



Open Science and Artificial Intelligence in Research Software Engineering

Lecturers: Daniel Garijo, Oscar Corcho

**ETSI Informáticos, Ontology Engineering Group,
Universidad Politécnica de Madrid, Spain**

<https://oeg.fi.upm.es/>

Session 1: Introduction*

✉ daniel.garijo@upm.es

🐦 @dgarijov

This work is licensed under the license

CC BY-NC-SA 4.0 International

<http://purl.org/NET/rdflicense/cc-by-nc-sa4.0>



You are free:

- to Share — to copy, distribute and transmit the work
- to Remix — to adapt the work

Under the following conditions

- Non-commercial – You cannot use it for commercial purposes, nor for training inside a commercial company
- Attribution — You must attribute the work by inserting
- “[source <http://www.oeg-upm.net/>]” at the footer of each reused slide
- A credits slide stating: “These slides are partially based on ‘Open Science and Artificial Intelligence in Research Software Engineering: Introduction’ by Daniel Garijo and Oscar Corcho”
- Share-Alike

Be respectful

- Always respect the opinions of others
- In the unlikely case of unrespectful interactions with lecturers or other colleagues (e.g., via chat, recordings appearing in social media without explicit permission, etc.), these will be handled appropriately



Code Of Conduct by [Nick Youngson](https://www.picpedia.org/highway-signs/images/code-of-conduct.jpg)
<https://www.picpedia.org/highway-signs/images/code-of-conduct.jpg>



Daniel Garijo

@dgarijov

daniel.garijo@upm.es

<https://w3id.org/people/dgarijo/>



Oscar Corcho

@ocorcho

ocorcho@fi.upm.es

<https://oeg.fi.upm.es//index.php/es/teachers/11-ocorcho/index.html>

This is a **new** course:

- We will be testing **new materials**
- We will be doing **new exercises**
- We will be experimenting with **state of the art** techniques
 - Natural language processing, knowledge representation, etc.



Therefore:

- Share with us your **main struggles/pain points**
- Ask us if you get **stuck**
 - We may include small additional tutorials!
- Focus on learning **the methods**
 - Technologies change rapidly
 - You need to learn to adapt



- Understand **why you are interested in the course**
- An **overview**:
 - Open Science (OS): FAIR principles
 - Research Software Engineering (RSE)
 - The role of AI in RSE and OS
- Course **structure**
- **Evaluation** method

- Are you **registered** in this course?
- Have you passed the **recommended courses**?
 - Artificial Intelligence
 - Semantic Web / Knowledge Graphs
 - Software Engineering
- Are you familiar with the following **technologies**?
 - GitHub
 - Jupyter Notebooks/Python
 - Docker
 - Apache Spark
 - RDF / RDFLib
- What are you **interested** in the most?
- What would you like to **learn**?

[job] Semantic Software Engineer and Linked Data Backend Developer

Auer, Sören <SRSO=fthn=ZD=tib.eu=Auer@relay.fi.upm.es>

para semantic-web@w3.org, Linking

Dear all,

In the context of the European Data Spaces initiative, the Open Research Knowledge Centre has several openings at TIB – Leibniz Information Centre for Science and Technology (<https://tib.eu>)

- Research Software Engineer: <https://tib.eu/stellenangebot-47-2022-eng>
- Linked Data Backend developer in the field of Research Data Management

Feel free to get in touch with me or the colleagues mentioned in the openings in case

 Research Software Engineers International

Blog

Research Software Engineers

This is the website of the international research software engineering community.

Research Software Engineers are people who combine professional software expertise with a deep understanding of research. They go by various job titles but the term Research Software Engineer is fast gaining international recognition.

On 13th October 2022 we've celebrated the second *International RSE Day!*

[Read more ...](#)

*International RSE Day is the second Thursday of October each year.
The next International RSE Day will be 12th October 2023!*



nature

Explore content

About the journal

Publish with us

Subscribe

nature > career q&a > article

<https://www.nature.com/articles/d41586-022-01516-2>

CAREER Q&A | 31 May 2022

Why science needs more research software engineers

Ten years after their profession got its name, research software engineers seek to swell



- Understand **why** you are interested in the course
- **An overview:**
 - Open Science (OS): FAIR principles
 - Research Software Engineering (RSE)
 - Artificial Intelligence, Research Software and Open Science
- Course **structure**
- **Evaluation** method

Changes towards **Open Science***

*with slides from ["The Scientific Paper of the Future" training materials for the OntoSoft project](#) (Yolanda Gil et al)

Open Data



Open Licenses



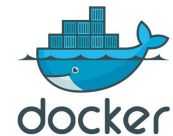
Open Code



Open Models



Open Containers



NATURE|METRICS SURVEY 2010

METRICS SURVEY RESULTS

Thinking about all of the possible measures of scientific contribution that are possible, please select your top 5 priorities.

	No. of times chosen	Relative ranking
Publication in high-impact journals	92	2.61
Grants earned	65	1.73
Training and mentoring students and postdocs	63	1.71
No. of citations on published research	58	1.62
No. of publications	53	1.38
Teaching courses	41	1.18
Collaborative work outside of your department/institution	37	0.97
Development of research resources for the scientific community	31	0.89

The outputs from scientific research are many and varied, including: research articles reporting new knowledge, data, reagents, and software; intellectual property; and highly trained young scientists. Funding agencies, institutions that employ scientists, and scientists themselves, all have a desire, and need, to assess the quality and impact of scientific outputs. It is thus imperative that scientific output is measured accurately and evaluated wisely. (San Francisco Declaration on Research Assessment, <https://sfedora.org/read/>)





Impactstory

Impact



usage

downloads
views



peer-review

expert opinion



citations



alt-metrics

storage
links
bookmarks
conversations

nature research

Data availability statements and data citations policy: guidance for authors

Policy summary

All manuscripts reporting original research must include a data availability statement. Authors are also encouraged to include formal data

nature.com > scientific data

SCIENTIFIC DATA



nature.com



PLOS ONE

Availability of Software

PLOS supports the development of open source software and believes that, for submissions appropriate open source standards will ensure that the submission conforms to (1) our requirement that another researcher can reproduce the experiments described, (2) our aim to promote open science, and (3) that PLOS journals can be built upon by future researchers. Therefore, if new software or a new analysis pipeline that the software conforms to the [Open Source Definition](#), have deposited the following three items as Supporting Information:

- **The associated source code of the software described by the paper.** This should be licensed under a suitable license such as BSD, LGPL, or MIT (see <http://www.copdess.org>). Commercial software such as Mathematica and MATLAB does not preclude a paper from being open source.
- **Documentation for running and installing the software.** For end-user applications this is a prerequisite; for software libraries, instructions for using the application program interface are sufficient.
- **A test dataset with associated control parameter settings.** Where feasible, results from running the test data should not have any dependencies — for example, a database dump.

Acceptable archives should provide a public repository of the described software. The code should be accessible for creating user accounts, logging in or otherwise registering personal details. The repository should host more than 1,000 projects. Examples of such archives are: [SourceForge](#), [Bioinformatics.Org](#), [Savannah](#), [GitHub](#) and the [Codehaus](#). Authors should provide a direct link to the deposited software.

COPDESS

Coalition on Publishing Data
in the Earth and Space
Sciences

COPDESS Suggested Author Instructions and Best Practices for Journals

The Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) develops and recommends best practices for journal author instructions around data and identifiers as a resource to the community. These best practices are consistent with and based on the COPDESS Statement of Commitment and have been developed with guidance from participants in COPDESS.

[Data Policy Statement](#)

[Data Citation](#)

[Sample Citation and Identification](#)

[Crossref Funder Registry](#)

[ORCIDs](#)

[Presentations on Best Practices](#)



EOSC

What the European Open Science Cloud is

The ambition of the European Open Science Cloud (EOSC) is to develop “Web of FAIR Data and services’ for science in Europe. EOSC will be a multi-disciplinary environment where researchers can publish, find and re-use data, tools and services, enabling them to better conduct their work.

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren *JPH*
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles


The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;

Increasing interest in DOING science



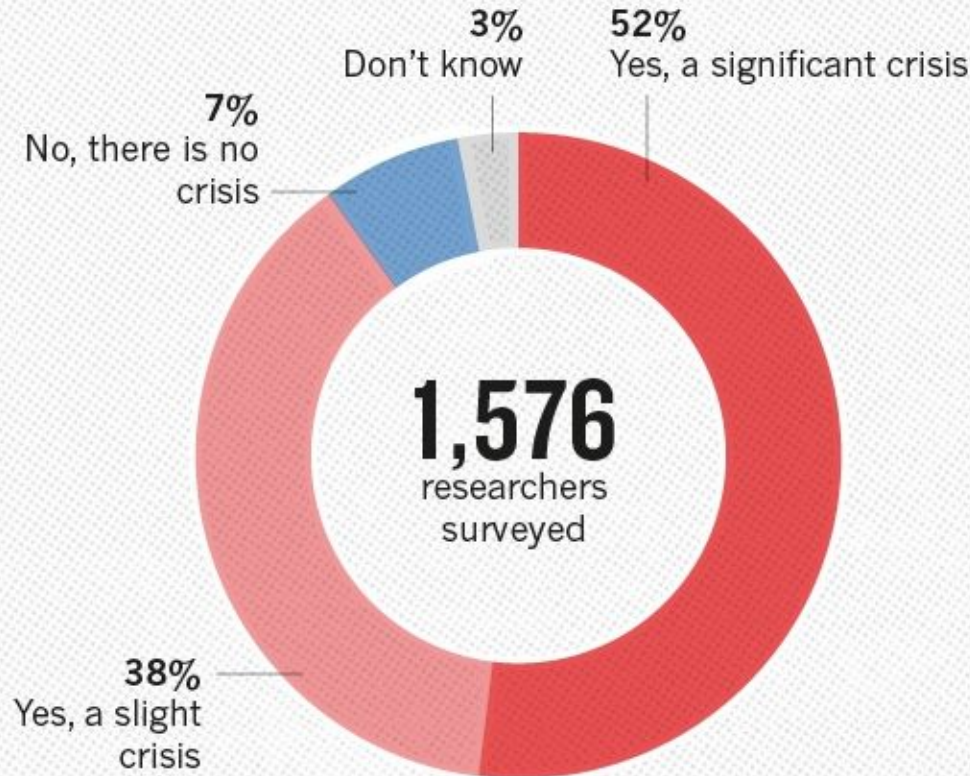
Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach

Richard A. Rocca , Gregory Magoon, David F. Reynolds, Thomas Krahn, Vincent O. Tilroe, Peter M. Op den Velde Boots, Andrew J. Grierson



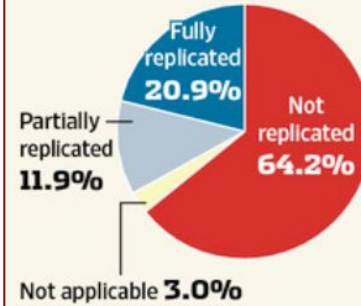
Published: July 24, 2012 • DOI: 10.1371/journal.pone.0041634

IS THERE A REPRODUCIBILITY CRISIS?



No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.



Source: Nature Reviews Drug Discovery



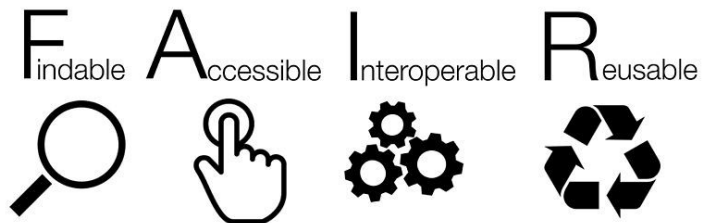
Retraction Watch

Tracking retractions as a window into the scientific process



<https://www.force11.org/group/fairgroup/fairprinciples>

<https://doi.org/10.1038/sdata.2016.18>



Implementing FAIR Data Principles: The Role of Libraries

What are the FAIR Data Principles?

The FAIR Data Principles are a set of guiding principles in order to make data findable, accessible, interoperable and reusable (FAIR). In short, these principles provide guidance for scientific data management and research and are relevant to all stakeholders in the current digital ecosystem. They directly address data producers and data publishers to promote maximum use of research data. Research libraries can use the FAIR Data Principles as a framework for leading and extending research data services.

What is FAIR DATA?



To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

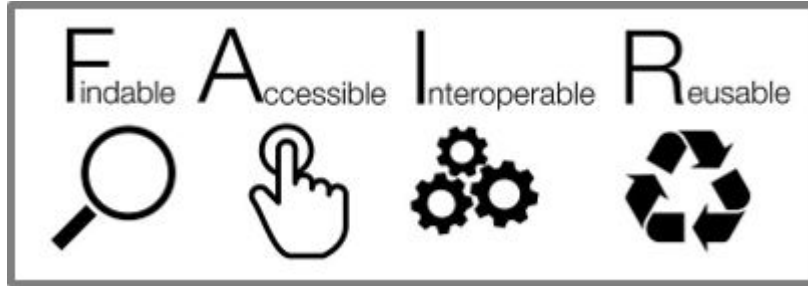
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.



Data (initially) [1]



Research Software (FAIR4RS)



Methods (FAIR Wfs)



Semantic artefacts

Other guidelines:

- Guidelines for Transparency and Openness Promotion (TOP) [2]
- Reproducibility Enhancement Principles (REP) [3]
- ...



[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

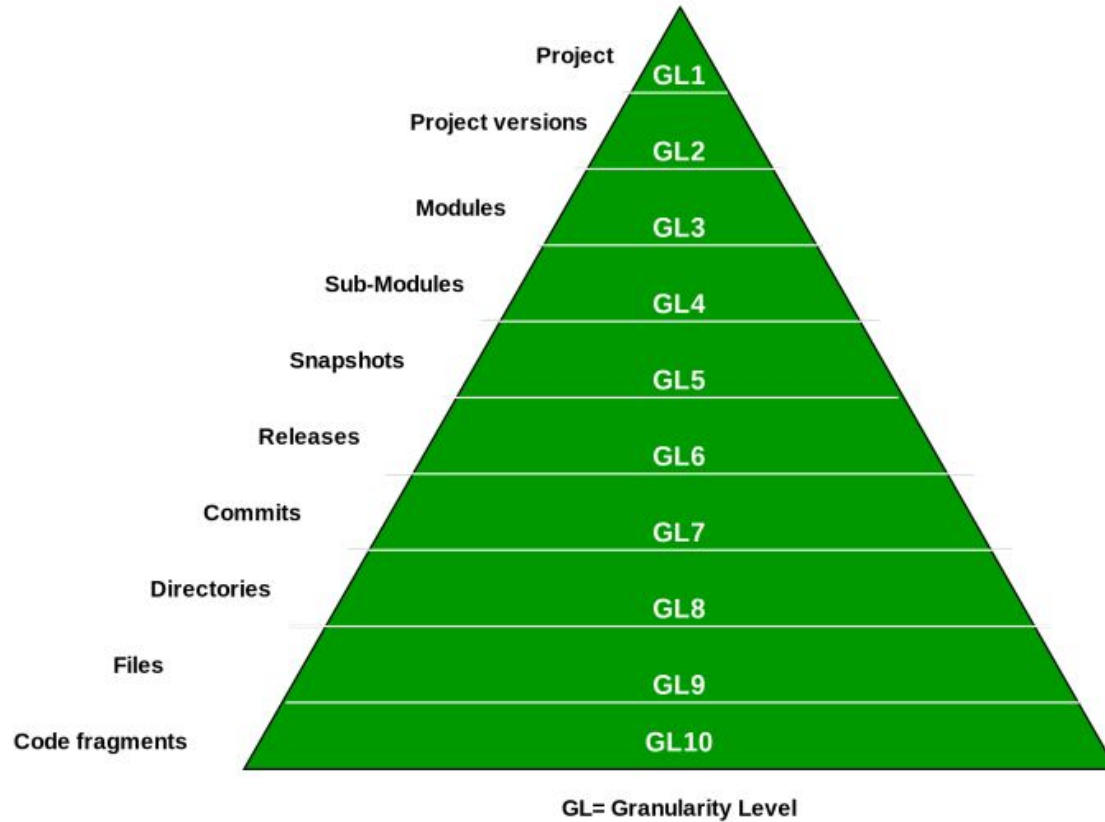
[2] <https://www.cos.io/initiatives/top-guidelines>

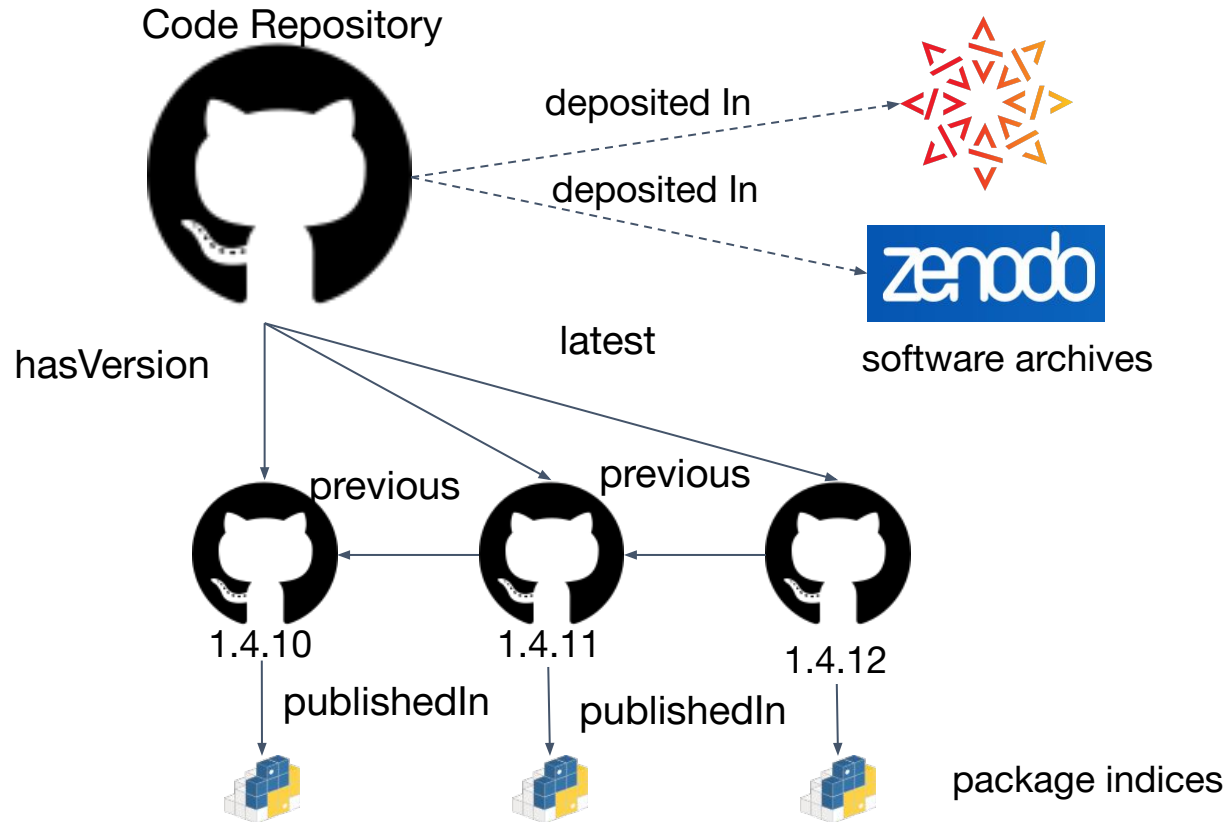
[3] Stodden, V *et al* Enhancing reproducibility for computational methods
<https://www.science.org/lookup/doi/10.1126/science.aah6168>

Defining **Research Software**

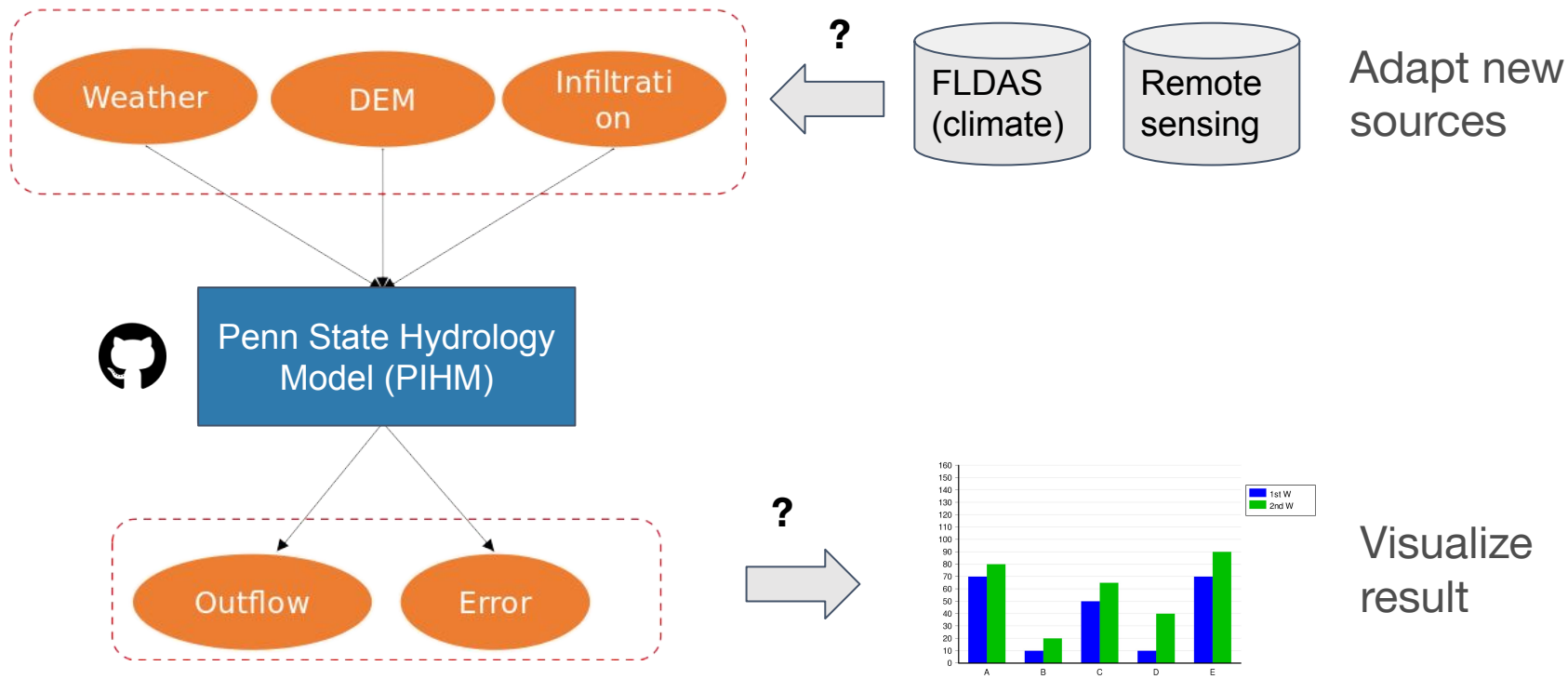
Research Software includes source code files, algorithms, scripts, computational workflows and executables that were created during the research process or for a research purpose. Software components (e.g., operating systems, libraries, dependencies, packages, scripts, etc.) that are used for research but were not created during or with a clear research intent should be considered software in research and not Research Software. This differentiation may vary between disciplines.

Article: Chue Hong, Neil P., Katz, Daniel S., Barker, Michelle, Lamprecht, Anna-Lena, Martinez, Carlos, Psomopoulos, Fotis E., Harrow, Jen, Castro, Leyla Jael, Gruenpeter, Morane, Martinez, Paula Andrea, Honeyman, Tom, Struck, Alexander, Lee, Allen, Loewe, Axel, van Werkhoven, Ben, Jones, Catherine, Garijo, Daniel, Plomp, Esther, Genova, Francoise, ... RDA FAIR4RS WG. (2022). FAIR Principles for Research Software (FAIR4RS Principles) (1.0). <https://doi.org/10.15497/RDA00068>





Research software associated resources: Inputs and outputs



Software images are created from configuration files (e.g., Dockerfiles)

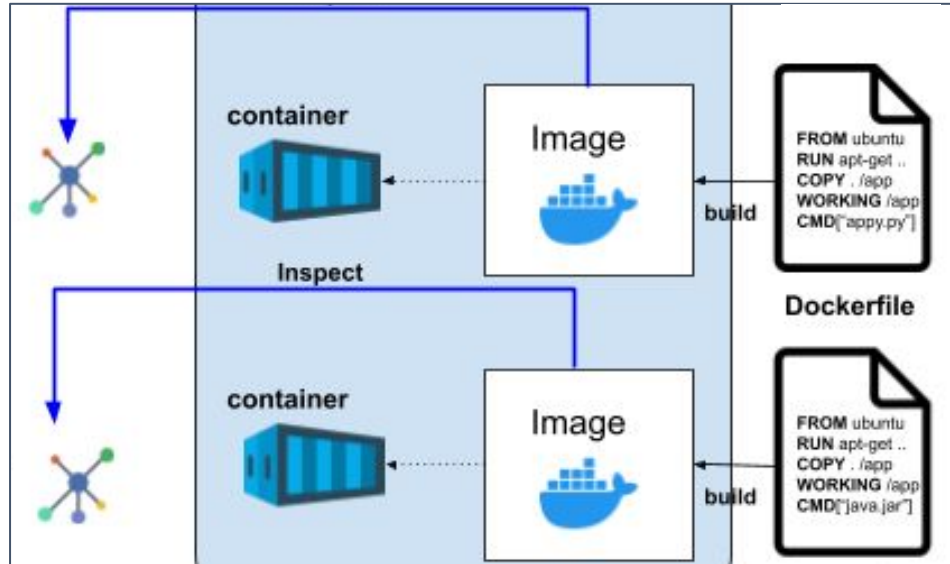


Fig. by Jhon Toledo

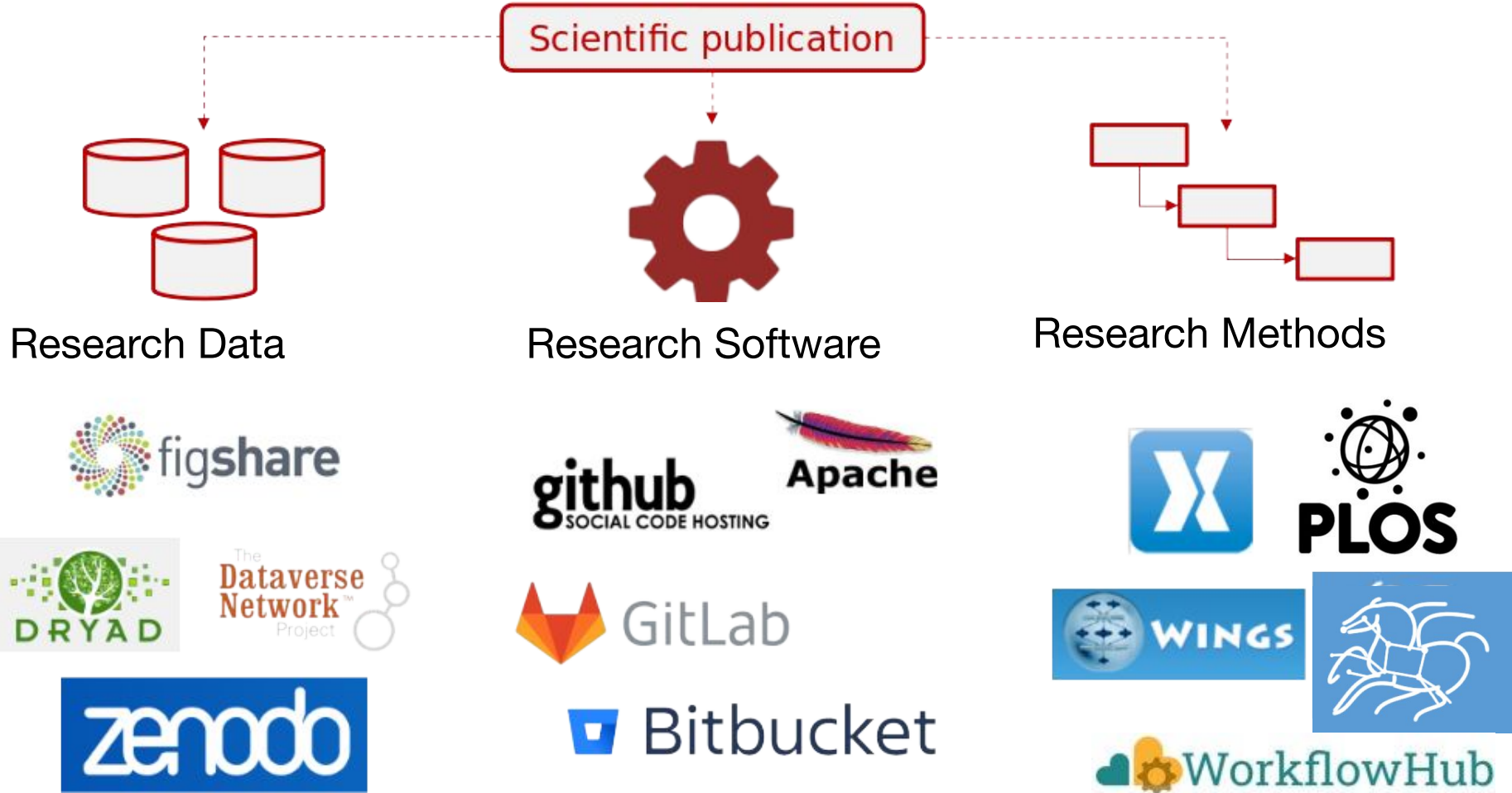
Initial effort transforming part of DockerHub: <https://dockerpedia.inf.utfsm.cl/>

Research Software and **Open Science**

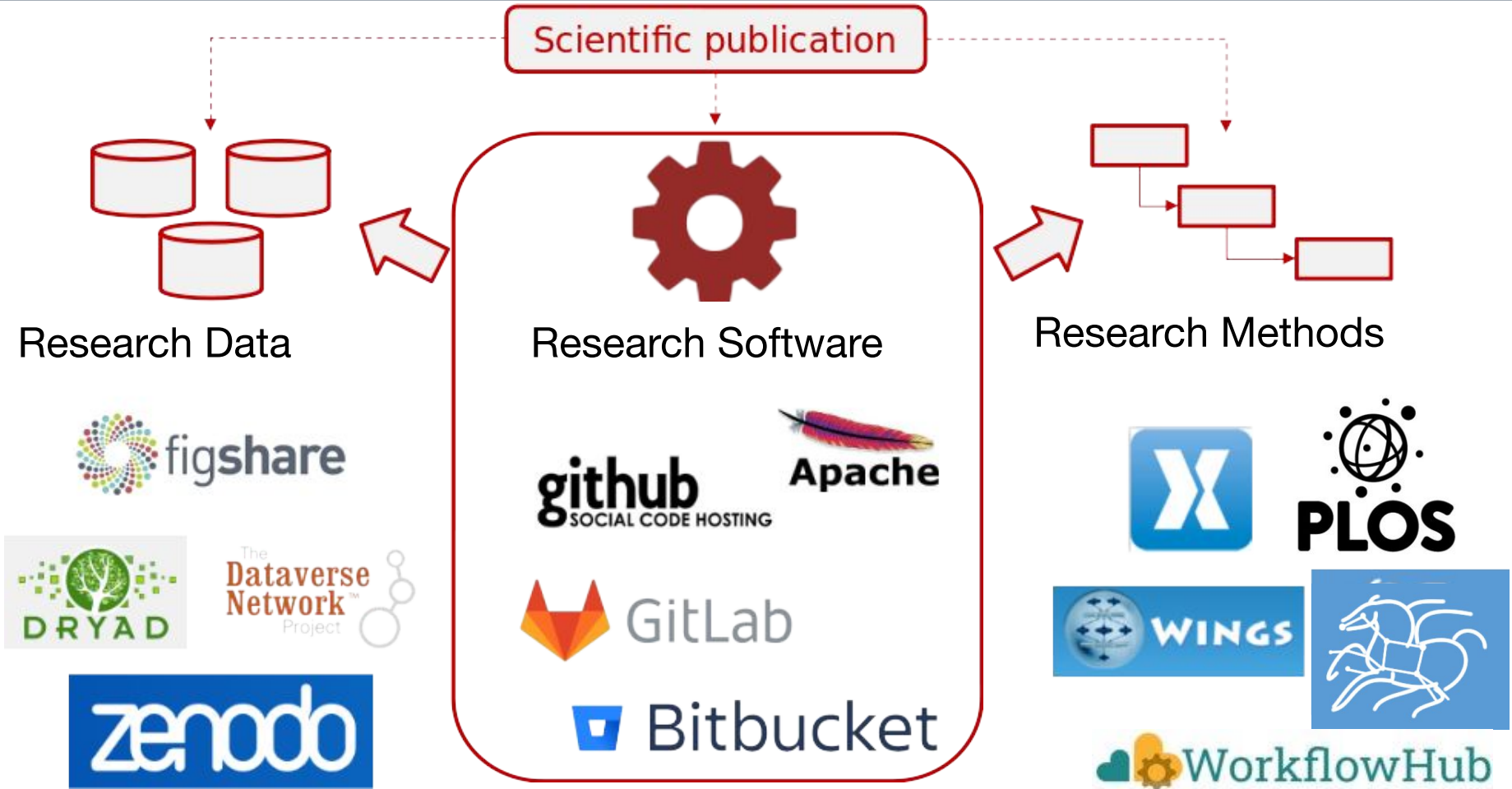
Research Software is one of the pillars of Open Science

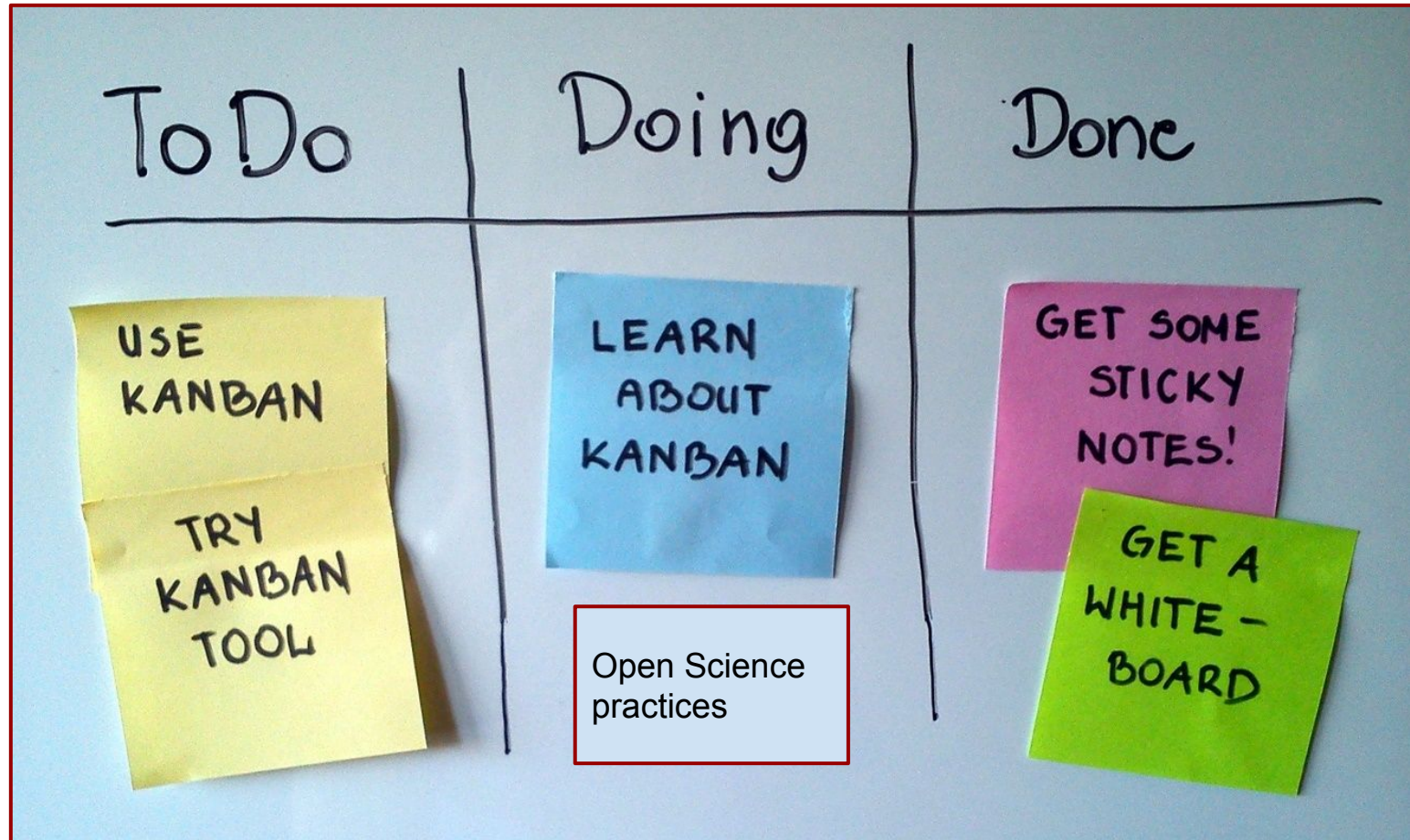


Research Software is one of the pillars of Open Science



Research Software is one of the pillars of Open Science





Artificial Intelligence **for Open Science and** **Research Software Engineering**

FAIR enables reusability and **machine-readability**. Some benefits include:

- Impact
- AI can train more efficiently
- Credit
- Reproducibility

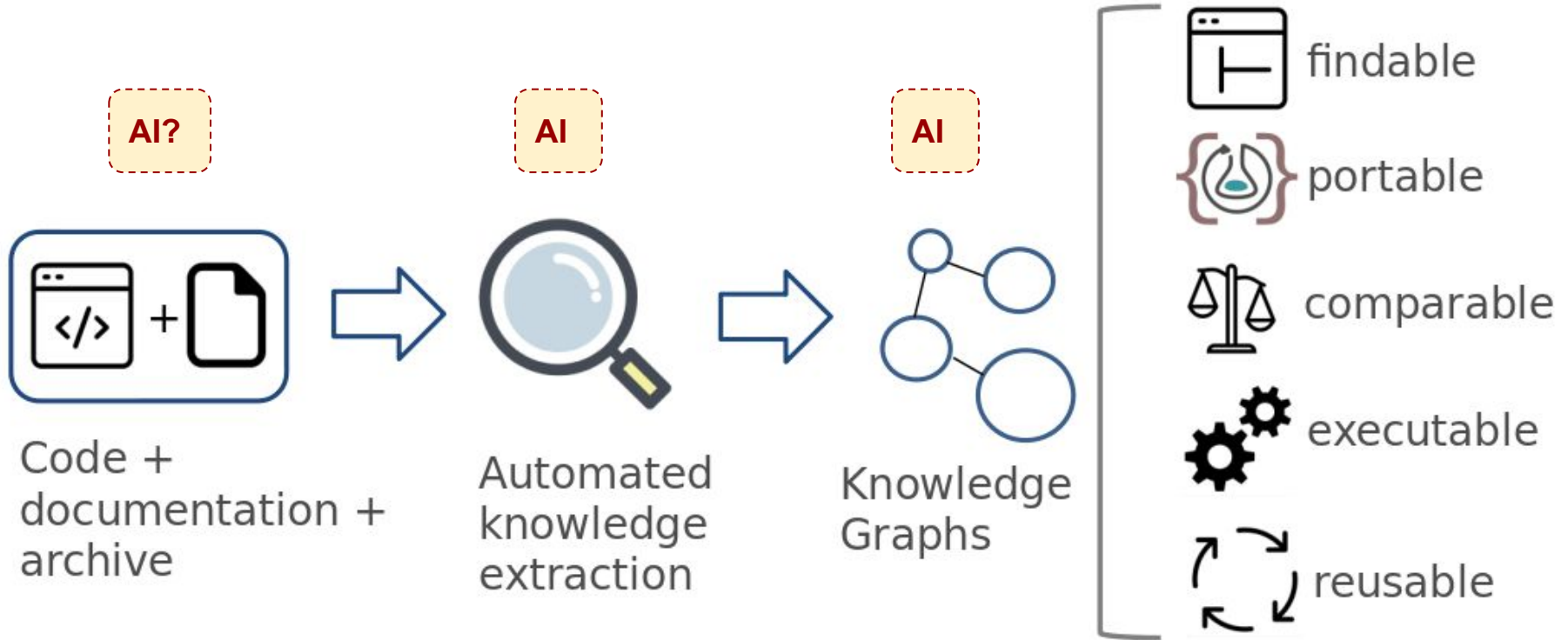
Making things FAIR may be **time consuming**

- Using AI techniques for assistance
- Enhance existing collaborative datasets

Expected **improvements**:

- Query, link
- Compare, inspect, scan
- Ease reuse
- Assist





- Understand **why you are interested in the course**
- An **overview**:
 - Open Science (OS): FAIR principles
 - Research Software Engineering (RSE)
 - Artificial Intelligence, Research Software and Open Science
- **Course structure**
- **Evaluation** method

1. Introduction
 - 1.1. Motivación: Reproducibilidad y los principios FAIR para datos y software científicos / Motivation: Reproducibility and the FAIR principles for research data and software
 - 1.2. Software y preservación de datos: repositorios de software y registros de metadatos/ Software and data conservation: code repositories and metadata registries
 - 1.3. Iniciativas abiertas para gestionar datos y software / Open initiatives for managing Data and Software
 - 1.4. Descripción de datos y software para reproducibilidad y reutilización / Describing data and software for reproducibility and reuse
2. Métodos computacionales / Computational scientific methods
 - 2.1. Notebooks / Computational notebooks
 - 2.2. Infraestructura computacional y contenedores de software / Computational infrastructure and software containers
 - 2.3. Flujos de trabajo científicos / Scientific workflows
 - 2.4. Experimentos a gran escala: planificación y paralelización / Large-scale experiments: planning and parallelization
 - 2.5. Composición de flujos de trabajo y razonamiento / Workflow composition and reasoning
 - 2.6. Aprendiendo a manejar Infraestructuras abiertas de investigación / Getting started with open research infrastructures.
3. Provenance en Ingeniería de Software de Investigación / Provenance in Research Software Engineering
 - 3.1. Introducción a provenance / Introduction to provenance
 - 3.2. El estándar W3C PROV / The W3C PROV standard
 - 3.3. Repositorios abiertos para recolectar provenance / Open repositories for collecting provenance
 - 3.4. Aplicaciones para capturar y usar provenance en la Web / Applications for capturing and exploiting
4. Grafos de Conocimiento Científicos / Scientific Knowledge Graphs provenance in the Web.
 - 4.1. Representación de Datos y software procesable por máquinas / Machine-readable data and software representation
 - 4.2. Capturando el contexto de un experimento: Objetos de Investigación / Capturing the context of research: Research Objects
 - 4.3. Aplicaciones y ejemplos sobre grafos de Conocimiento Científicos / Applications for Scientific Knowledge Graphs
5. Aprendizaje automático en Ingeniería del Software de Investigación / Machine Learning for Research Software Engineering
 - 5.1. Clustering y clasificación de software científico y sus metadatos / Clustering and classification of research software and its metadata
 - 5.2. Reconocimiento de Entidades en software científico / Named entity recognition in research software
 - 5.3. Aplicaciones de aprendizaje automático en Ingeniería del Software de Investigación / Applications of machine learning models in Research Software Engineering

Wednesdays, 5-7 pm (14 sessions + exam)

Calendar is available on the web:

[https://www.etsiinf.upm.es/docs/estudios/calendario/55_00%20borrador
calendario_grado_2022_23%20semestre%20par_v0.pdf](https://www.etsiinf.upm.es/docs/estudios/calendario/55_00%20borrador_calendario_grado_2022_23%20semestre%20par_v0.pdf)

This course is heavily based on practical exercises

- An **individual** assignment (20%)
 - Domain: text analysis
- A **group** assignment (35%)
 - Assignment (25%)
 - Presentation (10%)
 - Domain: knowledge extraction and representation
- A **written** test (45%)
 - Assignments must be passed in order to attend the final test
 - Based heavily on the methods exercised in the assignments

Lecturing team

- Oscar Corcho (coordinator and lecturer)
- Daniel Garijo (teaching assistant)

Online platforms

Participation in online forums will be positively evaluated

Moodle (**work in progress**):

<https://moodle.upm.es/titulaciones/oficiales/course/view.php?id=21504#section-0>

Start preparing for the first task (text extraction and analysis)

- Create an online GitHub repository for project deliverables
- Add an Open license
- Add a readme file
- Familiarize yourself with Grobid (<https://github.com/kermitt2/grobid>)
 - Text extraction tool from PDF
- Select up to 10 papers as your input dataset

Deadline **8th Feb, 2023**



Open Science and Artificial Intelligence in Research Software Engineering

Lecturers: Daniel Garijo, Oscar Corcho

**ETSI Informáticos, Ontology Engineering Group,
Universidad Politécnica de Madrid, Spain**

<https://oeg.fi.upm.es/>

Session 1: Introduction*

✉ daniel.garijo@upm.es

🐦 @dgarijov