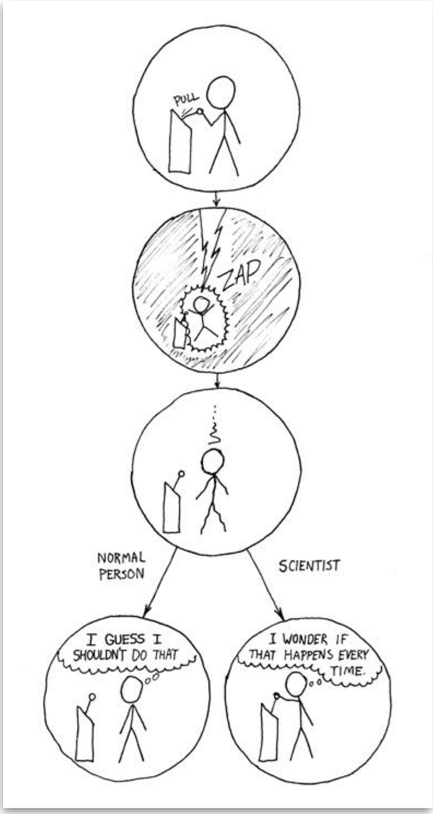
Making Software FAIR: A machine-assisted workflow for the research software lifecycle

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]

Paris, 29th January 2025

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


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

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

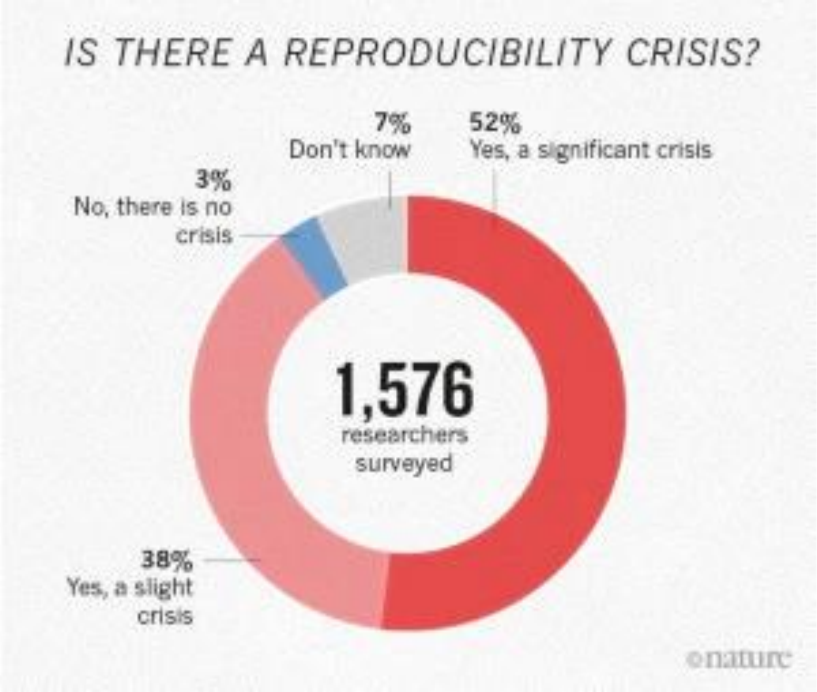
– Popper, 1935 –

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.

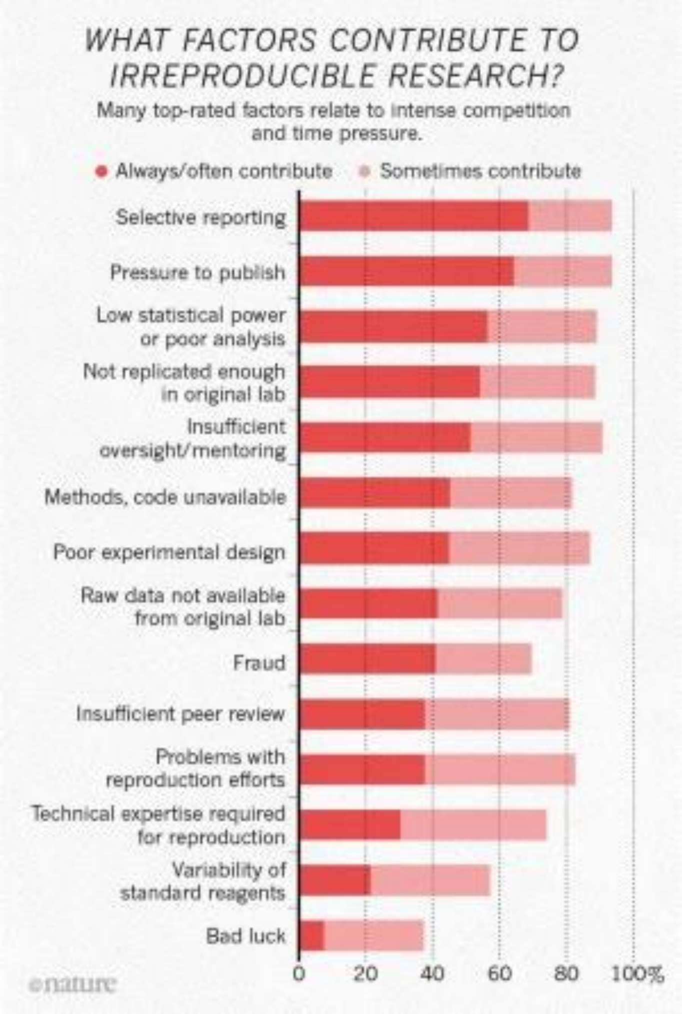
The majority replied that there is a significant reproducibility crisis



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

Reproducibility 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).

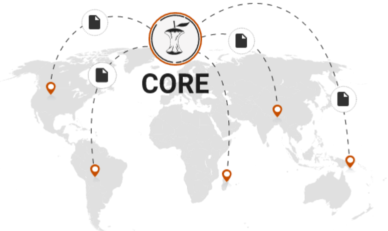
To this day, much open research software fails to meet FAIR principles and software resources are mostly not explicitly linked from the manuscripts that introduced them or used them.

SoFAIR project mission and partners

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

- 2 year CHIST-ERA project
- 5 partners:
  - CORE, The Open University, UK
  - INRIA, FR: (for Software Heritage and HAL)
  - Brno University of Technology, CZ
  - IBL-PAN, Poland
  - Europe PMC
- 4 countries

What is CORE



CORE's mission is

to index open access research worldwide and deliver unrestricted access for all.

We are here to support you and to advance the Open Access / Open Research movement

WE ARE

the world's **most used scholarly database** of open access research papers with >30 million monthly active users

WE ARE

a **not-for-profit** scholarly infrastructure dedicated to the open access mission, **adopters of POSI** principles.

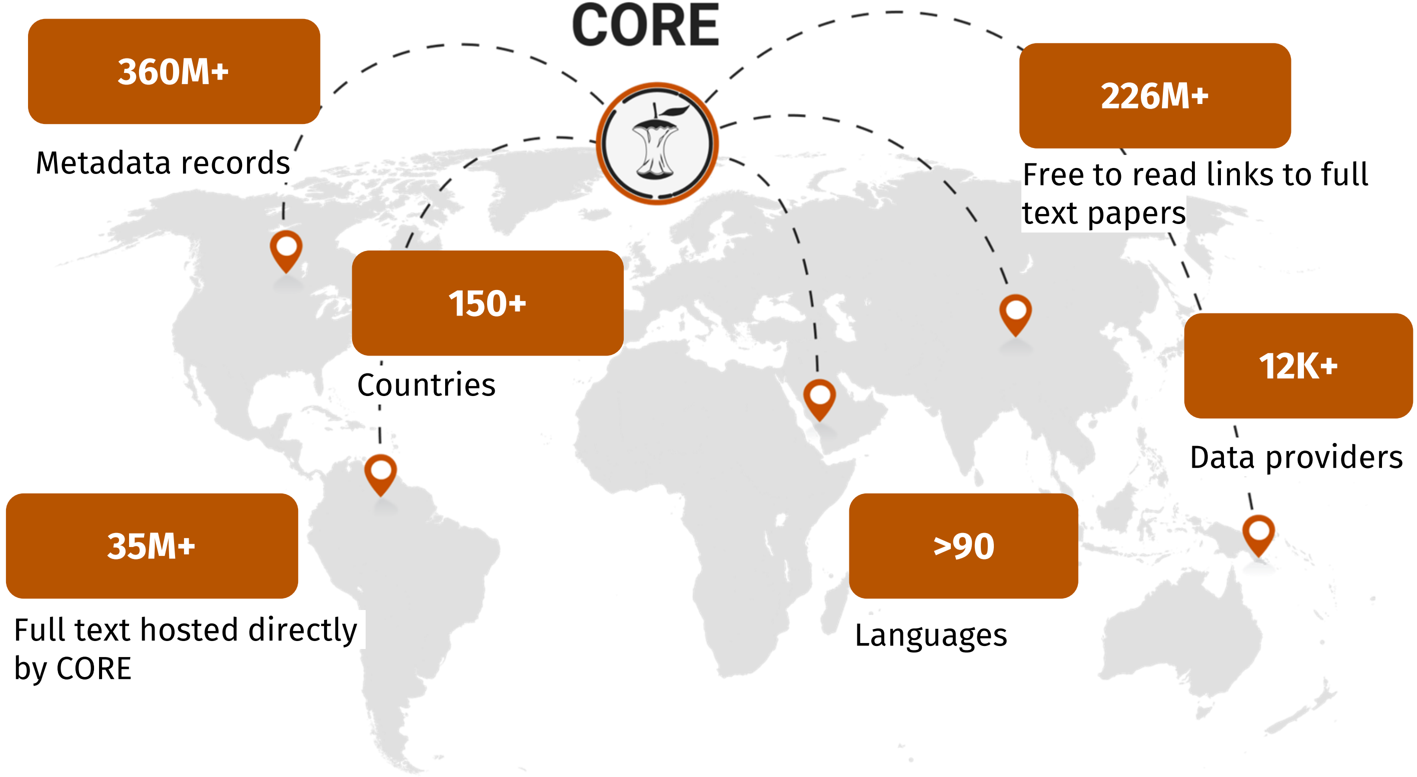
WE

provide solutions for content management, discovery and scalable machine access to research.

WE

serve the global network of repositories and journals by increasing discoverability and reuse of open access content.

CORE



360M+

Metadata records

150+

Countries

35M+

Full text hosted directly by CORE

226M+

Free to read links to full text papers

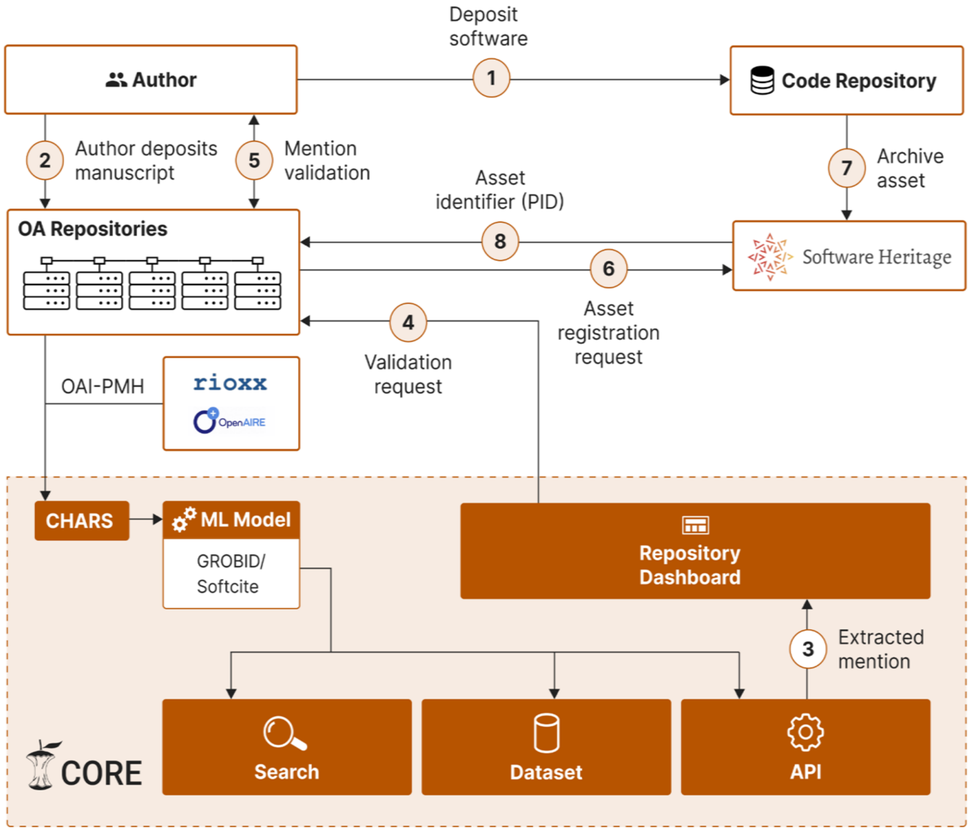
12K+

Data providers

>90

Languages

SoFAIR overall workflow

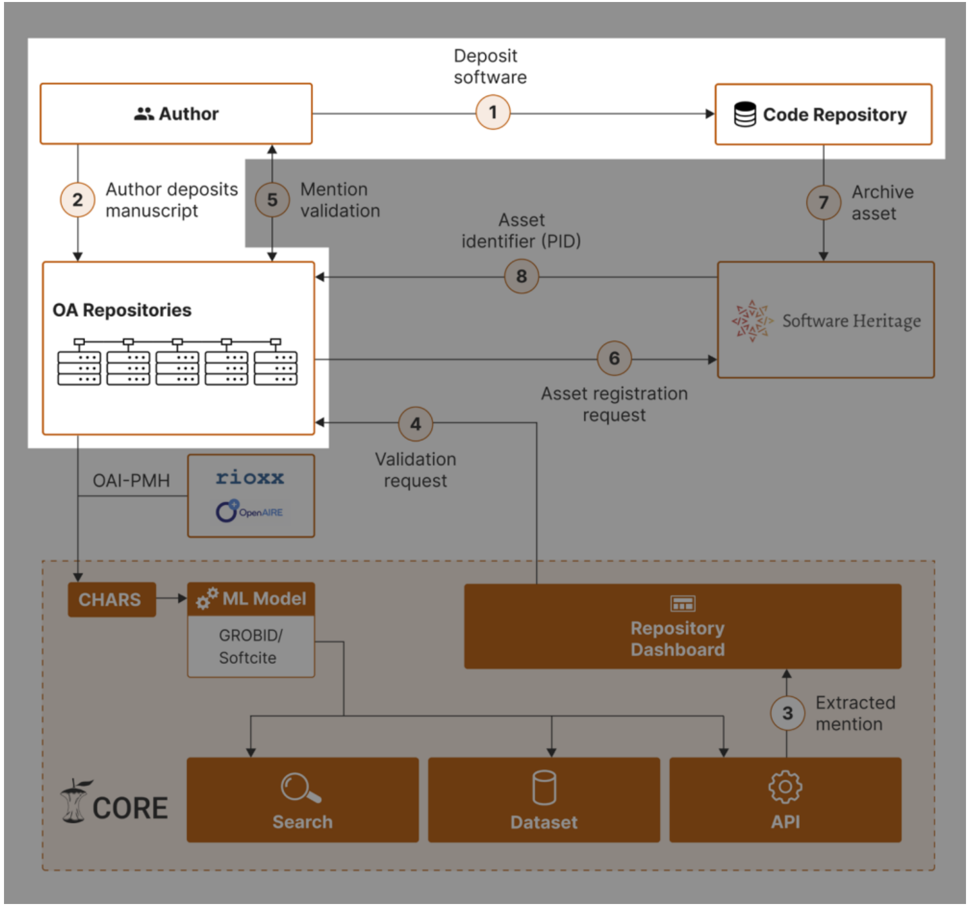


https://sofairoa.github.io/documentation/

SoFAIR

STEP 1-2:

Manuscript deposited by author & Software developed in forge on a code repository

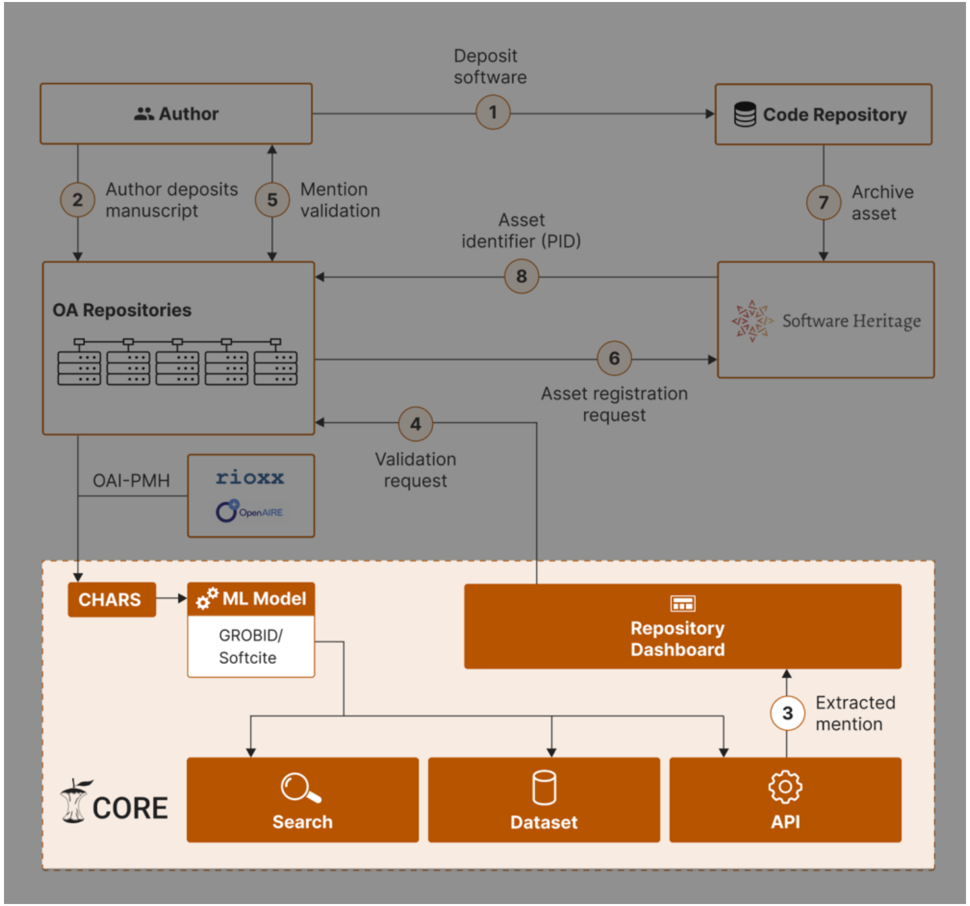


https://sofairoa.github.io/documentation/

SoFAIR

STEP 3:

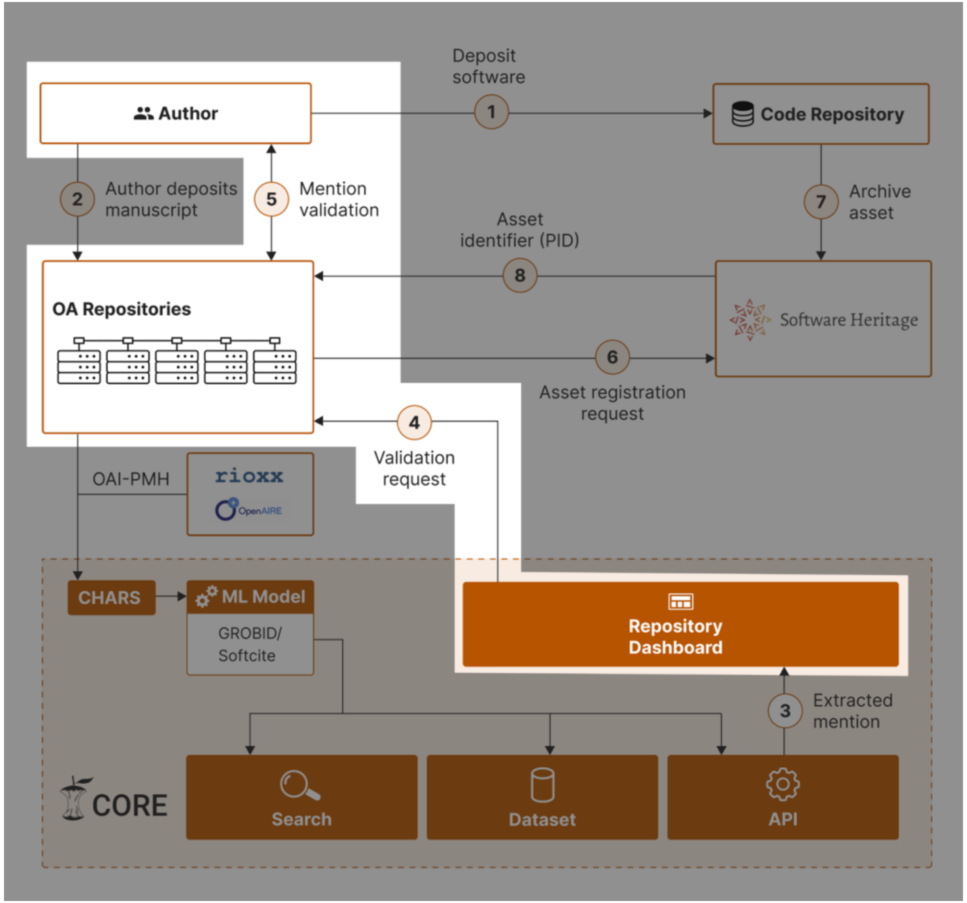
ML models for software extraction and integration in CORE



https://sofairoa.github.io/documentation/

SoFAIR

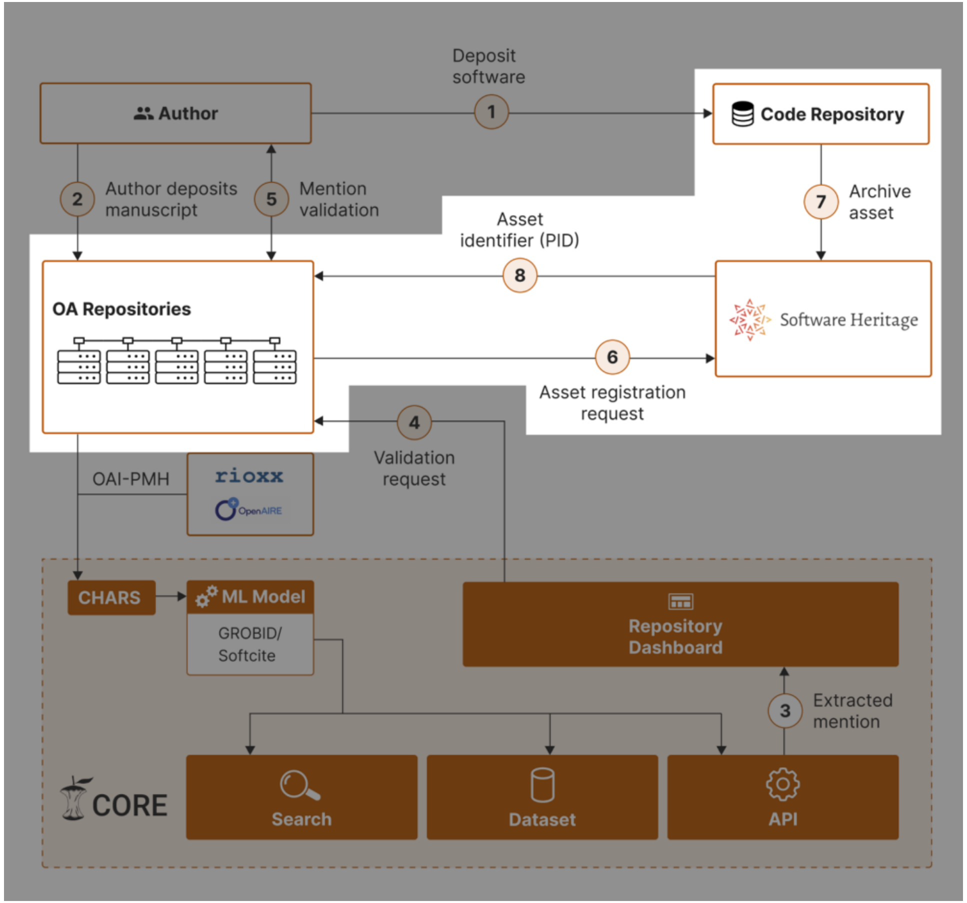
STEPS 4–5:
Validation of extracted software assets



https://sofairoa.github.io/documentation/

SoFAIR

STEPS 6-8:
Registration and archival of software assets



https://sofairoa.github.io/documentation/

SoFAIR

Conclusions

- ➔ Identifying and archiving software assets mentioned in research manuscripts is one of the preconditions for solving the reproducibility crisis.
- ➔ Using AI models to extract software mentions from manuscripts.
- ➔ Building a new SoFAIR workflow that will enable better management of SW assets leveraging CORE and Software Heritage