

Introduction: Open Science and FAIR in Software Engineering and Requirements Engineering

Oliver Karras, Alessio Ferrari, Davide Fucci, and Davide Dell'Anna

oliver.karras@tib.eu, alessio.ferrari@isti.cnr.it, davide.fucci@bth.se, d.dellanna@uu.nl

32nd IEEE International Requirements Engineering 2024 Conference – Exploring New Horizons: Expanding the Frontiers of Requirements Engineering

June 24th, 2024, Reykjavik, Iceland

What is Open Science?

Open science is the movement to make scientific research (including publications, data, physical samples, and software) and its dissemination **accessible** to all levels of society, amateur or professional.

- Open methodology
- Open source
- Open data
- Open access
- Open peer review (identities, reports, community participation)
- Open educational resources



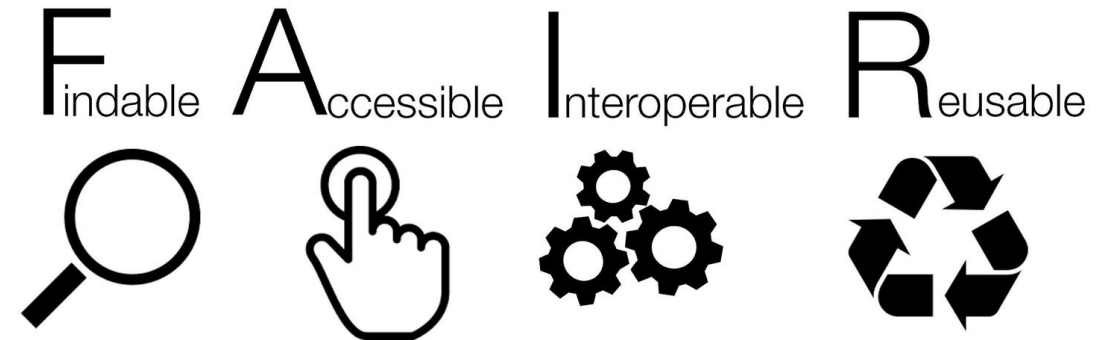
What are the FAIR Principles?

Scientists increasingly rely on computational support to deal with increasing amounts of data.

Focus on **Digital** Assets (Data & Knowledge)

FAIR Principles^[2]:

- **F**indability: Rich metadata, reachable by humans and machines, unique identifiers (DOI)
- **A**ccessibility: Open protocols for data access
- **I**nteroperability: Integration with other data
- **R**euse: Replicable experiments, understandable data (through metadata)



Why Do We Need Open Science and FAIR?

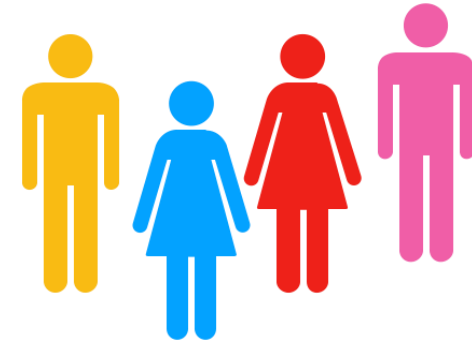
1. Transparency and Reproducibility:

- Open science helps verify the validity of research results
- Scientists can reproduce experiments, which is crucial for the **credibility**



2. Collaboration and Innovation:

- Collaboration among scientists, and **interdisciplinary** communication
- Accelerate discoveries and novel **ideas**



3. Equity and Inclusion:

- Democratize access to knowledge, and enrich viewpoints through **global access**
- Science accessible to the general public can increase scientific literacy, leading to an **informed society**



Current Initiatives in Software Engineering

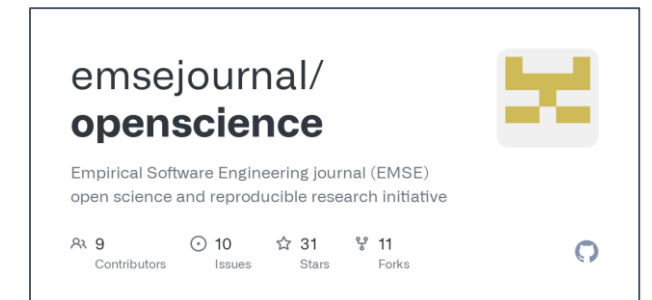
Emphasis on Data, **Tools**, Research Results



The ACM badge system (2010)



Artifact Evaluation Tracks



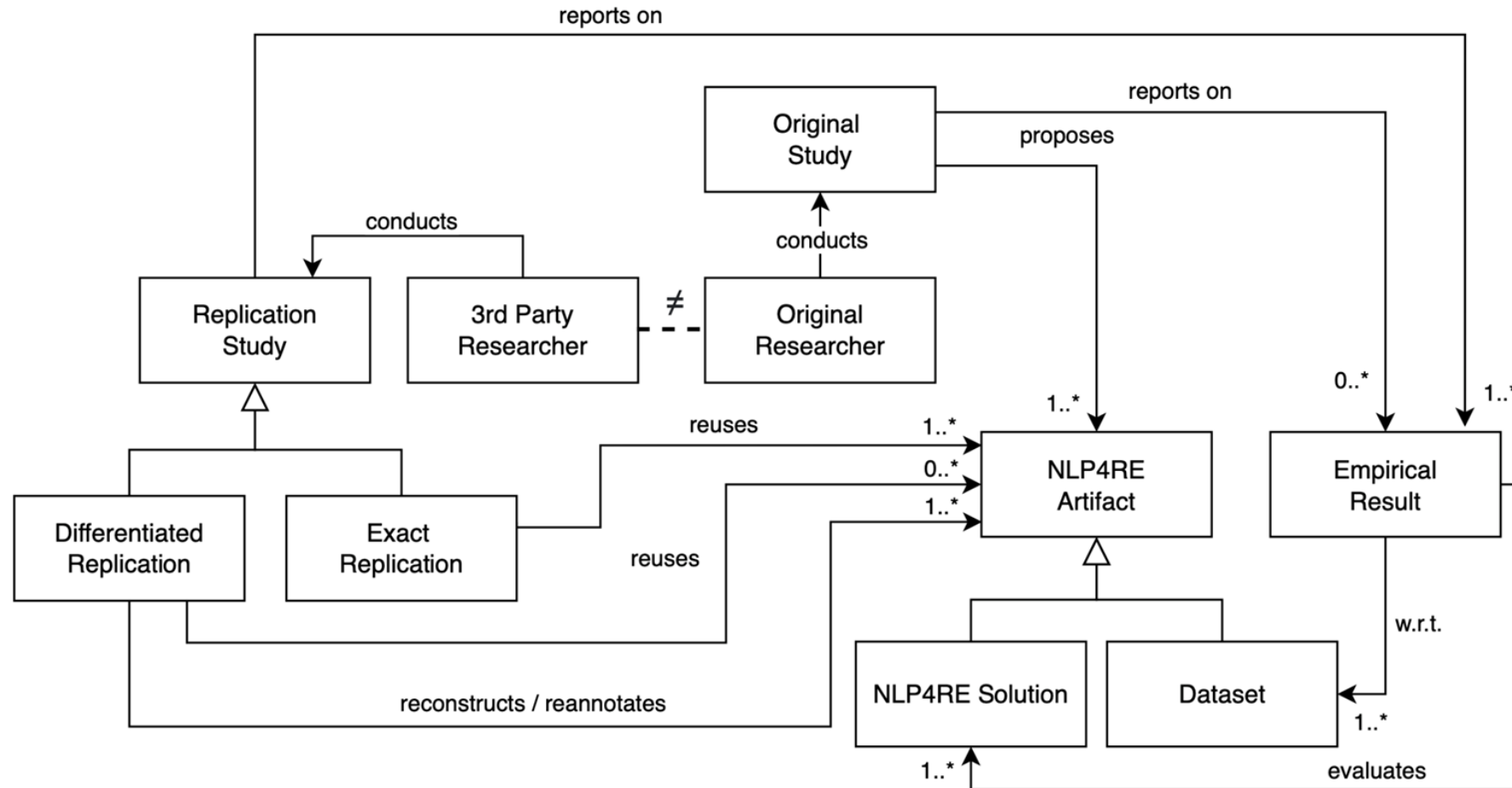
Home > ACM Journals > ACM Transactions on Software Engineering and Methodology > CFP

Sections

Replicated Computational Results (RCR) Report

Journal Initiatives

Replication and Reuse: Terminology



Example Problems in RE: Replication and Tool Reconstruction

ID	Description
Rec1	The reconstruction-relevant information and implementation details of the original approach can be ambiguous, imprecise, and incomplete.
Rec2	If a tool was partially or fully developed and/or evaluated using proprietary data, then there is no guarantee that the reconstructed tool would be identical to the original one since the used data cannot be accessed for reconstruction purposes.
Rec3	Communication with the original authors is not always useful since the actual information sources may not be available anymore.
Rec4	The continuous evolution of the NLP ecosystem entails that some libraries become outdated, unavailable, or not maintained anymore.
Rec5	Tools are typically developed as prototypes and not maintained in the long term.
Rec6	Tool reconstruction is not (yet) valued as a self-standing research contribution, and hence researchers are discouraged to replicate tools overtime.

Poor information

Proprietary data

Communication

Updates and
maintenance

Poor value

Example Problems in RE: Dataset Annotation

ID	Description
Ann1	Some RE-specific categorization tasks lack solid theories that can guide the annotation process.
Ann2	Besides annotation experience and theoretical knowledge, the lack of domain knowledge can limit the accuracy of the annotations.
Ann3	The annotation activity is time consuming due to factors such as language barriers, different individuals' background, and fatigue.
Ann4	The annotation protocol can evolve and thereby necessitate the re-annotation of the data which might, again, cause additional time and effort.
Ann5	Theoretical and practical training resources and opportunities are limited and not adequate for training novice annotators who are often trained during the annotation task by more experienced annotators.
Ann6	The lack of benchmarks entails that annotated datasets enabling comparison against the state-of-the-art are scarce.
Ann7	Available imbalanced datasets pose the challenge of both understanding the minority class and consequently the annotation of new examples thereof.
Ann8	Determining the right amount of context to be shared alongside the annotation raw data is essential and can significantly affect the annotation results.
Ann9	Motivating the annotators poses another challenge since an immediate observation of the impact of a given annotation task is not always possible.
Ann10	Annotators are often not experienced in managing the social aspects or resolving conflicts originating from power, authority, or other social relations.

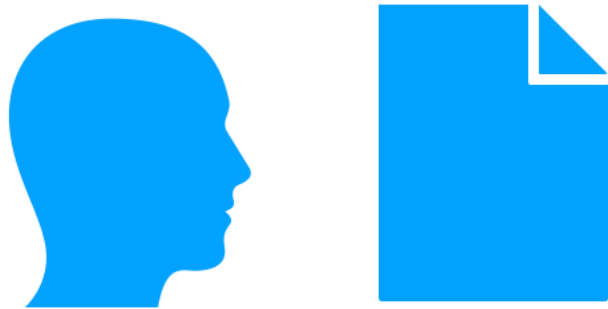
Domain Knowledge

Guidelines

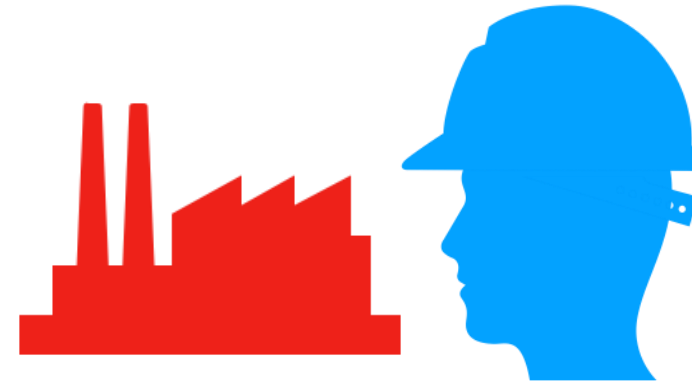
Data Quality

Lack of Benchmarks

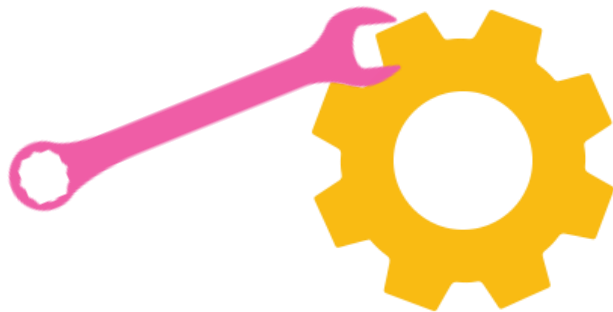
What do We Need?



Clear papers, Clear Communication



Involvement of companies and experts



Tool maintenance



Reward for open science and replication

Thanks to...



Sallam Abualhaija



Fatma Basak Aydemir



Fabiano Dalpiaz



Davide Dell'Anna



Xavier Franch



Davide Fucci