



# Archiving and referencing source code with Software Heritage

Roberto Di Cosmo

## ► To cite this version:

Roberto Di Cosmo. Archiving and referencing source code with Software Heritage. 2020. hal-02526083v1

**HAL Id: hal-02526083**

**<https://hal.science/hal-02526083v1>**

Preprint submitted on 31 Mar 2020 (v1), last revised 1 Jun 2020 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Archiving and referencing source code with Software Heritage

Roberto Di Cosmo<sup>1</sup>[0000–0002–7493–5349]

Software Heritage, Inria and University of Paris, France roberto@dicosmo.org

**Abstract.** Software, and software source code in particular, is widely used in modern research. It must be properly archived, referenced, described and cited in order to build a stable and long lasting corpus of scientific knowledge. In this article we show how the Software Heritage universal source code archive provides a means to fully address the first two concerns, by archiving seamlessly all publicly available software source code, and by providing *intrinsic persistent identifiers* that allow to reference it at various granularities in a way that is at the same time convenient and effective. We call upon the research community to adopt widely this approach.

**Keywords:** Software source code · archival · reference · reproducibility

## 1 Introduction

Software source code is *an essential research output*, and there is a growing general awareness of its importance for supporting the research process [6,22,15]. Many research communities focus on the issue of *scientific reproducibility* and strongly encourage making the source code of the artefact available by archiving it in publicly-accessible long-term archives; some have even put in place mechanisms to assess research software, like the *Artefact Evaluation* process introduced in 2011 and now widely adopted by many computer science conferences [7], and the *Artifact Review and Badging* program of the ACM [4]. Other raise the complementary issues of making it easier to discover existing research software, and giving academic credit to authors [20,16,17].

As a first step, it is important to clearly identify the different concerns that come into play when addressing software, and in particular its source code, as a research output, that can be classified as follows:

**Archival:** software artifacts must be properly **archived**, to ensure we can *retrieve* them at a later time;

**Reference:** software artifacts must be properly **referenced** to ensure we can *identify* the exact code, among many potentially archived copies, used for reproducing a specific experiment;

**Description:** software artifacts must be equipped with proper **metadata** to make it easy to *find* them in a catalog or through a search engine;

**Citation:** research software must be properly **cited** in research articles in order to give *credit* to the people that contributed to it.

As already pointed out in the literature, these are not only different concerns, but also *separate* ones. Establishing proper *credit* for contributors via *citations* or providing proper metadata to *describe* the artifacts requires a *curation* process [5,2,10] and is way more complex than simply providing stable, intrinsic identifiers to *reference* a precise version of a software source code for reproducibility purposes [16,3,12]. Also, as remarked in [15,3], research software is often a thin layer on top of a large number of software dependencies that are developed and maintained outside of academia, so the usual approach based on institutional archives is not sufficient to cover all the software that is relevant for reproducibility of research.

In this article, we focus on the first two concerns, *archival* and *reference*, showing how they can be addressed fully by leveraging the Software Heritage universal archive [1], and also mention some recent evolutions in best practices for embedding *metadata* in software development repositories.

In Section 2 we briefly recall what is Software Heritage and what makes it special; in Section 3 we show how researchers can easily ensure that any relevant source code is archived; in Section 4 we explain how to use the *intrinsic identifiers* provided by Software Heritage to enrich research articles, making them

more useful and appealing for the readers, and providing stable links between articles and source code in the web of scientific knowledge we are all building. Finally, we point to ongoing collaborations and future perspectives in Section 5.

## 2 Software Heritage: the universal archive of software source code

Software Heritage [13,1] is a non profit, long term universal archive specifically designed for software source code, and able to store not only a software artifact, but *also its full development history*.

Software Heritage’s mission is to collect, preserve, and make easily accessible the source code of *all* publicly available software. Among the strategies designed for collecting the source code there is the development of a large scale automated crawler for source code, whose architecture is shown in Figure 1.

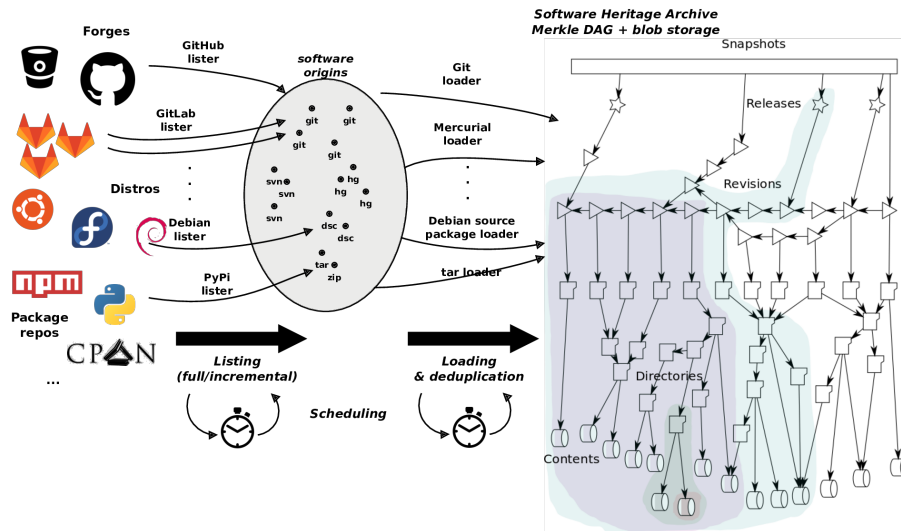


Fig. 1: Architecture of the Software Heritage crawler

We recall here a few key properties that set Software Heritage apart from all other scholarly infrastructures:

- it *proactively* archives *all software*, making it possible to store and reference any piece of publicly available software relevant to a research result, independently from any specific field of endeavour, and even when the author(s) did not take any step to have it archived [13,1];
- it stores the source code with its development history in a uniform data structure, a Merkle DAG [18], that allows to provide uniform, *intrinsic* identifiers for the billions of software artifacts of the archive, independently of the version control system or package format used [12].

At the time of writing this article, the Software Heritage archive contains over 7 billions unique source code files, from more than 100 million different software origins<sup>1</sup>. It provides the ideal place to *preserve research software artifacts*, and offers powerful mechanisms to *enhance research articles* with precise references to relevant fragments of your source code. Using Software Heritage is straightforward and involves very simple steps, that we detail in the following sections.

<sup>1</sup> See <https://archive.softwareheritage.org> for the up to date figures.

### 3 Archiving and self archiving

In a research article one may want to reference different kinds of source code artifacts: some may be popular open source components, some may be general purpose libraries developed by others, and some may be one own's software projects.

All these different kinds of software artifacts can be archived extremely easily in Software Heritage: it's enough that their source code is hosted on a publicly accessible repository (Github, Bitbucket, any GitLab instance, an institutional software forge, etc.) using one of the version control systems supported by Software Heritage, currently Subversion, Mercurial and Git <sup>2</sup>.

For source code developed on popular development platforms, chances are that the code one wants to reference is already archived in Software Heritage, but one can make sure that the archived version history is fully up to date, as follows:

- go to <https://save.softwareheritage.org>,
- pick the right version control system in the drop-down list, enter the code repository url <sup>3</sup>,
- click on the Submit button (see Figure 2).



The image shows a web form titled 'Save Code Now'. It has two main input fields: 'Origin type' and 'Origin url'. The 'Origin type' field is a dropdown menu with 'git' selected. The 'Origin url' field is a text input box. To the right of these fields is a 'Submit' button.

Fig. 2: The Save Code Now form

That's all. No need to create an account or disclose personal information of any kind. If the provided URL is correct, Software Heritage will archive the repository shortly after, with its full development history. If it is hosted on one of the major forges we already know, this process will take just a few hours; if it is in a location we never saw before, it can take longer, as it will need to be manually screened <sup>4</sup>.

#### 3.1 Preparing source code for self archiving

In case the source code is one own's, before requesting its archival it is important to structure the software repository follow well established good practices for release management [19]. In particular one should add README and AUTHORS files as well as licence information following industry standard terminology [14,21].

Future users that find the artifact useful might want to give credit by citing it. To this end, one might want to provide instructions on how one prefers the artifact to be cited. We would recommend to also provide structured metadata information in machine readable formats. While practices in this area are still evolving, one can use the CodeMeta generator available at <https://codemeta.github.io/codemeta-generator/> to produces metadata conformant to the CodeMeta schema: the JSON-LD output can be put at the root of the project in a **codemeta.json** file. Another option is to use the Citation File Format, CFF (usually in a file named **citation.cff**).

<sup>2</sup> For up to date information, see <https://archive.softwareheritage.org/browse/origin/save/>

<sup>3</sup> Make sure to use the clone/checkout url as given by the development platform hosting your code. It can easily be found in the web interface of the development platform.

<sup>4</sup> It is also possible to request archival programmatically, using the Software Heritage API, which can be quite handy to integrate in a Makefile; see <https://archive.softwareheritage.org/api/1/origin/save/> for details.

## 4 Referencing

Once the source code has been archived, the Software Heritage *intrinsic identifiers*, called SWH-ID, fully documented online and shown in Figure 3, can be used to reference with great ease any version of it.

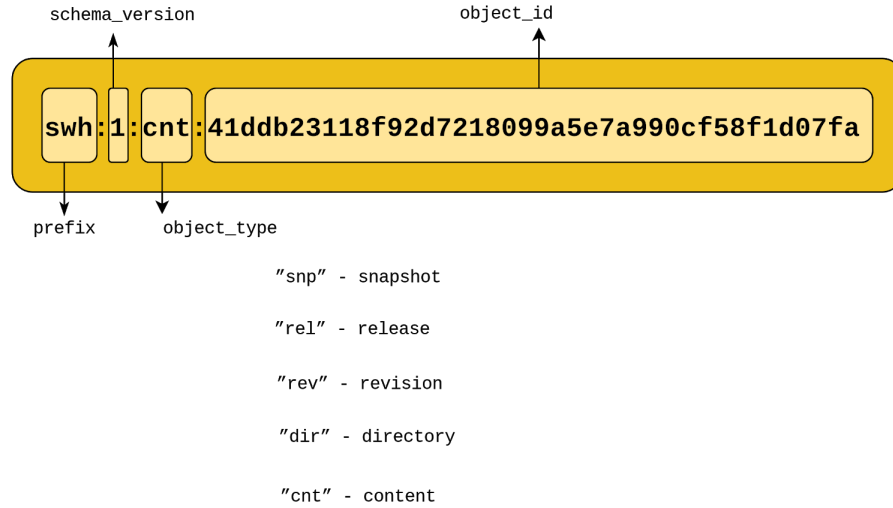


Fig. 3: Schema of the core Software Heritage identifiers

SWH-IDs are URIs with a very simple schema: the `swh` prefix makes explicit that these identifiers are related to Software Heritage; the colon (`:`) is used as separator between the logical parts of identifiers; the schema version (currently 1) is the current version of this identifier schema; then follows the type of the objects identified and finally comes a hex-encoded (using lowercase ASCII characters) cryptographic signature of this object, computed in a standard way, as detailed in [11,12].

These core identifiers may be equipped with the *qualifiers* that carry contextual *extrinsic* information about the object:

**origin** : the *software origin* where an object has been found or observed in the wild, as an URI;

**visit** : persistent identifier of a *snapshot* corresponding to a specific *visit* of a repository containing the designated object;

**anchor** : a *designated node* in the Merkle DAG relative to which a *path to the object* is specified;

**path** : the *absolute file path*, from the *root directory* associated to the *anchor node*, to the object;

**lines** : *line number(s)* of interest, usually within a content object

The combination of the core SWH-IDs with these qualifiers provides a very powerful means of referring in a research article to all the software artefacts of interest.

To make this concrete, in what follows we use as a running example the article *A “minimal disruption” skeleton experiment: seamless map and reduce embedding in OCaml* by Marco Danelutto and Roberto Di Cosmo [8] published in 2012. This article introduced an elegant library for multicore parallel programming that was distributed via the `gitorious.org` collaborative development platform, at `gitorious.org/parmap`. Since Gitorious has been shut down a few years ago, like Google Code and CodePlex, this example is particularly fit to show why pointing to an *archive* of the code is better than pointing to the collaborative development platform where it is developed.

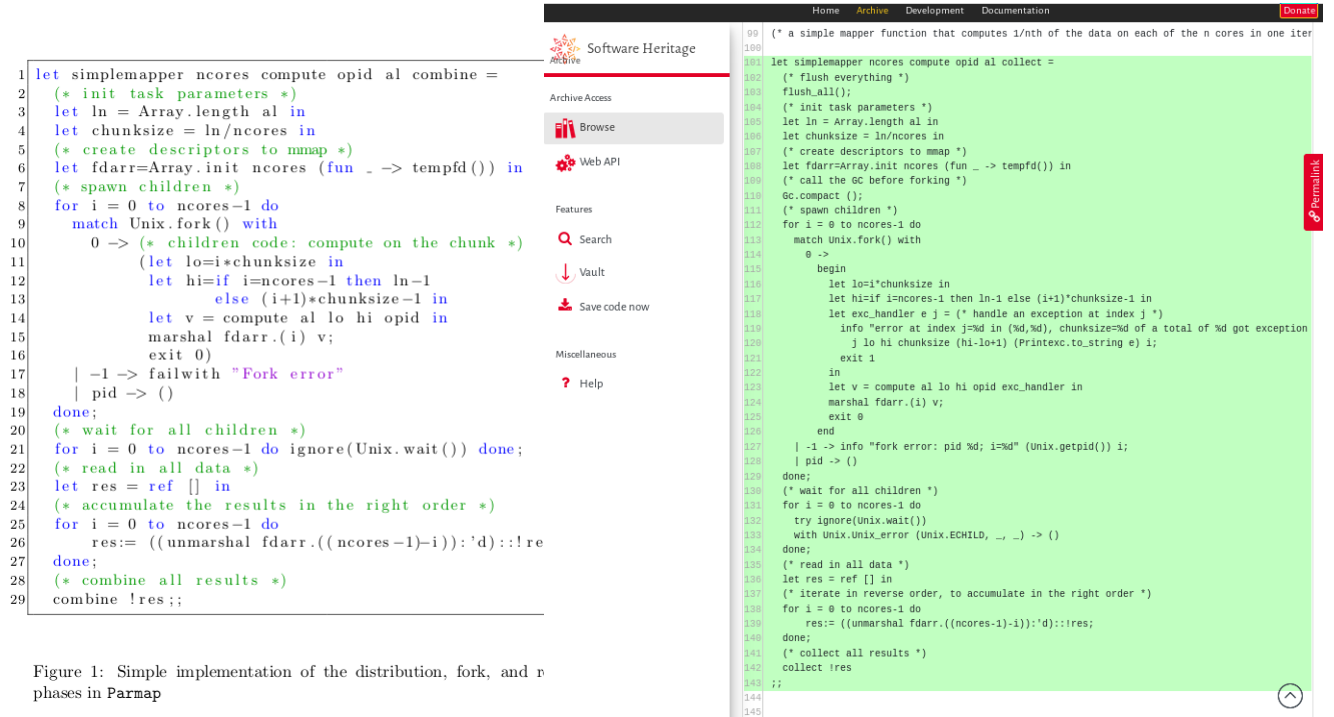


Figure 1: Simple implementation of the distribution, fork, and r phases in Parmap

(a) as presented in the article [8]

(b) as archived in Software Heritage

Fig. 4: Code fragment from the published article compared to the content in the Software Heritage archive

#### 4.1 Specific version

The Parmap article describes a *specific version* of the Parmap library, the one that was used for the experiments reported in the article, so in order to support reproducibility of these results, we need to be able to pinpoint precisely the state(s) of the source code used in the article.

The exact revision of the source code of the library used in the article has the following SWH-ID:

```
swh:1:rev:0064fbd0ad69de205ea6ec6999f3d3895e9442c2;
origin=https://gitorious.org/parmap/parmap.git;
visit=swh:1:snp:78209702559384ee1b5586df13eca84a5123aa82
```

This identifier can be turned into a clickable URL by prepending to it the prefix `https://archive.softwareheritage.org/` (one can try it by clicking on this link).

#### 4.2 Code fragment

Having a link to the exact archived revision of a software project is important in all research articles that use software, and the core SWH-IDs allow to drill down and point to a given directory or even a file content, but sometimes, like in our running example, one would like to do more, and pinpoint a fragment of code inside a specific version of a file. This is possible using the `lines=` qualifier available for identifiers that point to file content.

Let's see this feature at work in our running example, showing how the experience of studying or reviewing an article can be greatly enhanced by providing pointers to code fragments.

In Figure 1 of [8], which is shown here as Figure 4a, the authors want to present the core part of the code implementing the parallel functionality that constitutes the main contribution of their article. The

usual approach is to typeset in the article itself *an excerpt of the source code*, and let the reader try to find it by delving into the code repository, which may have evolved in the mean time. Finding the exact matching code can be quite difficult, as the code excerpt is *often edited* a bit with respect to the original, sometimes to drop details that are not relevant for the discussion, and sometimes due to space limitations.

In our case, the article presented 29 lines of code, slightly edited from the 43 actual lines of code in the Parmap library: looking at 4a, one can easily see that some lines have been dropped (102-103, 118-121), one line has been split (117) and several lines simplified (127, 132-133, 137-142).

Using Software Heritage, the authors can do a much better job, because the original code fragment can now be precisely identified by the following Software Heritage identifier:

```
swh:1:cnt:d5214ff9562a1fe78db51944506ba48c20de3379;
origin=https://gitorious.org/parmap/parmap.git;
visit=swh:1:snip:78209702559384ee1b5586df13eca84a5123aa82;
anchor=swh:1:rev:0064fbd0ad69de205ea6ec6999f3d3895e9442c2;
path=/parmap.ml;
lines=101-143
```

This identifier will **always** point to the code fragment shown in Figure 4b.

The caption of the original article shown in Figure 4a can then be significantly enhanced by incorporating a clickable link containing the SWH-ID shown above: it's all is needed to point to the exact source code fragment that has been edited for inclusion in the article, as shown in Figure 5. The link contains, thanks to the SWH-ID qualifiers, all the contextual information necessary to identify the context in which this code fragment is intended to be seen.

Simple implementation of the distribution, fork, and recollection phases in `Parmap` (slightly simplified from the the actual code in the version of `Parmap` used for this article)

Fig. 5: A caption text with the link to the code fragment and its contextual information

When clicking on the hyperlinked text in the caption shown above, the reader is brought seamlessly to the Software Heritage archive on a page showing the corresponding source code archived in Software Heritage, with the relevant lines highlighted (see Figure 4b).

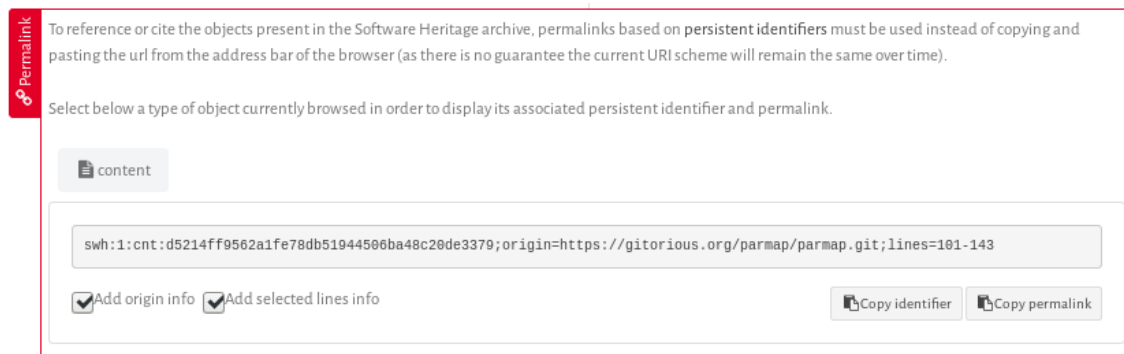
### 4.3 Getting the SWH-ID

A fully qualified SWH-ID is rather long, and it needs to be, as it contains quite a lot of information that is essential to convey. In order to make it easy to use SWH-IDs, we provide a very simple way of getting the right SWH-ID without having to type it by hand. Just browse the archived code in Software Heritage and navigate to the software artifact of interest. Clicking on the *permalinks vertical red tab* that is present on all pages of the archive, opens up a tab that allows to select the identifier for the object of interest: an example is shown in Figure 6.

The two buttons on the bottom right allow to copy the identifier or the full permalink in the clipboard, and to paste it in an article as needed.

### 4.4 Generating and verifying SWH-IDs

An important consequence of the fact that SWH-IDs are *intrinsic identifiers* is that they can be generated and verified *independently* of Software Heritage, using `swh-identify`, an open source tool developed



**Permalink**

To reference or cite the objects present in the Software Heritage archive, permalinks based on persistent identifiers must be used instead of copying and pasting the url from the address bar of the browser (as there is no guarantee the current URI scheme will remain the same over time).

Select below a type of object currently browsed in order to display its associated persistent identifier and permalink.

☐ content

swh:1:cnt:d5214ff9562a1fe78db51944506ba48c20de3379;origin=https://gitorious.org/parmap/parmap.git;lines=101-143

☒ Add origin info ☒ Add selected lines info

Fig. 6: Obtaining a Software Heritage identifier using the permalink box on the archive Web user interface

by Software Heritage, and distributed via PyPI as `swh.model`, with the stable version at the time of writing being {this one}.

Version 1 of the SWH-IDs uses git-compatible hashes, so if the source code that one wants to reference uses git as a version control system, one can create the right SWH-ID by just prepending `swh:1:rev:` to the commit hash. This comes handy to automate the generation of the identifiers to be included in an article, as one will always have code and article in sync.

## 5 Perspectives for the scholarly world

We have shown how Software Heritage and the associated SWH-IDs enables the seamless archival of all publicly available source code. It provides for all kind of software artifacts the *intrinsic identifiers* that are needed to establish long lasting, resilient links between research articles and the software they use or describe.

All researchers can use *right now* the mechanisms presented here to produce improved and enhanced research articles. More can be achieved by establishing collaborations with academic journals, registries and institutional repositories and registries, in particular in terms of description and support for software citation. Among the initial collaborations that have been already established, we are happy to mention the cross linking with the curated mathematical software descriptions maintained by the `swMath.org` portal [5], and the curated deposit of software artefacts into the HAL french national open access portal [10], which is performed via a standard SWORD protocol interface, an approach that is currently being explored by other academic journals.

We believe that the time has come to see software become a first class citizen in the scholarly world, and Software Heritage provides a unique infrastructure to support an open, non profit, long term and resilient web of scientific knowledge.

### 5.1 Acknowledgements

This article is a major evolution of the research software archival and reference guidelines available on the Software Heritage website [9] resulting from extensive discussions that took place over several years with many people. Special thanks to Alain Girault, Morane Gruenpeter, Antoine Lambert, Julia Lawall, Arnaud Legrand, Nicolas Rougier and Stefano Zacchiroli for their precious feedback on these issues and/or earlier versions of this document.



## References

1. J.-F. Abramatic, R. Di Cosmo, and S. Zacchiroli. Building the universal archive of source code. *Communications of the ACM*, 61(10):29–31, Sept. 2018.
2. A. Allen and J. Schmidt. Looking before leaping: Creating a software registry. *Journal of Open Research Software*, 3(e15), 2015.
3. P. Alliez, R. Di Cosmo, B. Guedj, A. Girault, M.-S. Hacid, A. Legrand, and N. Rougier. Attributing and referencing (research) software: Best practices and outlook from inria. *Computing in Science Engineering*, 22(1):39–52, Jan. 2020. Available from <https://hal.archives-ouvertes.fr/hal-02135891>.
4. Association for Computing Machinery. Artifact review and badging. <https://www.acm.org/publications/policies/artifact-review-badging>, Apr. 2018. Retrieved April 27th 2019.
5. S. Bönisch, M. Brickenstein, H. Chrapary, G. Greuel, and W. Sperber. swMATH - A new information service for mathematical software. In *MKM/Calculus/DML*, volume 7961 of *Lecture Notes in Computer Science*, pages 369–373. Springer, 2013.
6. C. L. Borgman, J. C. Wallis, and M. S. Mayernik. Who’s got the data? interdependencies in science and technology collaborations. *Computer Supported Cooperative Work*, 21(6):485–523, 2012.
7. B. R. Childers, G. Fursin, S. Krishnamurthi, and A. Zeller. Artifact Evaluation for Publications (Dagstuhl Perspectives Workshop 15452). *Dagstuhl Reports*, 5(11):29–35, 2016.
8. M. Danelutto and R. Di Cosmo. A “Minimal Disruption” skeleton experiment: Seamless map & reduce embedding in OCaml. *Procedia CS*, 9:1837–1846, 2012.
9. R. Di Cosmo. How to use Software Heritage for archiving and referencing your source code: guidelines and walkthrough. Available at <https://hal.archives-ouvertes.fr/hal-02263344>, Apr. 2019.
10. R. Di Cosmo, M. Gruenpeter, B. P. Marmol, A. Monteil, L. Romary, and J. Sadowska. Curated Archiving of Research Software Artifacts : lessons learned from the French open archive (HAL). Presented at the International Digital Curation Conference, submitted to IJDC, Dec. 2019.
11. R. Di Cosmo, M. Gruenpeter, and S. Zacchiroli. Identifiers for digital objects: the case of software source code preservation. In *Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, USA*, Sept. 2018.
12. R. Di Cosmo, M. Gruenpeter, and S. Zacchiroli. Referencing source code artifacts: a separate concern in software citation. *Computing in Science & Engineering*, 22(2):33–43, Mar. 2020.
13. R. Di Cosmo and S. Zacchiroli. Software Heritage: Why and how to preserve software source code. In *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017*, Sept. 2017.
14. F. S. F. Europe. REUSE software. <https://reuse.software>, Sept. 2019. Accessed on 2019-09-24.
15. K. Hinsén. Software development for reproducible research. *Computing in Science and Engineering*, 15(4):60–63, 2013.
16. J. Howison and J. Bullard. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9):2137–2155, 2016.
17. A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpi, N. Chue Hong, C. Goble, and S. Capella-Gutierrez. Towards FAIR principles for research software. Preprint:1–23, 2019. Preprint.
18. R. C. Merkle. A digital signature based on a conventional encryption function. In *Advances in Cryptology - CRYPTO '87, A Conference on the Theory and Applications of Cryptographic Techniques, Santa Barbara, California, USA, August 16-20, 1987, Proceedings*, pages 369–378, 1987.
19. E. S. Raymond. Software release practice HOWTO. [https://www.tldp.org/HOWTO/html\\_single/Software-Release-Practice-HOWTO/](https://www.tldp.org/HOWTO/html_single/Software-Release-Practice-HOWTO/), Jan. 2013. Accessed on 2019-06-05.
20. A. M. Smith, D. S. Katz, and K. E. Niemeyer. Software citation principles. *PeerJ Computer Science*, 2:e86, 2016.
21. SPDX Workgroup. Software package data exchange licence list, 2019. <https://spdx.org/licenses-list>, retrieved 30 March 2020.
22. V. Stodden, R. J. LeVeque, and I. Mitchell. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science and Engineering*, 14(4):13–17, 2012.