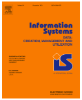
Reproducibility practices and initiative encouraged by Information Systems



Information Systems

DATA: CREATION, MANAGEMENT AND UTILIZATION

Dr. Juan J. Lastra-Díaz
Reproducibility Editor of Information Systems

NLP & IR Research Group (jlastra@invi.uned.es)
Department of Computer Languages and Systems
Universidad Nacional de Educación a Distancia (UNED)

December 15, 2020

Aims and Scope of Information Systems

Information Systems publishes articles concerning any research topic related or applied to data management and information systems

Information Retrieval
Machine Learning
Web Semantics
Scientific computing
Visual and Audio Information Systems
Business Process Management
Data Mining
Internet and cloud data management
Data Science

What is the aim of our reproducibility initiative?

Publishing reproducible papers to mitigate the reproducibility crisis [8]

- Pioneering reproducibility initiative[1] launched by IS editors in 2016
- It is based on a long experience on reproducibility tools like Reprozip [2] and reproducible works in SIGMOD DB-reproducibility [7, 4] since 2008
- Unique initiative for publishing reproducible experiments which are provided by the original authors and independently confirmed

Reproducible paper = Invited companion paper providing a detailed step-by-step reproducibility protocol based on a reproducibility dataset (software & data) to allow the exact replication of its parent paper, which is co-authored by its blind reviewers

[1] [1] F. Chirigati, R. Capone, R. Rampin, J. Freire, D. Shasha, A collaborative approach to computational reproducibility, Inf. Syst. 59 (2016) 95–97

Why is this reproducibility initiative so important?

- Current reproducibility crisis [10, 8] is undermining two pillars of the science ⇒ rigour & replicability of results
- A large number of papers cannot be reproducible because of:
  - Methods and/or experimental setup are not well detailed
  - Author's original data and software are not publicly available (non published, licensing restrictions, out of date, etc.)
  - Missing instructions for building, setting up, and running the software
  - Missing instructions for building final data and figures from raw output data (postprocessing & data analysis)
- Lack of independent replication studies ⇒ previous results & conclusions are copied and accepted without confirmation

Two reproducibility studies in the fields of NLP and Physics

- Wieling et al. [11] (2018) review 395 recent NLP papers concluding:
  - only 36.2% of the 2016 revised works provided their source code, and only $\frac{1}{10}$ of them could be reproduced exactly ⇒ repro. ratio $< 4\%$
  - “even if the source code and data are available, there is no guarantee that the results are reproducible“

- Stodden et al. [9] (2018) review 306 recent JCS[2] papers concluding:
  - “only about 6% (17 articles) of articles gave information making some artifacts available” ⇒ they emailed the remaining 298 authors
  - “we did not receive a reply from 37% of the authors”
  - “we received a reply but did not receive any artifacts from 48% of authors”
  - “roughly 15% supplied some artifacts to us” (55 of 306 articles)
  - “For the 55 articles with artifacts, we fully replicated none; partially replicated 32.7% (18); ran 54.5% (30); were able to build 3.6% (2); and had no progress on 9.1% (5)”

[2] Journal of Computational Physics, Elsevier

What are the main consequences of this crisis?

The lack of reproducibility resources hampers the research in multiple ways:

- It hampers the integration of newcomers (e.g. PhD students)
- It hampers the confirmation of previous results
- It discourages the evaluation of author's results
  ⇒ encourages the copy of unconfirmed results and conclusions
- It significantly increases the time and costs of any research work for the replication of the methods and experiments from other authors
- It prevents the sound and rigorous progress of any line of research
- It hampers the knowledge sharing in any research team
  ⇒ authors could be unable of reproducing their own results exactly

How does Information Systems (and Elsevier) want to help?

- Encouraging the development of reproducibility software & resources by rewarding the authors (and reviewers) with an additional paper
- Encouraging an independent replication and confirmation of previous results ⇒ visual reproducibility badging in the near future[3]
- Encouraging the adoption of good reproducibility practices since the very beginning of any research work

  ⇒ Beyond a Reproducibility-centered research methodology

- Encouraging the long-term reproducibility of any research work
- Increasing the impact and visibility of any research work
- Serving as a pilot program for other Elsevier journals

[3] Standard subscribed by main publishers, https://www.niso.org/standards-committees/reproducibility-badging

Definitions adopted by NISO[4] and subscribed by the ACM[5]

Repeatability same team, same experimental setup & software

⇒ authors obtain the same results with their software artifacts on different trials

Reproducibility different team, same experimental setup & software

⇒ other team obtains the same results with the author-created software artifacts

Replicability different team, same experimental setup & different software

⇒ other team obtains the same results with its own software artifacts by replicating the original author's methods

[4] https://www.niso.org/standards-committees/reproducibility-badging

[5] ACM has recently swapped the reproducibility and replicability concepts to match the NISO standard, https://www.acm.org/publications/policies/artifact-review-and-badging-current

What is a computational reproducible experiment?

Reproducible experiment

reproducibility dataset + reproducibility protocol

Reproducibility dataset

single & self-contained & publicly available & fixed collection of software and data ⇒ it enables the independent replication of reported experiments

Reproducible paper

article introducing and confirming a set of reproducible experiments (co-authored by reviewers)

Weakly reproducible experiment

reported conclusions are confirmed but not all results are reproduced exactly

Strongly reproducible experiment

reported results and conclusions are confirmed exactly

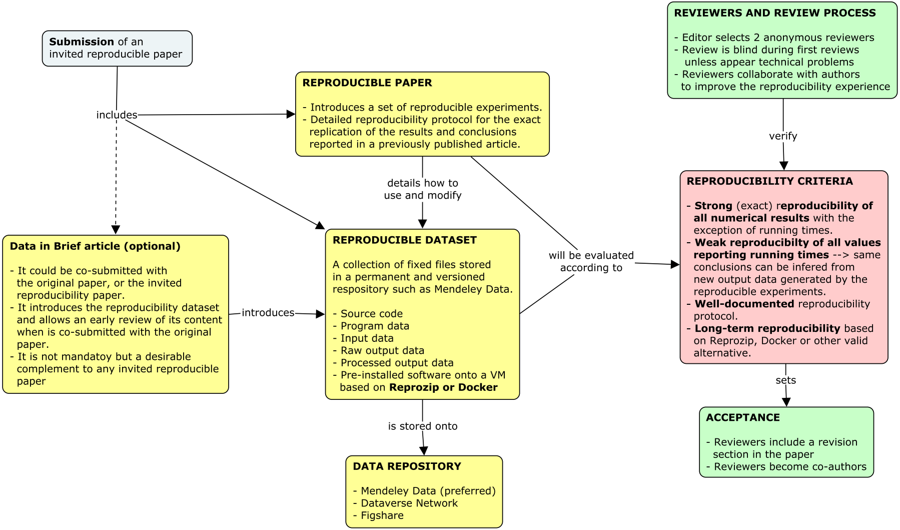
Adaptable reproducible experiment

a strongly reproducible experiment which is able to evaluate unexplored experimental setups

Scope of our reproducibility initiative

- Submission of reproducible papers is only by invitation
- We mainly invite reproducible papers for IS articles.
  → However, we could consider related articles in other Elsevier journals
- What are our ultimate goals?
  - Encouraging, capturing and disseminating reproducible experiments
  - Encouraging a reproducibility-centered research methodology
- Which article types are good candidates for a reproducible paper?
  - Any research article introducing reproducible experiments
  - Large experimental surveys
  - Any research article setting state-of-the-art results
  - Large evaluation campaigns setting standard experimental setups

Submission workflow



Submission of an invited reproducible paper

REPRODUCIBLE PAPER

- Introduces a set of reproducible experiments.
- Detailed reproducibility protocol for the exact replication of the results and conclusions reported in a previously published article.

REVIEWERS AND REVIEW PROCESS

- Editor selects 2 anonymous reviewers
- Review is blind during first reviews unless appear technical problems
- Reviewers collaborate with authors to improve the reproducibility experience

Data in Brief article (optional)

- It could be co-submitted with the original paper, or the invited reproducibility paper.
- It introduces the reproducibility dataset and allows an early review of its content when is co-submitted with the original paper.
- It is not mandatoy but a desirable complement to any invited reproducible paper

REPRODUCIBLE DATASET

A collection of fixed files stored in a permanent and versioned respository such as Mendeley Data.

- Source code
- Program data
- Input data
- Raw output data
- Processed output data
- Pre-installed software onto a VM based on **Reprozip or Docker**

REPRODUCIBILITY CRITERIA

- **Strong** (exact) r**eproducibility of all numerical results** with the exception of running times.
- **Weak reproducibilty of all values reporting running times** --> same conclusions can be infered from new output data generated by the reproducible experiments.
- **Well-documented** reproducibility protocol.
- **Long-term reproducibility** based on Reprozip, Docker or other valid alternative.

DATA REPOSITORY

- Mendeley Data (preferred)
- Dataverse Network
- Figshare

ACCEPTANCE

- Reviewers include a revision section in the paper
- Reviewers become co-authors

Why should any author adopt these guidelines?

- For increasing the quality, rigor, impact and credibility of all their scientific communications
- For contributing to making comparisons with their work easier
- For encouraging the adoption, citation and reuse of their research
- For speeding up the integration of newcomers (e.g. graduate students)

Practical long-term computational reproducibility

Practical long-term reproducibility	=	Docker → most successful & lightweight VM and/or Reprozip [2] + Docker → meta tool for building VMs

Can we reproduce most of experimental setups? Yes, we can but we should ...

- fix all data & software versions in our reproducibility dataset
- remove any randomness in your experimental setup → strong reproducibility,
- check your experimental setup to force at least its weak reproducibility

Random training of ML models ⇒ training + evaluation steps

- Random ML model ⇒ at most weakly reproducible, unless reproducible training being forced
- Deterministic evaluation of ML models ⇒ strongly reproducibility

Some basic recommendations

- Main goal = to be able to reproduce your results in a couple of hours
- Adopt a reproducibility mindset = to consider the reproducibility of your work since the very beginning of your research
- Write a reproducibility appendix (lab notebook) detailing everything
  ⇒ see our reproducibility guidelines for details
- Design and document your reproducibility dataset
  ⇒ co-submitted data paper (e.g. DiB paper) or dataset appendix
- Create a single and self-contained reproducibility dataset (Mendeley, Dataverse, or FigShare). Docker images could be stored into Docker Hub.
- All your data, figures, results, and conclusions should be reproducible
  ⇒ use data processing scripts (R-language or Python scripts)
- Automate your experimental setup & data pipeline (driver-program)
- Use open-source or free software for academics (check licensing)
- Test your reproducibility protocol (appendix) with a newcomer (e.g. graduate student) → best way of teaching reproducibility practices

Some examples of IS reproducible papers

Authors	Topic	Reproducibility SW	Verification	Results
Wolke et al. [12]	Dynamic resource allocation in cloud data centers	Reprozip + Docker + Python programs	*Raw output files. *R-language script for data analysis *HTML report	Strongly reproducible
Lastra-Díaz et al. [5]	New semantic measures library; benchmarks of semantic measures libraries; and word similarity benchmarks	Reprozip based on Docker + Java-based program	*Raw output files. *Final output files. *R-language script for data analysis and figures	Adaptable (strongly) reproducible
Fariña et al. [3]	Indexing methods for repetitive document collections	Docker + C++-based program	*PDF report. *Supplementary data tables	Weakly reproducible with minor corrigendum
Lastra-Díaz et al. [6]	Benchmarks of ontology-based methods and word embeddings on word similarity	Reprozip based on Docker + Java-based program	*Raw output files. *Final output files. *R-language script for data analysis. *HTML report	Adaptable (strongly) reproducible with minor corrigendum

The End

Reproducible science demands a strong commitment from everyone: authors, reviewers, editors, and publishers

Help us to make it happen !!!

Thank you very much !!!

[1] Chirigati, F., Capone, R., Rampin, R., Freire, J., Shasha, D., 2016a.
A collaborative approach to computational reproducibility.
Information Systems 59, 95–97.

[2] Chirigati, F., Rampin, R., Shasha, D., Freire, J., 2016b.
ReproZip: computational reproducibility with ease, in: Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD), bigdata.poly.edu. pp. 2085–2088.

[3] Fariña, A., Martínez-Prieto, M.A., Claude, F., Navarro, G., Lastra-Díaz, J.J., Prezza, N., Seco, D., 2019.
On the reproducibility of experiments of indexing repetitive document collections.
Information systems 83, 181–194.

[4] Freire, J., Bonnet, P., Shasha, D., 2012.
Computational reproducibility: state-of-the-art, challenges, and database research opportunities, in: Proceedings of the 2012 ACM SIGMOD international conference on management of data, dl.acm.org. pp. 593–596.

[5] Lastra-Díaz, J.J., García-Serrano, A., Batet, M., Fernández, M., Chirigati, F., 2017.
HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset.
Information Systems 66, 97–118.

[6] Lastra-Díaz, J.J., Goikoetxea, J., Hadj Taieb, M., García-Serrano, A., Ben Aouicha, M., Agirre, E., Sánchez, D., 2021.
A large reproducible benchmark of ontology-based methods and word embeddings for word similarity.
Information Systems 96, 101636.
https://doi.org/10.1016/j.is.2020.101636.

[7] Manolescu, I., Afanasiev, L., Arion, A., Dittrich, J., Manegold, S., Polyzotis, N., Schnaitter, K., Senellart, P., Zoupanos, S., Shasha, D., 2008.
The repeatability experiment of SIGMOD 2008.
SIGMOD Rec. 37, 39–45.

[8] Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.A., 2017.
A manifesto for reproducible science.
Nature Human Behaviour 1, 0021.

[9] Stodden, V., Krafczyk, M.S., Bhaskar, A., 2018.
Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility, in: Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–5.

[10] Stodden, V.C., 2010.
Reproducible research: Addressing the need for data and code sharing in computational science.
Computing in Science & Engineering 12, 8–12.

[11] Wieling, M., Rawee, J., van Noord, G., 2018.
Reproducibility in Computational Linguistics: Are We Willing to Share?
Computational Linguistics 44, 641–649.

[12] Wolke, A., Bichler, M., Chirigati, F., Steeves, V., 2016.
Reproducible experiments on dynamic resource allocation in cloud data centers.
Information Systems 59, 98–101.