Key topic

PIDs

Scientific domains

Earth and enviromental sciences

Life science

Photon & Neutron Sciences

Social and Humanities

Leading organisation

Software Heritage

Referencing software source code artifacts: identifiers for digital object

Context and materials

Software identification refers to multiple practices depending if you focus more on describing (i.e. attributing credit to authors) or on referencing software. **PIDs used to reference data sets, such as DOIs, are useful to reference a software as a project** (i.e. the software as a concept, not a digital object). But referencing software artifacts (i.e. digital objects) with different levels of granularity calls for specific identifiers. **Therefore identifiers in Software Heritage allow to reference a specific version of the source code of a project, at different levels of granularity: a snapshot, a release, a directory, down to a single file.**
SWHID are unique identifiers intrinsically bound to the software components. The difference between extrinsic and intrinsic identifiers lies in the way the relation between identifier and designated object is created and maintained. SWHID don't rely on an external register. Thus, end-users can recompute identifiers on retrieved objects and verify the match.
An agreement on a standard therefore, in June 2023, a first stable version of the SWHID specification was published: it describes precisely how SWHIDs are computed. A working group gathering experts from different institutions had been launched in March 2023. The relevance of SWHIDs goes way beyond source code and Software Heritage.
The attribution of SWHID raises the fact that deduplication has to be built-in, the database of the archive itself has to implement an ad hoc data model. Thus, any software artifacts encountered in the wild gets added to Software Heritage only if a corresponding node with a matching intrinsic identifier is not already available in the graph—file content, commits, entire directories or project snapshots are all deduplicated, incurring storage costs only once.
The FAIR-IMPACT project is an opportunity to elaborate compelling use cases in order to help the stakeholders adopt best practices for software referencing.

Challenges that need to be addressed

The related challenges vary somewhat depending on the stakeholder at hand. We have defined separate sets of challenges for end-users, service and infrastructure providers and policy-makers.
The end-user oriented challenges lie in understanding why software calls for a specific PID, as software and data are distinct concerns, and in understanding the type of PID that is compatible with a particular need. For example, an end-user may need to cite a software as a project or to cite a fragment of source code. Yet, finding the exact matching code can be quite difficult, as the code excerpt is often edited a bit with respect to the original, e.g. to drop details that are not relevant for the discussion or due to space limitations.
The challenges of service and infrastructure (archives, aggregators, catalogs) **providers relate to ensuring precise identification of software artifacts for reuse and reproducibility and in implementing workflows that ease the use of SWHIDs**, e.g. the French repository HAL provides a deposit service that allows to archive a software via a SWHID and the Image Processing On Line journal (IPOL) has decided to deposit systematically in the Software Heritage archive all the software artifacts associated to the articles it publishes. Other identified challenges is monitoring the adoption of SWHID among users via dedicated indicators, offering guidance to end-users, making the use of SWHID part of "current" science in all the academic fields, not only in computing science, and dealing with a posteriori curation for non-referenced software artifacts.
Policy-makers oriented challenges relate to ensuring that proper support actions are offered which promote the adoption of SWHID for software artifacts as early as possible in the software lifecycle and which assist in software archiving, as a prerequisite to reference software artifacts. In addition, regular monitoring efforts are needed to gauge the gap between current practices and recommendations.

Expected impact and outputs of the Use Case

Dissemination of software referencing good practices among stakeholders, partners and end-users from different scientific communities. It will result in a better understanding of the different use-cases related to SWHID.

Contributors

Sabrina Granger
INRIA

Josefine Nordling
CSC