RDA Working Group: Software Source Code Identification

(this is a joint effort, coordinated with FORCE11)

Charter (A concise articulation of what issues the WG will address within a 18 month time frame and what its “deliverables” or outcomes will be.)

Software, and in particular source code, plays an important role in science: it is used in all research fields to produce, transform and analyse research data, and is sometimes itself an object of research and/or an output of research.

Unlike research data and scientific articles, though, software source code has only very recently been recognised as important subject matter in a few initiatives related to scholarly publication and archiving. These initiatives are now working on a variety of plans for handling the identification of software artifacts.

At the same time, unlike research data and scientific articles, the overwhelming majority of software source code is developed and used outside the academic world, in industry and in developer communities where software is routinely referenced, in practice, through methods that are totally different from the ones used in scholarly publications.

The objective of this working group is to bring together a broad panel of stakeholders directly involved in *software identification*.

The planned output will be concrete *recommendations* for the academic community to ensure that the solutions that will be adopted by the academic players are compatible with each other and especially with the software development practice of tens of millions of developers worldwide.

The output of this working group is highly relevant for the broader RDA community, because most research datasets are created and/or transformed using software, so a common standard for software identification will enable better traceability and reproducibility of research data.

Value Proposition (A specific description of who will benefit from the adoption or implementation of the WG outcomes and what tangible impacts should result)

The planned outcomes of the working group are recommendations and guidelines for software artifact identification (in particular in its source code form), targeted specifically at scholarly stakeholders that are willing to integrate software artifact into their workflow: scientific publishers, institutional repositories, and archives.

We believe that bringing together a broad panel of stakeholders is the best approach to avoid fragmentation in the emerging scholarly software identification landscape.

We also believe that connecting scholarly players with the daily practice of software development in industry will ease the adoption by these emerging scholarly initiatives of standards that are compatible with the well established practice of software development worldwide.

To this end, we plan to engage a dialogue with software industry bodies and software foundations that are working on standard approaches for identification of software components, like the Linux Foundation. An endorsement from such organizations would have a significant positive impact, as a shared standard will allow one to refer to both research and industry software in exactly the same way.

Engagement with existing work in the area (A brief review of related work and plan for engagement with any other activities in the area)

The initial participants of the working group are member of, or have direct connections with the following related initiatives:

- ***FORCE11 Software Citation Implementation WG***
  This group builds on the previous FORCE11 Software Citation Working Group, which developed and published an initial set of software citation principles (https://doi.org/10.7717/peerj-cs.86). The activities of the Software Citation Implementation Working Group will be conducted with relevant stakeholders (publishers, librarians, archivists, funders, repository developers, other community forums with related working groups, etc.) to: endorse the principles; develop sets of guidelines for implementing the principles; help implement the principles; and test specific implementations of the principles.
- ***Software Heritage***
  The Software Heritage archive provides unique, intrinsic, persistent identifiers for over 7 billion software source code artifacts worldwide, and is tightly connected with industry players working on source code qualification (Intel, Microsoft, Google, GitHub, Nokia Bell Labs, etc.)
- ***swMath***
  swMath is a project that has indexed and referenced over 20.000 research software projects in Mathematics
- ***DataCite***
  DataCite, working with about 100 members and 1,500 repositories, is providing persistent identifiers in the form of DOIs to scholarly outputs, including software.

- ***FREYA***
  The European Commission-funded FREYA project provides persistent identifier infrastructure for the European Open Science Cloud, and is working on increasing the adoption of persistent identifiers, including software.
- ***OpenAire***
  OpenAIRE is the European infrastructure in support of Open Science. It fosters and monitors the adoption of Open Science across Europe and beyond, at the level of the Countries for legal issues, and cross-boundaries to address research community specific requirements. In particular, it is building a portal indexing all open access articles, and will soon expand its scope to cover scientific software. **Work Plan** (A specific and detailed description of how the WG will operate including)

Related RDA Groups

We have identified the following initial list of RDA groups whose activity and scope is related to this working group:

- PID IG
- Reproducibility IG
- Data versioning WG
- Research Data Provenance IG
- Research Data Repository Interoperability WG
- Repository Platforms for Research Data IG

The target outcome of the working group is composed of the following documents that can be separated into two categories medium-term goals and long-term goals:

Medium-term goals (M12)

- An initial collection of software identification use cases and software identifier schemas.
- An overview of the different contexts in which software artifact identification is relevant, including
  - Scientific reproducibility
  - Fine grained reference to specific code fragments from scientific articles or documentation
  - Description of dependency information
  - Citation of software projects for proper credit attribution

Long-term goals (M18)

- Call out other RDA groups, in particular those working on citation and versioning issues, for consultation on the draft guidelines
- A set of guidelines for persistent software artifact identification, in each of the above contexts

Mode of operation

- Open a GitHub repository where issues are used to discuss topics that will be discussed and meetings are documented.
- Schedule a monthly on-line conf-call or group-mail informing the advancement made during the month and opening issues to discussion.
- Schedule meetings during the 13th, 14th, 15th and 16th plenaries (18M)

Timeline

Apr 19: [13th plenary] first meeting start discussion on medium-term goals
May 19 - Aug 19: medium-term goals
Sep 19: [14th plenary] progress report
Oct 19 - Feb 20: medium-term goals and long-term goals
Mar 20: [15th plenary] medium-term goals report and draft Long-term deliverable
Apr 20 - aug 20: long-term goals
Sep 20: [16th plenary] outputs publication

Adoption Plan (A specific plan for adoption or implementation of the WG outcomes within the organizations and institutions represented by WG members, as well as plans for adoption more broadly within the community. Such adoption or implementation should start within the 18 month timeframe before the WG is complete.)

Adoption by organizations and institutions represented by WG members

The first key step to broad adoption is to get the guidelines endorsed and adopted by all the initiatives that are represented in this working group: they are significant catalysers for adoption in the academic community.

Adoption by the academic community

The software identification guidelines are a stepping stone for software citation, where an identifier is needed to specify the exact software referenced, therefore its recommendations

will be the first output formalizing the way software source code should be referenced in the academic community. Potentially, the adoption of the software identification guidelines will provide a consensual solution to identifying software when citing software. It will be the first document produced by the academic community for software identification in a time when software is starting to be considered a legitimate product of research and its adoption will ensure a standardized approach to identify software in scholarly workflows that is compatible with the well established practice of software development.

Initial Membership (A specific list of initial members of the WG and a description of initial leadership of the WG.)

First Name	Last Name	Email	Institution	Role
Roberto	Di Cosmo	roberto@dicosmo.org	Inria/Software Heritage	co-Chair
Neil	Chue Hong	N.ChueHong@software.ac.uk	SSI	
Martin	Fenner	martin.fenner@datacite.org	Datacite FREYA	
Daniel S.	Katz	d.katz@ieee.org	University of Illinois	
Andrea	Dell'Amico		OpenAIRE (ISTI-CNR, Italy)	
Peter	Doorn		DANS	
Suenje	Dallmeier-Tiessen		CERN	
Wolfram	Sperber		swMATH	
Brian	Matthews		STFC	
Morane	Gruenpeter		Software Heritage/Crossminer	