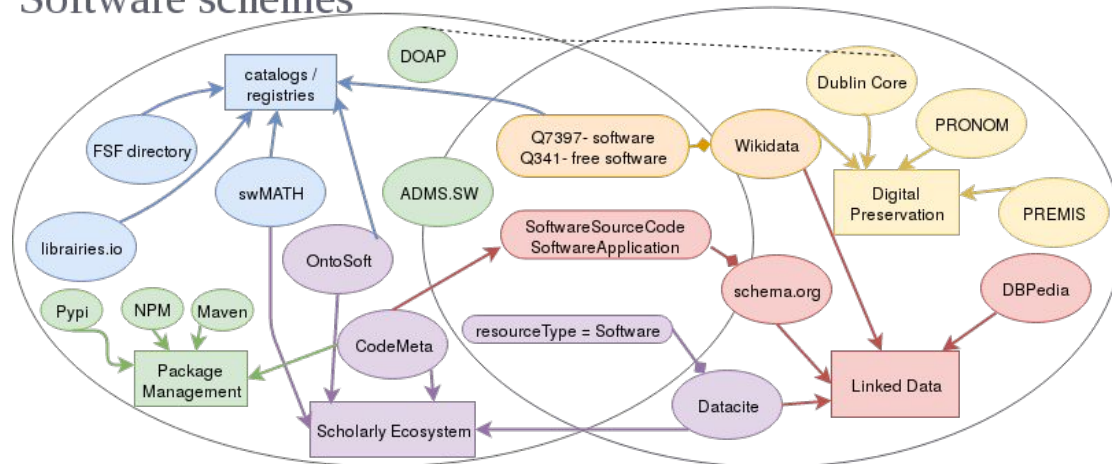
Gruenpeter M. and Thornton K. (2018) Pathways for Discovery of Free Software (slide deck from LibrePlanet 2018).
https://en.wikipedia.org/wiki/File:Pathways-discovery-free.pdf accessed on 6.11.2020.

The Future of Research Communications and e-Scholarship

CodeMeta generator

- An open source tool to create codemeta.json files
  - Use it directly on the CodeMeta hosted version
  - Contributions are welcome on the code repository

Contributed to the community by

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

FAIRsFAIR report

Link to report on Zenodo

Link to open consultation document

There will be a free access webinar on Monday November 23rd at 3PM CEST - register here

October 16, 2020

Project milestone   Open Access

M2.15 Assessment report on 'FAIRness of software'

Gruenpeter, Morane; Di Cosmo, Roberto; Koers, Hylke; Herterich, Patricia; Hooft, Rob; Parland-von Essen, Jessica; Tana, Jonas; Aalto, Tero; Jones, Sarah

Software has an important place in academia and as such it has an important place in the FAIR ecosystem. Software can be used throughout the research process; however it can also be an outcome of the research process. Distinguishing between these different roles is essential for any assessment of the `FAIRness of software'.

This is the first milestone of the FAIRsFAIR project focused specifically on software as a digital object. In this report we discuss the state-of-the-art of software in the scholarly ecosystem in general and in the FAIR literature in particular. We identify the challenges of different stakeholders when it comes to finding and reusing software. Furthermore, we present an analysis of nine resources that call for the recognition of software in academia and that present guidelines or recommendations to improve its status - either by becoming more FAIR or by improving the curation of software in general. With this analysis we demonstrate to what extent each of the FAIR principles is seen as relevant, achievable and measurable; and in what sense it benefits software artifacts. Finally, we present 10 high-level recommendations for organizations that seek to define FAIR principles or other requirements for research software in the scholarly domain.

Feedback and suggestions will be most welcome as comments on the public Google Doc version of this report
https://docs.google.com/document/d/1yvdLSP6oH3XozVy4CJtThzGNHkseCBdvmxfruDYLB6Q/edit?usp=sharing

Preview

Page : 1 of 63   — + 50%

![FAIRSFAIR]

Fostering Fair Data Practices in Europe

Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

M2.15 ASSESSMENT REPORT ON 'FAIRness of software'

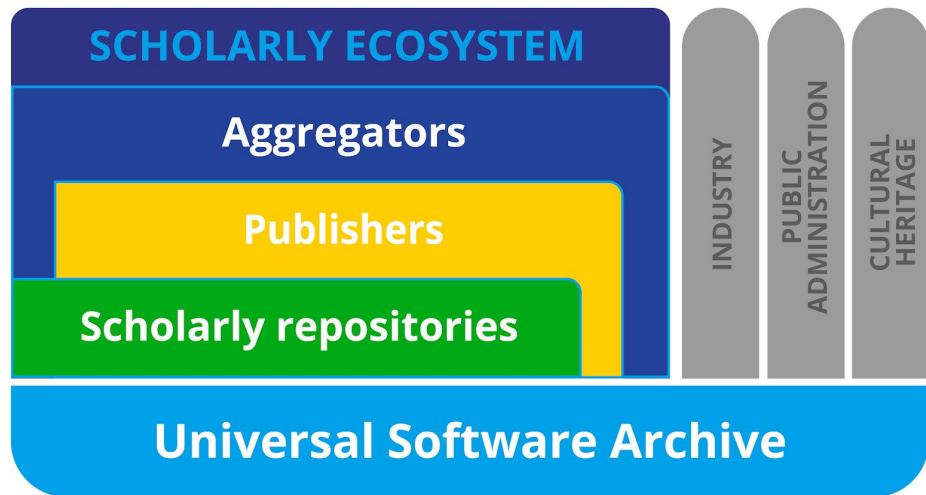| Work Package | WP2 -FAIR practices: semantics, interoperability and services |
| Lead Author (Org) | Morane Gruenpeter (INRIA) |
| Contributing Author(s) (Org) | Roberto Di Cosmo (INRIA), Hylke Koers (SURF), Patricia Herterich |

EOSC Scholarly Infrastructures for Research Software

- **Chairs**
  - Roberto Di Cosmo, Software Heritage
  - José Benito Gonzalez Lopez, Zenodo
- **Participants**
  - Representatives from 9 infrastructures:
  - Archives
    - HAL
    - Software Heritage
    - Zenodo
  - Publishers
    - Dagstuhl
    - eLife
    - IPOL
  - Aggregators
    - OpenAIRE
    - scanR
    - swMATH

SCHOLARLY ECOSYSTEM

Aggregators

Publishers

Scholarly repositories

Universal Software Archive

INDUSTRY

PUBLIC ADMINISTRATION

CULTURAL HERITAGE

Link to document
(community consultation ended on the 10.11.2020)

research data sharing without barriers
rd-alliance.org

EOSCArchitecture
Executive Board Working Group

RDA
RESEARCH DATA ALLIANCE

EOSC Scholarly Infrastructures for Research Software

- **Four Pillars**
  - **Archive, Reference, Describe, Credit**
- **State of the Art**
  - Best Practices & Open Problems
  - Cross Cutting Concerns

- **The Road ahead**
  - Requirements & Criteria
  - 13 Workflows / Use Cases examples
- **Recommendations**
  - Standards & Tools
  - Policy recommendations
  - Long term perspectives

Archive	Reference	Describe	Cite
Long term preservation	Link and access	Search and find	Publish

Related Projects - ReSA

Research Software Alliance (ReSA)

Vision: Research software: recognised and valued as a fundamental and vital component of research worldwide

Mission: To bring research software communities together to collaborate on the advancement of research software.

Task Forces:

- Software landscape analysis
- Evidence for the importance of research software
- Register of research software funding opportunities

Let's start

Full room discussion or in groups depending on how many people. 25' and 10' wrap up

- Introduce yourself to your neighbours (name, affiliation, why do source code interest you?)
- Software practices collection:
  - Do you or your organization create software? Use software?
  - Do you or your organization follow institutional or community best practices with the source code you create? (an old (2020) example is the Software Release Practice by E.S Raymond)

Next steps for the SSC IG

Open questions (to answer on the notes):

- What subjects would you like to **discuss** during the **next plenaries**?

- What types of **materials** would be helpful to have on the SSC IG **wiki page**?

- Would you like the **mailing list updates** to be more frequent and if so, what are the topics you would like to see on the mailing list?

Join us on the mailing list:

https://www.rd-alliance.org/groups/software-source-code-ig

research data sharing without barriers
rd-alliance.org

Thanks for joining