Existing efforts and practices related to Software Source Code in academia

Software Source Code IG

Morane Gruenpeter
Inria, Software Heritage (France)

10th November 2020 - RDA 16th Virtual Plenary Meeting
Housekeeping

- Collaborative notes https://tinyurl.com/y2kunpf5
- These slides 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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong>: Interest Group goals and past activities</td>
<td>10</td>
</tr>
<tr>
<td><strong>Ice-breaker</strong>: Why are you interested in software source code?</td>
<td></td>
</tr>
<tr>
<td><strong>Software source Code IDentification (SCID) WG output:</strong></td>
<td>15</td>
</tr>
<tr>
<td>Use cases and identifiers schemes for software source code identification</td>
<td></td>
</tr>
<tr>
<td><strong>FAIR4 Research Software WG</strong>: Introduction and Invite to join the discussion</td>
<td>5</td>
</tr>
<tr>
<td><strong>FORCE11 Software Citation</strong> Implementation Working Group update</td>
<td>10</td>
</tr>
<tr>
<td>● including the ongoing task forces (CodeMeta, journals, repositories...)</td>
<td></td>
</tr>
<tr>
<td><strong>Overview of other ongoing efforts related to software</strong></td>
<td>10</td>
</tr>
<tr>
<td><strong>Group activity</strong>: Collecting existing practices</td>
<td>30</td>
</tr>
<tr>
<td><strong>Next steps for the SSC Interest Group</strong></td>
<td>10</td>
</tr>
</tbody>
</table>
Co-chairs:
● Neil Chue Hong
● Julia Collins
● Roberto Di Cosmo
● Mingfang Wu

VP16 coordinator:
● Morane Gruenpeter

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.

https://www.rd-alliance.org/groups/software-source-code-ig
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](http://www.force11.net) (FAIR4RS WG)
    - Welcome to join the work defining FAIR principles for research software
  - RDA & FORCE11 [Software Source Code Identification WG](http://www.force11.net) (SCID IG)
    - Output published in September 2020
  - FORCE11 [Software Citation Implementation Working Group](http://www.force11.net) (SCIWG)
    - Ongoing WG about software citation

- Related software sessions during VP16:
  - [Software Source Code IG meeting](http://www.force11.net)
  - [FAIR software roadmap BOF meeting](http://www.force11.net)
  - [FAIR 4 Research Software](http://www.force11.net) (FAIR4RS)
  - [Computational Notebooks BOF meeting](http://www.force11.net)
  - [CURE-FAIR WG meeting](http://www.force11.net)
  - [Research Data Management in Engineering: Data Provenance and Research Software in Engineering IG meeting](http://www.force11.net)
Why software source 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

Go to the code!
Why are you interested in Software Source Code?

https://tinyurl.com/y2kunpf5
Joint RDA & FORCE11 effort

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

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

Secretariat Liaison
● Stefanie Kethers

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
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
## The use cases collection (a small excerpt)

<table>
<thead>
<tr>
<th>Actor</th>
<th>Use case description</th>
<th>Action</th>
<th>Identification target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive</td>
<td>Identify all the software artifacts I hold</td>
<td>Archiving, referencing</td>
<td>Release and smaller artifacts</td>
</tr>
<tr>
<td>Citation manager</td>
<td>Curate the software citation entries</td>
<td>Credit</td>
<td>Project, release</td>
</tr>
<tr>
<td>Curator / librarian / digital archivist</td>
<td>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)</td>
<td>Archiving</td>
<td>Project, release and smaller artifacts depending on the reference</td>
</tr>
<tr>
<td>Publisher</td>
<td>Create/retrieve identifiers quickly for use in the paper for all software including commercial packages.</td>
<td>Referencing, describing</td>
<td>Any item (all granularity levels)</td>
</tr>
<tr>
<td>Registry</td>
<td>Identify and curate the software entries I hold</td>
<td>Archiving, referencing, describing, credit</td>
<td>Project</td>
</tr>
<tr>
<td>Researcher as a software user (RSU)</td>
<td>Access and use SSC no longer available on a collaborative platform</td>
<td>Archiving</td>
<td>Snapshot, release, revision, directory</td>
</tr>
</tbody>
</table>
Identifiers schemes

HAL - ID

Digital Object Identifier

ASCL.net
Astrophysics Source Code Library

ARK
Archival Resource Key

Software Identification

Handle
Handle System identifiers

swMATH
an Information service for mathematical software

SWHID
Software Heritage identifiers

Wiki Item identifier (Qxxx)
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
<table>
<thead>
<tr>
<th>Granularity level (GL)</th>
<th>ID target</th>
<th>Extrinisc identifiers</th>
<th>Intrinsic identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ASCL</td>
<td>ARK</td>
</tr>
<tr>
<td>GL1</td>
<td>project</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GL2</td>
<td>project version</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>GL3</td>
<td>module</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL4</td>
<td>repository</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL5</td>
<td>repository snapshot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL6</td>
<td>release</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL7</td>
<td>commit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL8</td>
<td>directory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL9</td>
<td>file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GL10</td>
<td>Code fragment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DOI: 10.15497/RDA00053
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)
FORCE11

Software Citation Implementation WG (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
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
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
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
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
CodeMeta initiative

- A subset of schema.org
- An academic community discussing software metadata
- A crosswalk table - mapping the metadata landscape

Software schemes

**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](image_url)

Most fields are optional. Mandatory fields will be highlighted when generating Codemeta.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>My Software the software title</td>
</tr>
<tr>
<td>Description</td>
<td>My Software computes ephemerides and orbit propagation. It has been developed from early ‘80.</td>
</tr>
<tr>
<td>Creation date</td>
<td>YYYY-MM-DD</td>
</tr>
<tr>
<td>First release date</td>
<td>YYYY-MM-DD</td>
</tr>
</tbody>
</table>

![CodeMeta generator](image_url)
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
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

[Link to document](community consultation ended on the 10.11.2020)
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

<table>
<thead>
<tr>
<th>Archive</th>
<th>Reference</th>
<th>Describe</th>
<th>Cite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long term preservation</td>
<td>Link and access</td>
<td>Search and find</td>
<td>Publish</td>
</tr>
</tbody>
</table>
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

https://tinyurl.com/y2kunpf5

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)
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
Thanks for joining