Do Planeamento à Partilha:
Boas Práticas para Criar Datasets de Investigação com Impacto

Open Science Festival, 20.oct.2026

Ivo Silva Prof. Auxiliar (UMinho), Investigador (Centro ALGORITMI)

Cristiano Pendão Prof. Auxiliar (UTAD), Investigador (Centro ALGORITMI)

cristianop@utad.pt; ivo@di.uminho.pt



Universidade do Minho

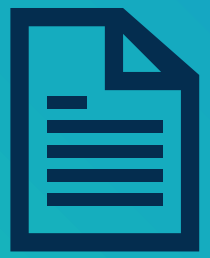


Outline

1. Why open datasets matter?
2. The current state of our research area
3. Guidelines for collecting a dataset
    1. Before (data collection)
    2. During (data collection)
    3. After (data collection)
4. Key Takeaways

Open Science
Beyond Open Access

OPEN ACCESS
Access to the paper
Read the findings and conclusions



OPEN DATA
Access to the evidence
Inspect and reuse the research data



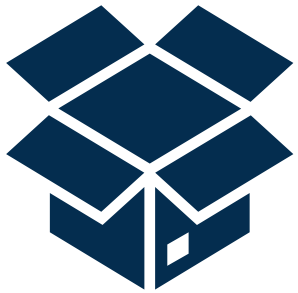
OPEN CODE
Access to the implementation
Understand and repeat the analysis



Open science extends openness to the research process.

Open Science

Why open datasets matter?



ACCESSIBILITY

- Open data
- Open methods (code)



REPRODUCIBILITY

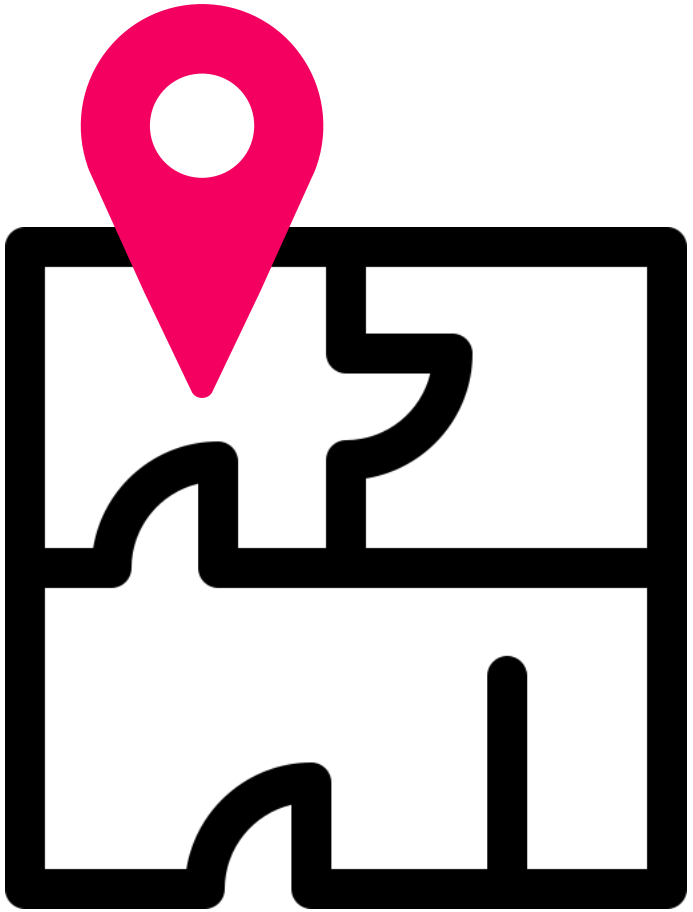
- Verify experiments
- Others can reproduce results



BENCHMARKING

- Comparison between systems

We first surveyed our research area:
Indoor Positioning and Localization



We first surveyed our research area:

Open Research Data in Indoor Positioning (ORDIP)

Discover Open Datasets
- IPIN conference papers
- Zenodo & Scopus
- Other sources

Table of Open Datasets
- Technologies & Signals
- Environments & Metadata
- Collection & Usage



Best Practices
- How to collect data
- How to share data
- How to reuse data

Public discussion of standard
- Sharing the results
- Identifying good/bad practices

https://www.sciencedirect.com/science/article/pii/S2542660524004268

Open Research Data in Indoor Positioning (ORDIP)

119
DATASETS

Analysed by 2 reviewers

~40%
NO PERSISTENT ID

Harder to find, cite, and reuse

44%
WITH DATA LOADING CODE

More difficult to reproduce without the code

35%
INCLUDE FLOOR PLANS

Missing context and metadata for reuse

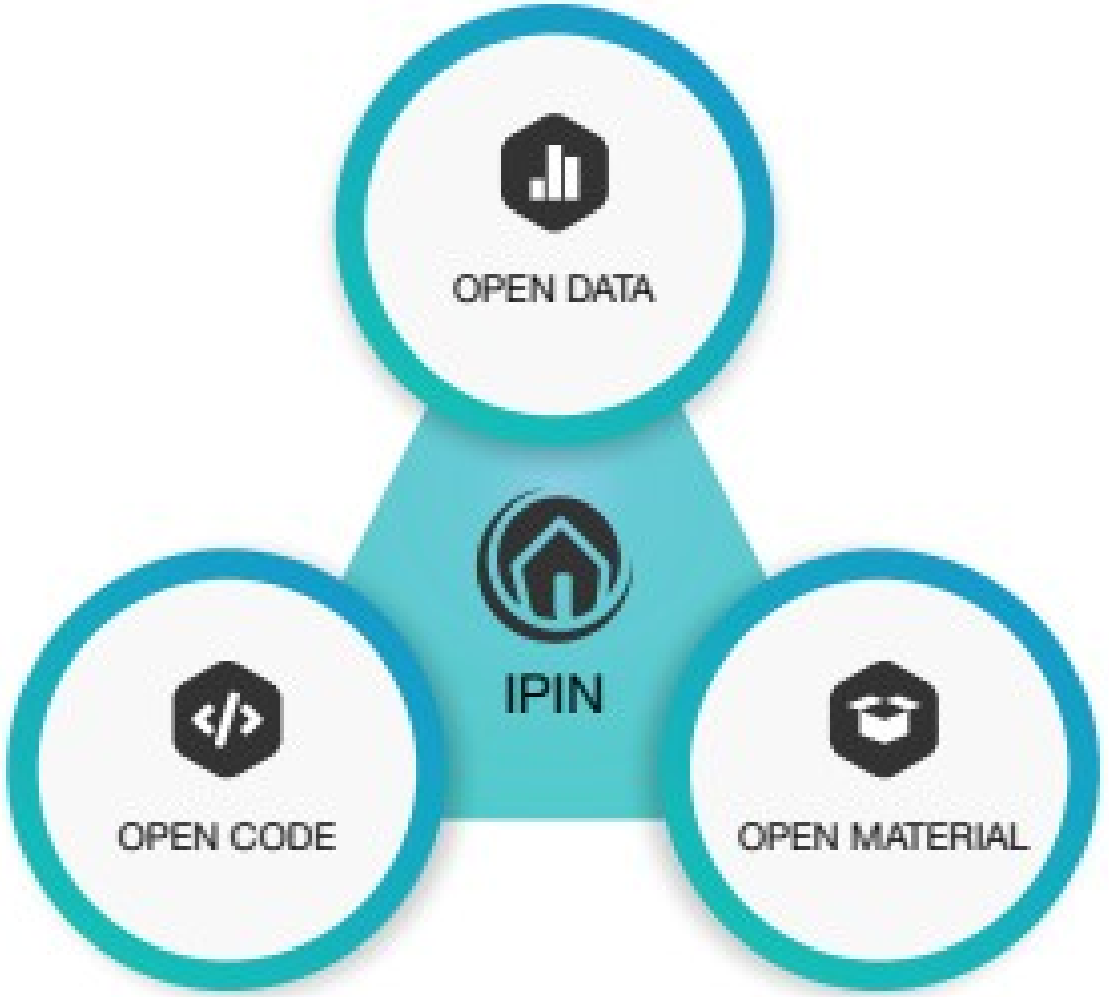
~46%
HAVE SEPARATE TRAINING SET

Essential for some methods

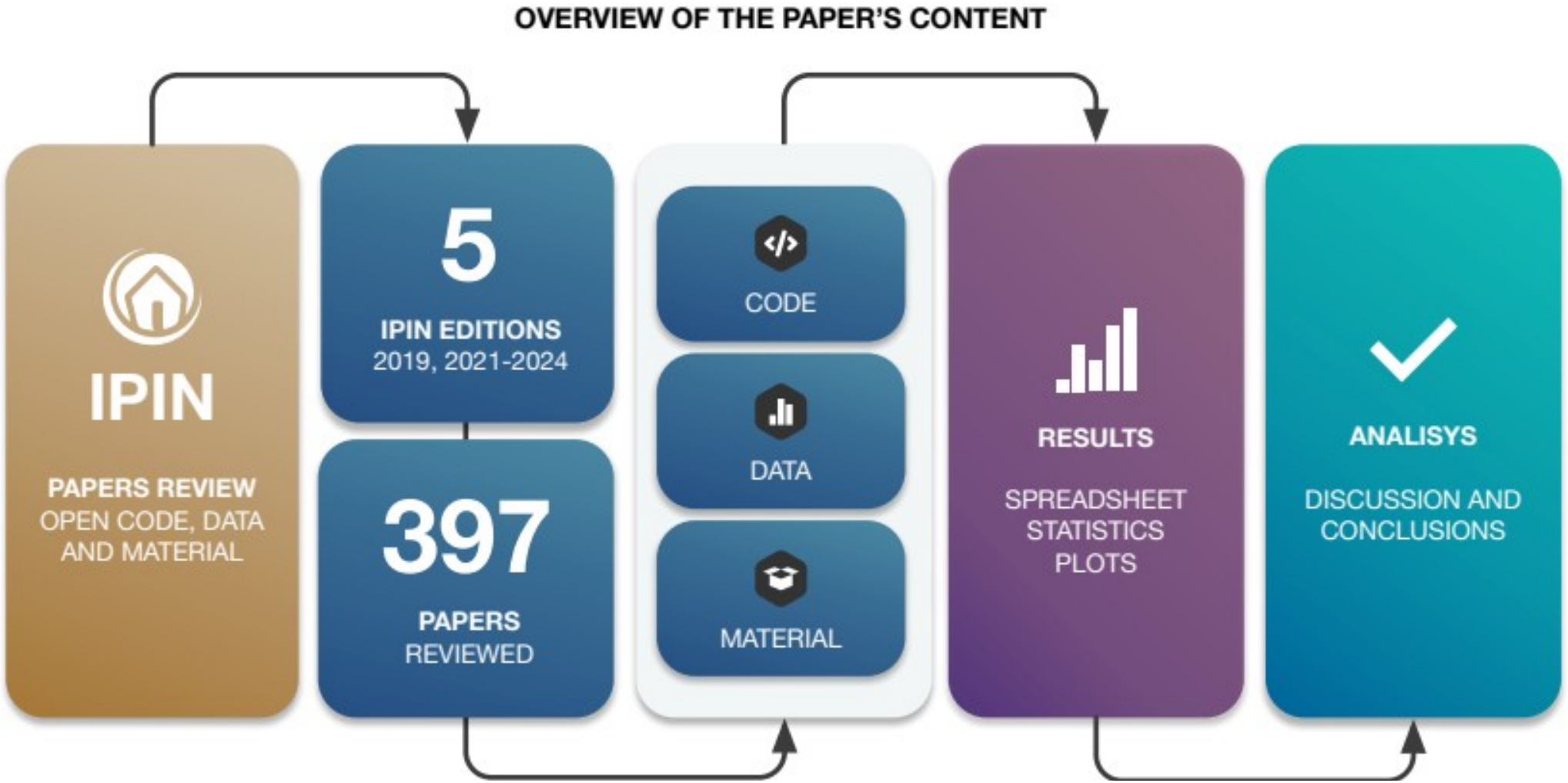
6.72%
PROVIDE 3x DATA SPLITS

Harder to compare methods fairly without train/val./test subsets

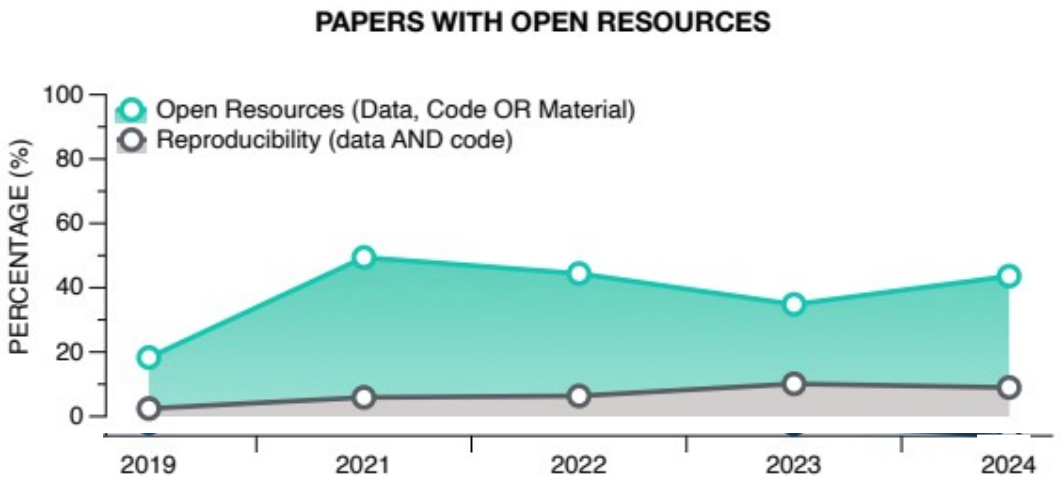
Open Science Practices in IPIN

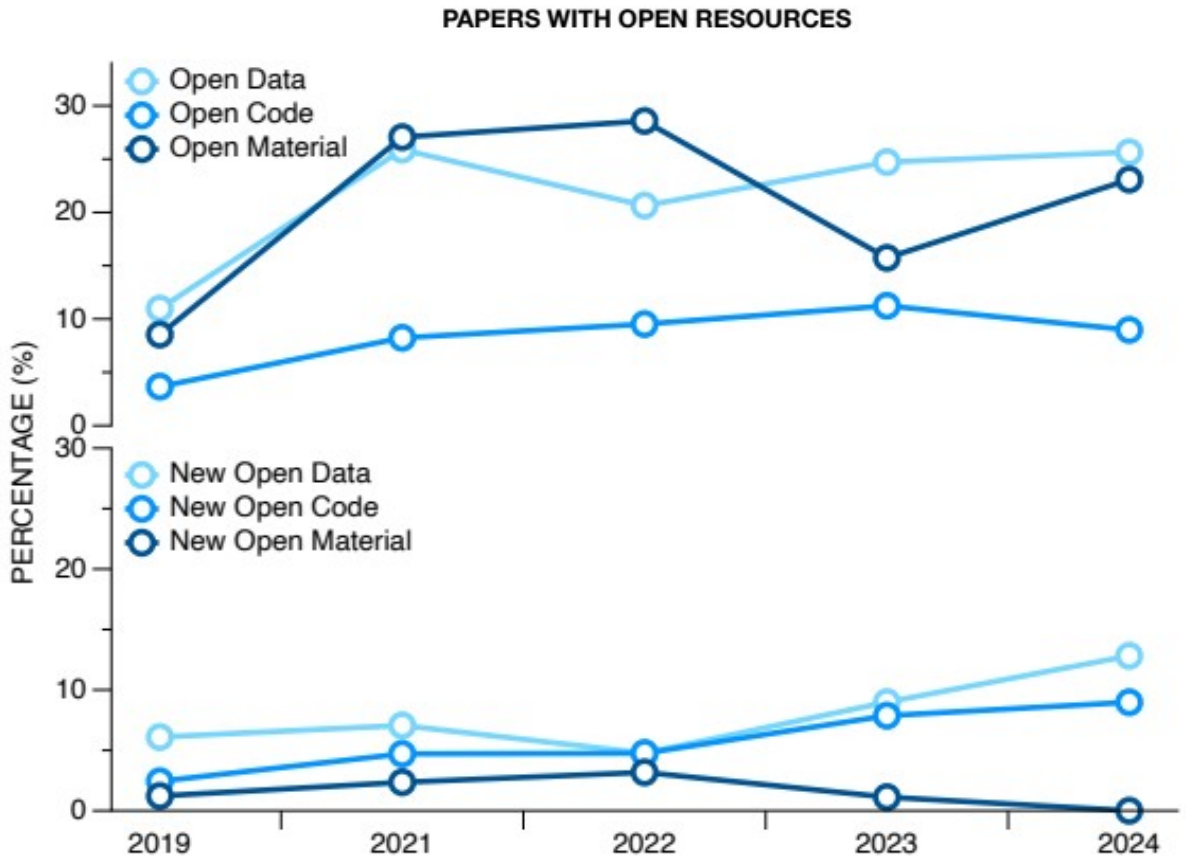


Open Science Practices in IPIN

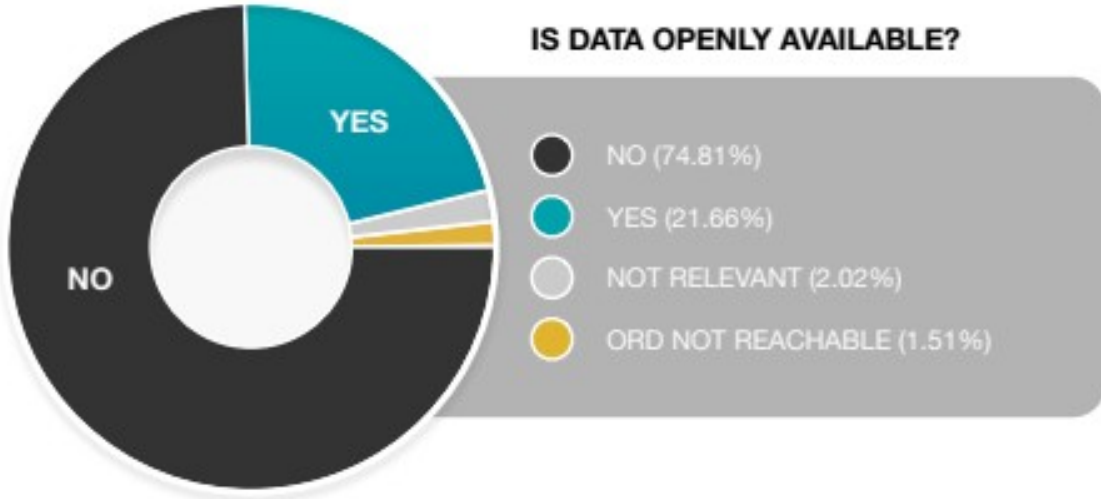


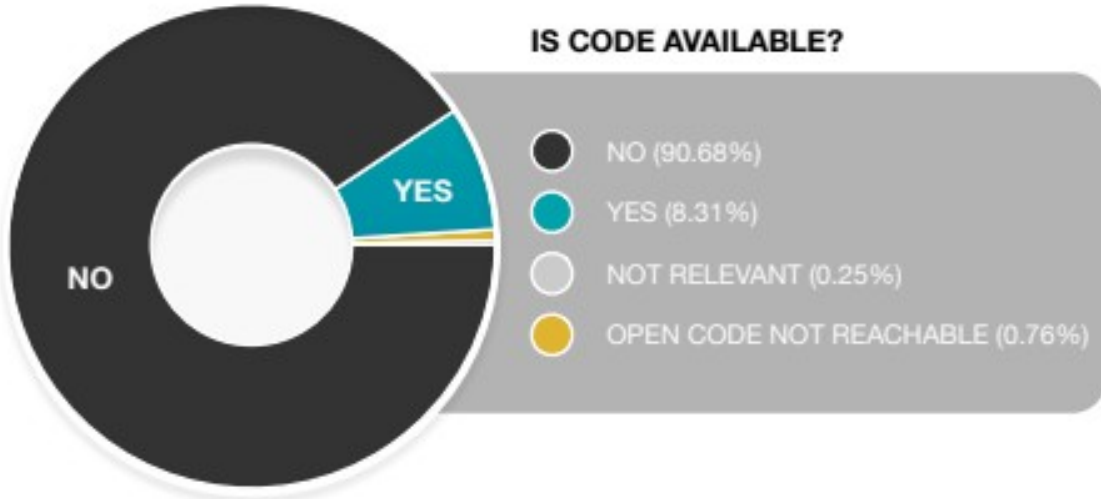
Open Science Practices in IPIN





Open Science Practices in IPIN



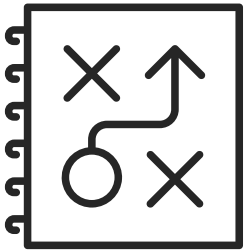


Guidelines for collecting a dataset



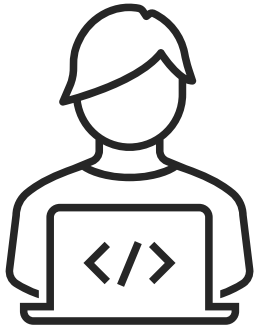
1. BEFORE

DATA COLLECTION



2. DURING

DATA COLLECTION



3. AFTER

DATA COLLECTION

PLAN ▶ **COLLECT** ▶ **SHARE**

01 Before: design for reuse

PLAN THE EXPERIMENT

Research question and intended reuse

🔍 Review existing datasets, identify gaps and plan/propose features

🛠 Prepare tools and materials

♻ Compliance with FAIR principles (Findable, Accessible, Interoperable, Reusable)



DESIGN THE DATA

🗂 Define metadata schema and file structure

Set units, timestamps and naming conventions

🔒 Privacy, consent and quality checks



02 During: record what happened

COLLECT DATA

Capture data:

- Reference positions
- Raw observations
- Timestamps



ANNOTATE

Document procedures:

- Photos/videos
- Device configs
- User instructions
- Routes and checkpoints



LOG

Record what happened

- Errors
- Missing samples
- Deviations from protocol



A dataset without its collection context is difficult to interpret.

03 After: publish a reusable package

FAIR COMPLIANCE

Ensure FAIR compliance:

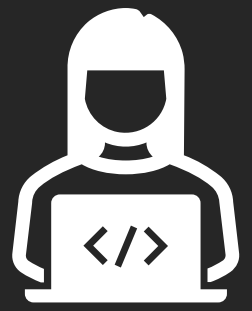
- 🔓 Open access
- 🏷️ Rich metadata
- Clear licensing



VALIDATE

Test and check:

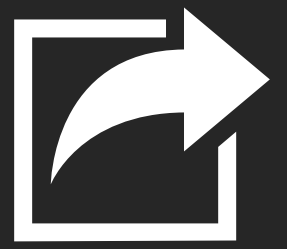
- Data in open format
- Validate dataset usability
- 💻 Provide code+examples



SHARE

Publish dataset:

- License
- Persistent identifier (DOI)
- 🏛️ Long-term repository (e.g., Zenodo)



Guidelines for dataset providers

Longevity	Deposit data in a repository committed to long term preservation, such as Zenodo, so it remains available beyond the project.
Documentation	Explain collection methods, variables, units and limitations so others can interpret the data correctly.
Format	Use open, accessible formats and provide loading examples to simplify reuse.
Licensing	State permissions and citation requirements so others know how they may reuse the data.
Persistent identifiers	Assign a DOI and identify releases so others can find and cite the exact dataset.

Guidelines for dataset users

Longevity	Retain a permitted copy of the version used so your analysis remains reproducible if access changes.
Documentation	Record selected samples, exclusions and processing steps so others can repeat the analysis.
Format	Preserve loading and conversion scripts so transformations remain transparent and repeatable.
Licensing	Respect reuse conditions and credit the creators, including original sources of derived datasets.
Persistent identifiers	Cite the dataset DOI and exact version, separately from its paper, so readers can trace your results.

Key Takeaways

- **Open science practices remain limited in indoor positioning research**: Progress is being made, but much remains to be done.
- **Follow FAIR principles**: Make shared data and code Findable, Accessible, Interoperable, and Reusable.
- **IPIN conference**:
  - We proposed the TPC to add incentives for authors following open science practices.
  - In IPIN2026, a Reproducibility Committee was established, implementing a Reproducibility Review, an optional review for papers that promote open science practices
- Special thanks to the team:
  - **Grigorios Anagnostopoulos** (HES-SO)
  - **Joaquín Torres-Sospedra** (UValència)
  - **Antonino Crivello** (ISTI, CNR)
  - **Paolo Barsocchi** (ISTI, CNR)
  - **Cristiano Pendão** (UTAD)
  - **Ivo Silva** (UMinho)

Obrigado!

https://opensciencefestival.netlify.app/

Do Planeamento à Partilha:

Boas Práticas para Criar Datasets de Investigação com Impacto

Open Science Festival, UMinho, 20.out.2026

Universidade do Minho