Making Software FAIR: A machine-assisted workflow for the research software lifecycle

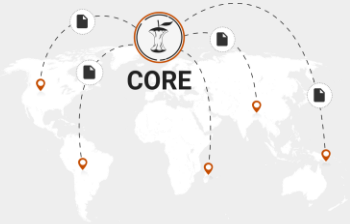
Petr Knoth[1], Laurent Romary[2], Patrice Lopez[3], Roberto Di Cosmo[2], Pavel Smrz[4], Tomasz Umerle[5], Melissa Harrison[6], Alain Monteil[2], Matteo Cancellieri[1], David Pride[1]

22nd October 2024

1: CORE, The Open University, United Kingdom;
2: Inria;
3: Science Miner;
4: Brno University of Technology;
5: Polish Academy of Sciences;
6: European Institute of Bioinformatics

SoFAIR

Big Scientific Data and Text Analytics group : AI for open and responsible research



CORE is the world's most used aggregator of **Open Access** papers, collating and enriching content from over **11,000 repositories**.

- **>30 Million** monthly active users (MAU)
- **34 Million** full-text research papers hosted by CORE.
- **260 Million** metadata records

Signatory of Principles of Open Scholarly Infrastructure **(POSI)**

CORE delivers **services** for HEIs, researchers, funders and commercial partners, offering seamless access to research.

Content discovery
- Search
- Recommender
- Discovery

Raw data services
- API
- Dataset
- FastSync

Managing content
- Repository Dashboard

Identifiers
- OAI Resolver

Providing seamless access to open research for humans and machines.

Commercial Partners

- Innovation and trends analysis
- Plagiarism detection
- Fact checking
- Finance
- Health

Institutional Members

32 supporting or sustaining members

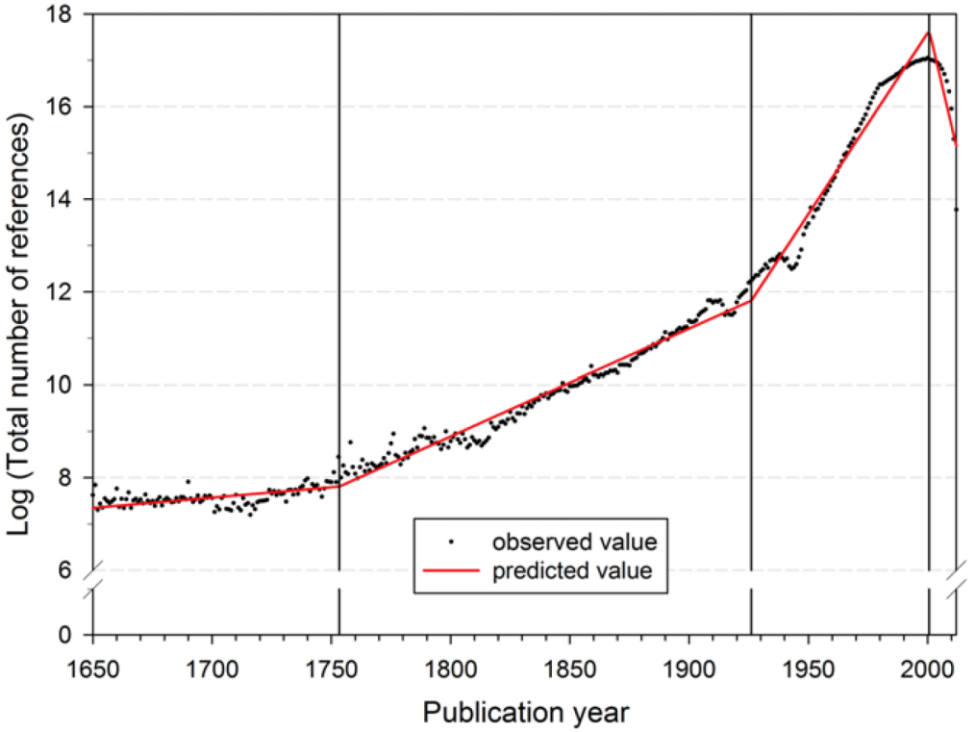
Research areas

- ➔ AI Applications in Research Evaluation (e.g. citation type classification, bibliometrics, impact assessment)
- ➔ Automatic Expert Finder systems (e.g. for peer-review and grant applications)
- ➔ Deduplication, document classification, rapid systematic reviews
- ➔ Research graphs: entity extraction (affiliation, author, etc.)
- ➔ Research recommender systems and academic search

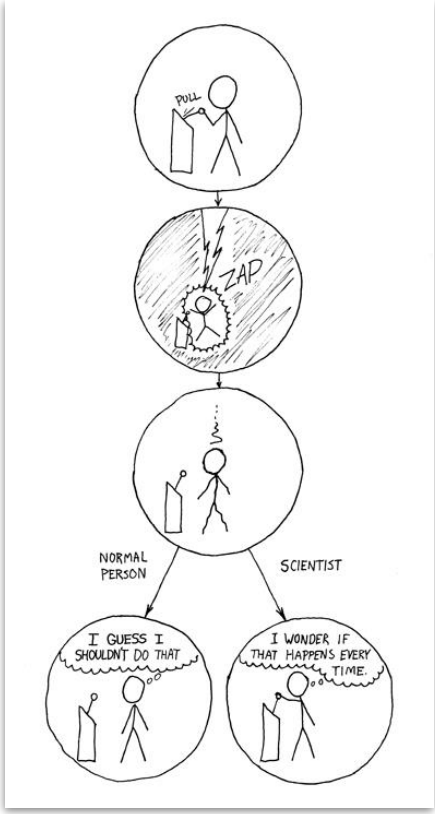
Dr. Petr Knoth : Senior Research Fellow in Text and Data Mining petr.knoth@open.ac.uk

Scientific output

Global scientific output
doubles every nine years
[Bornmann, 2015]



SoFAIR



“Single occurrences that cannot be reproduced are of no significance to science”

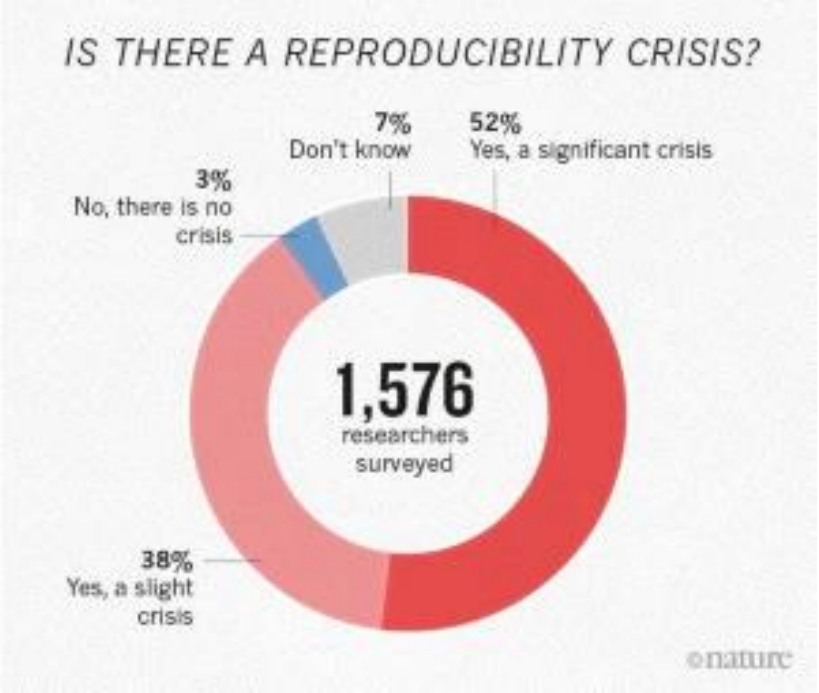
– Popper, 1935 –

Reproducibility crisis

More than **70%** of researchers have tried and failed to reproduce another scientist's experiments.

More than **50%** have failed to reproduce their own experiments.

The majority replied that there is a significant reproducibility crisis



Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). https://doi.org/10.1038/533452a

Reproducibility (according to Claerbout and Karrenbach, 1992)

“Reproducing”

means “running the same software on the same input data and obtaining the same results”

“Replicating”

means “writing and then running new software based on the description of a computational model or method provided in the original publication, and obtaining results that are similar enough …”

Reproducibility (according to ACM, 2016)

Repeatability — **same team | same experimental setup**

The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

Replicability — **different team | same experimental setup**

The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

Reproducibility — **different team | different experimental setup**

The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

Reproducibility (according to Goodman, 2016)

Methods reproducibility: provide sufficient detail about procedures and data so that the same procedures could be exactly repeated.

Results reproducibility: obtain the same results from an independent study with procedures as closely matched to the original study as possible.

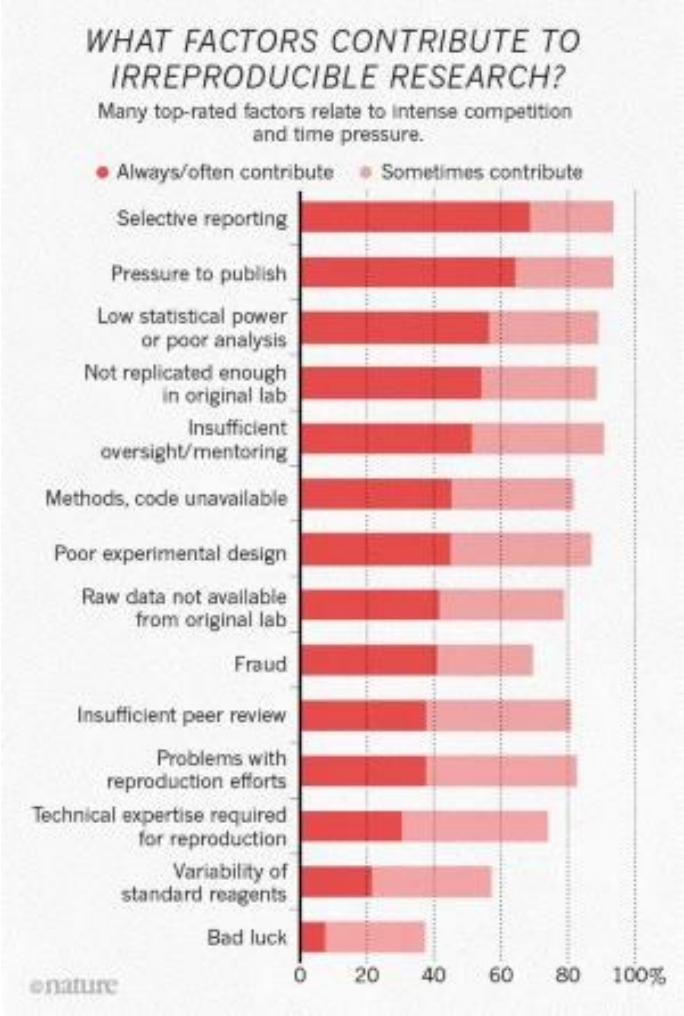
Inferential reproducibility: draw the same conclusions from either an independent replication of a study or a reanalysis of the original study.

Goodman	Claerbout	ACM
		Repeatability
Methods reproducibility	Reproducibility	Replicability
Results reproducibility	Replicability	Reproducibility
Inferential reproducibility		

Plesser HE (2018) **Reproducibility vs. Replicability: A Brief History of a Confused Terminology.** *Front. Neuroinform*. 11:76. doi: 10.3389/fninf.2017.00076

Reproducibility and SW

Unavailability of research software reported as the **6th** most significant reason for non-reproducibility.



Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). https://doi.org/10.1038/533452a

SoFAIR

The SoFAIR research problem

A key issue hindering discoverability, attribution and reusability of **open research software** is that its **existence often remains hidden within the manuscript of research papers.**

For these resources to become **first-class bibliographic records**, they first need to be identified and subsequently registered with persistent identifiers (PIDs) to be made FAIR (Findable, Accessible, Interoperable and Reusable).

To this day, much open research software fails to meet FAIR principles and software resources are mostly not explicitly linked from the manuscripts that introduced them or used them.

Vision - 1/2

Research software assets as first-class bibliographic records

Describe software assets using metadata

Assign PIDs for software assets

Make Software assets FAIR (SoFAIR :))

Motivation:

Incentivise good practices of software assets curation:

- Facilitate correct attribution
- Credit researchers for the creation of research software
- Reward research software creation in institutional promotion processes

Contribute to an open scholarly research graph connecting entities including papers, authors, institutions, data and software

Vision - 2/2

Scalable workflow for the software assets lifecycle for open repositories

Context

- Precious and limited time researchers have available
- Prevent re-typing of information
- Help researchers save time

Vision

- A **workflow for the management of the entire lifecycle of research software assets,** connecting and adapting existing open infrastructures and tools.
- **Establish a machine-assisted workflow embedded into widely used open scholarly infrastructures to assist researchers in identifying, describing, registering, linking and archiving research software.**

Machine assisted workflow for software assets lifecycle

Embed this workflow into established scholarly infrastructures, making the solution available to the global network of open repositories

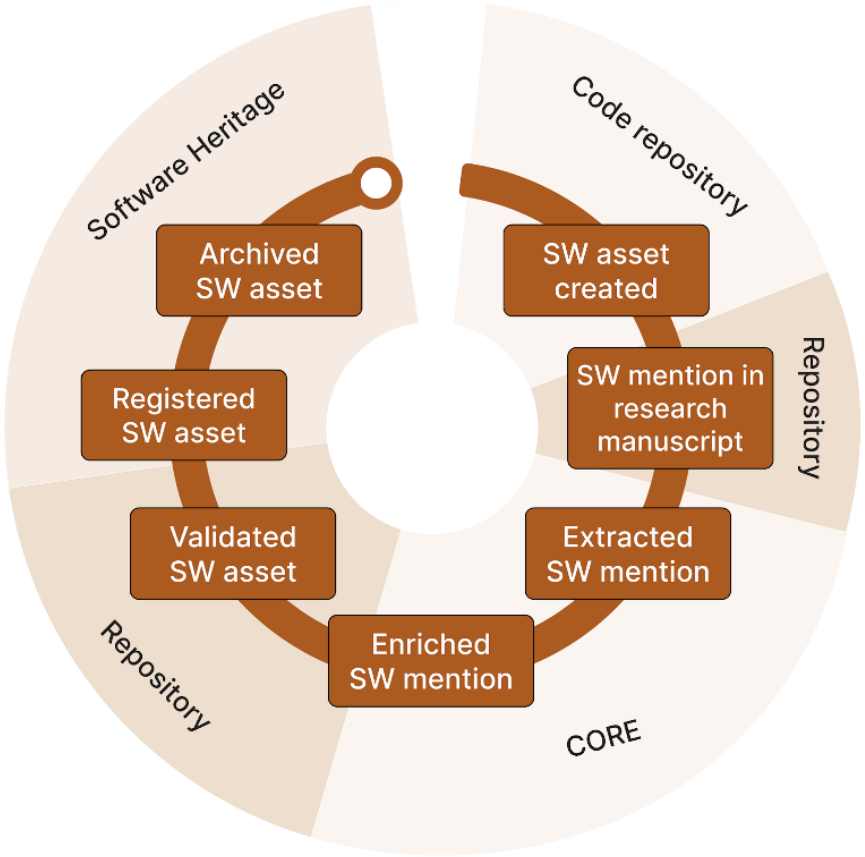
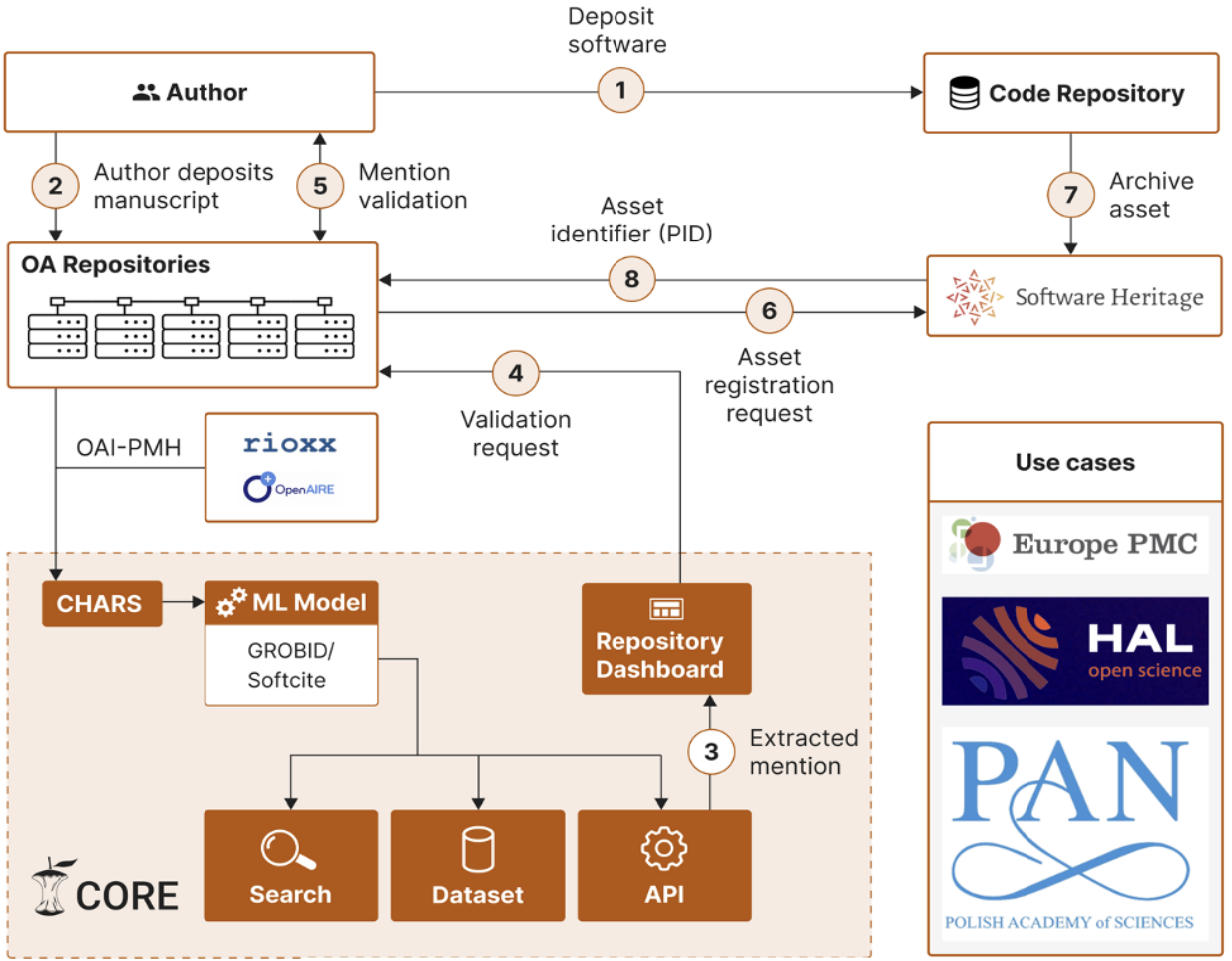


Figure 1: Software Asset Lifecycle

SoFAIR

Approach



SoFAIR

Infrastructures leveraged by the project

GROBID / Softcite	CORE (core.ac.uk)	HAL	Software Heritage
Progress beyond the state–of-the-art:			
SoFAIR will extend Softcite with models for multidisciplinary identification of software mentions from research manuscripts, including their disambiguation and enrichment.	CORE will enable the application of the developed ML-assisted workflow on both pre-existing and new open access content from any open repository in the world.	HAL will be used as a best practice example of a repository participating in the SoFAIR workflow, it will 1) participate in the routing of identified software mentions coming from CORE for validation by INRIA authors and 2) adapt its repository software to expose (over OAI-PMH) links between research manuscripts and software assets used in their creation.	Software Heritage will be embedded into the introduced workflow to support the registration of newly identified and validated software assets with PIDs and their subsequent archival.

Demonstrators and use cases

➔ Demonstrator 1 (EMBL-EBI):

Linking research studies to software in life sciences for Europe PMC

➔ Demonstrator 2 (INRIA):

Validating extracted software mentions within an institutional repository

➔ Case study in the digital humanities (IBL-PAN)

Key innovations

A novel machine-assisted workflow for software assets lifecycle management

New machine learning models for software mentions extraction and disambiguation

Scalable application of the technology across open repositories and relevance to both pre-existing and new software assets

Some of the main challenges we are facing

- ➔ Coming up with an effective and efficient annotation schema
- ➔ Performance (p/r) of our automatic software mention extraction and disambiguation models
- ➔ Scalability of our solution and demonstrating that it can be applied to both pre-existing and new articles.
- ➔ Routing discovered software mentions back to authors via repository software and making them to act on them to register software PIDs
- ➔ Propagating registered software PIDs to research outputs metadata
- ➔ Workflow integration

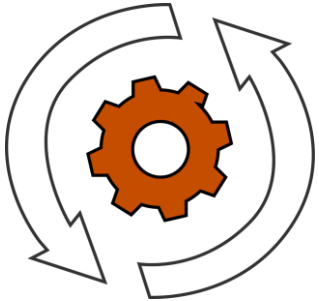
Expected impact

Increase the likelihood of both pre-existing and new research software assets being identified, registered and archived.

Lower the barriers for making research software FAIR

Facilitate seamless access to, and management of, increasing volumes of research software.

Contribute to equity and fairness in academia and increasing public trust in the research process.



Enable more researchers to cite and give credit to research software

Increase the ability of researchers to demonstrate their impact by quantifying the reuse of their research software

Facilitate and contribute to the reproducibility of research software

A solution for both pre-existing and new manuscripts

Reduce the barriers for entrepreneurs and enterprises to innovate

Incentivise researchers to develop better and more reusable software

Conclusions

- Recognising and archiving software assets mentioned in research manuscripts is one of the preconditions for solving the reproducibility crisis.
- NLP / AI offers new opportunities to address and help semi-automate this
- SoFAIR is developing a new workflow that will enable better management of SW assets leveraging CORE and Software Heritage