What is the solution Secure Data Facilities could offer to enable reproducibility for papers based on personal and confidential data?

Beate Lichtwardt and Cristina Magder
UK Data Service, UKDA, University of Essex

8 November 2023

2023 Research Methods eFestival
Hosted by NCRM
#RMeF23

Roadmap

- An introduction to UK Data Service (UKDS).
- Reproducibility in Secure Data Facilities - current situation.
- Journal requirements.
- Landscape changes.
- What could potential solutions look like?
- Scenario A: Reviewer allowed @ UKDS SecureLab.
- Scenario B: Reproducibility Service established @ UKDS SecureLab.
- Outlook.
- Discussion.

UK Data Service (UKDS)

- hosts UK's largest collection of social, economic and population research data.
- provides users with access, support, guidance and training to facilitate high quality social and economic research and education.
- is a partnership between UK Data Archive at University of Essex, Cathie Marsh Institute for Social Research at University of Manchester, Jisc, UK, EDINA from University of Edinburgh, and the Department of Information Studies and Centre for Advanced Spatial Analysis at University College London.
- supports the development of best practices for data preservation and sharing standards.
- ukdataservice.ac.uk

Stats about UK Data Service

9,000 datasets in the collection.

300 new datasets and new editions added each year.

48,000 registered users.

130,000 downloads worldwide p.a.

UKDS data collections are accessed every 6 minutes
24 x 7 x 365.

UK Data Service users

Academic researchers and students.

Government analysts.

Charities and foundations.

Business consultants & data analysists.

Independent research centers & think tanks.

UK Data Service data sources

- National statistical authorities.
- UK government departments.
- Intergovernmental organisations.
- Research institutes.
- Individual researchers.

Office for National Statistics

CENTRE FOR LONGITUDINAL STUDIES

iSER

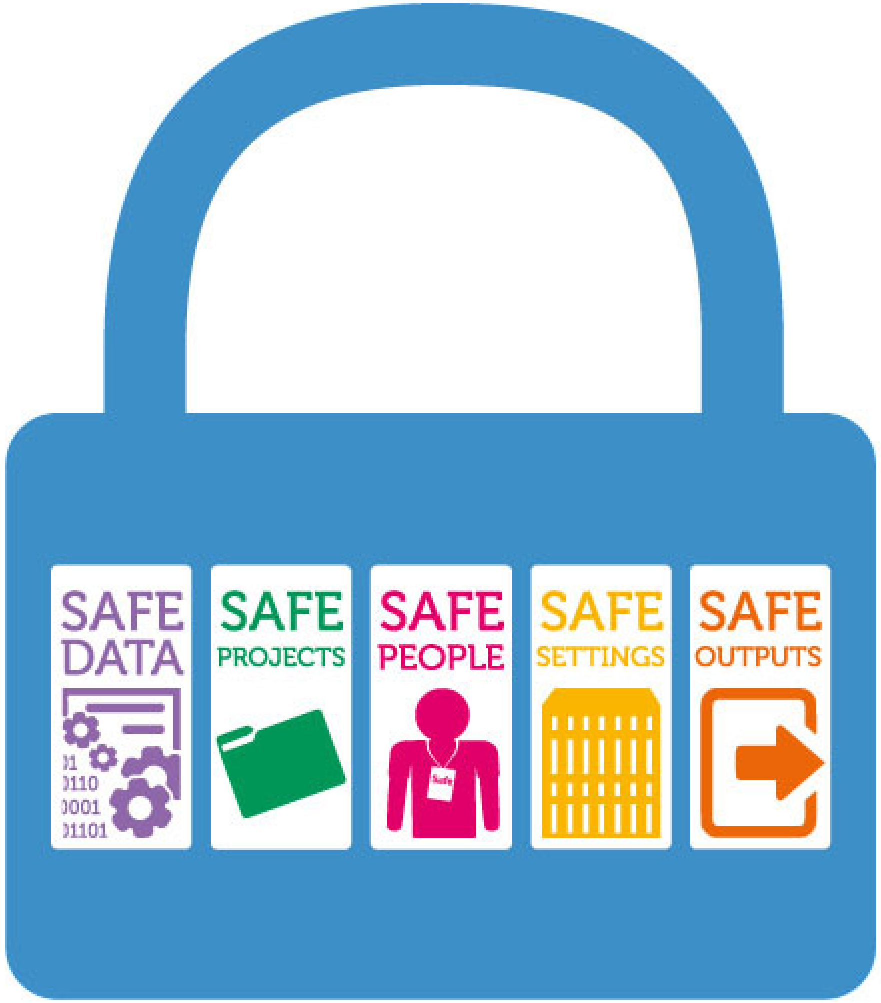
Department for Work & Pensions

INTERNATIONAL MONETARY FUND

Home Office

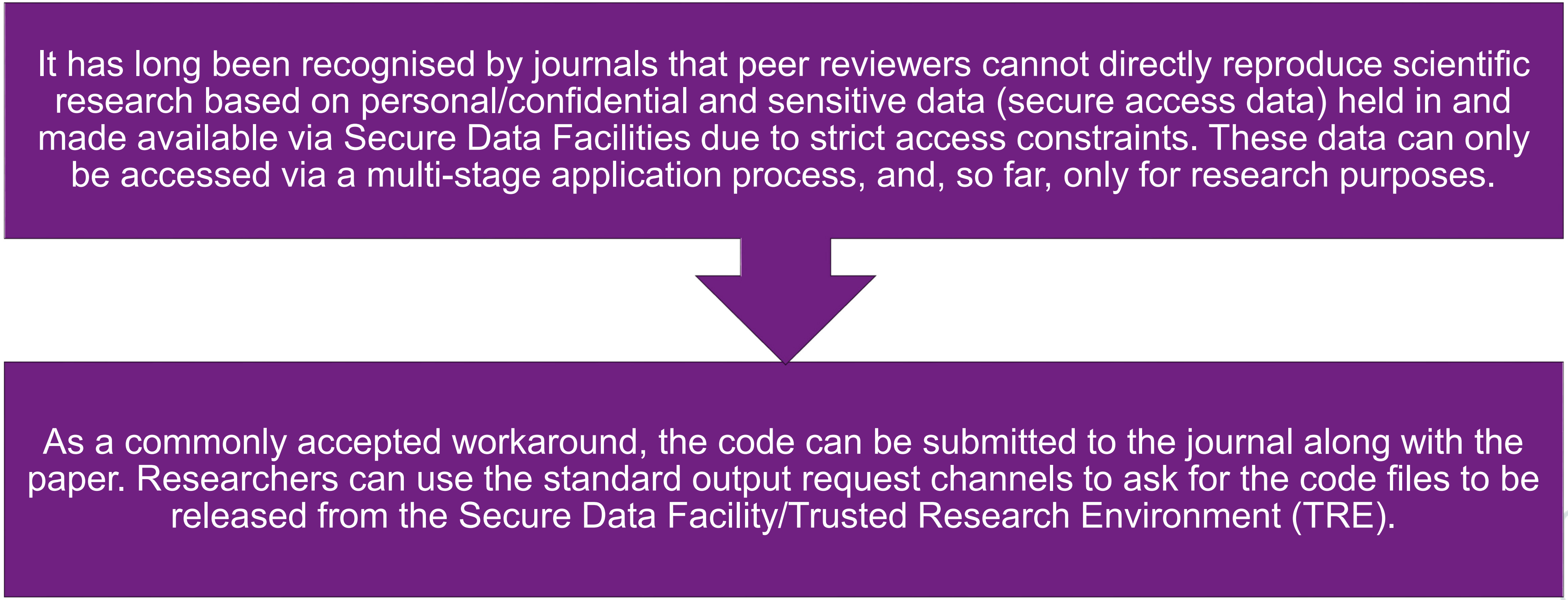
THE WORLD BANK

UKDS SecureLab



- Providing secure access to the most sensitive and confidential data since 2011.
- Using the Five Safes Framework to facilitate access.

Reproducibility in Secure Data Facilities - current situation



It has long been recognised by journals that peer reviewers cannot directly reproduce scientific research based on personal/confidential and sensitive data (secure access data) held in and made available via Secure Data Facilities due to strict access constraints. These data can only be accessed via a multi-stage application process, and, so far, only for research purposes.

As a commonly accepted workaround, the code can be submitted to the journal along with the paper. Researchers can use the standard output request channels to ask for the code files to be released from the Secure Data Facility/Trusted Research Environment (TRE).

Journal requirements - PLOS example

“PLOS journals require authors to make all data necessary to replicate their study’s findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data.

…

PLOS does not permit references to “data not shown.” Authors should deposit relevant data in a public data repository or provide the data in the manuscript.

…

PLOS recognizes that, in some instances, authors may not be able to make their underlying data set publicly available for legal or ethical reasons. This data policy does not overrule local regulations, legislation or ethical frameworks. Where these frameworks prevent or limit data release, authors must make these limitations clear in the Data Availability Statement at the time of submission.”

(Data Availability | PLOS ONE;
PLOS ONE is an inclusive journal community working together to advance science for the benefit of society, now and in the future. Founded with the aim of accelerating the pace of scientific advancement and demonstrating its value, we believe all rigorous science deserves to be published and should be discoverable, widely disseminated and freely accessible to all.)

Journal requirements - current situation

- “Typically, a ‘data access statement’ - saying that secure data can be accessed via the UK Data Service - would be fine. This would be accompanied by R scripts/DO files.” (SRT Attendee)

- **Data Availability Statement (example)**

  Data Availability: All available linked Millennium Cohort Study data can be accessed from the UK Data Service (University College London, UCL Institute of Education, Centre for Longitudinal Studies, SAIL Databank, NHS Wales. (2017). Millennium Cohort Study: NHS Patient Episode Database for Wales, Linked Administrative Datasets: ICD-10 Codes in Continuous Spells, 2001-2012: Secure Access. [data collection]. UK Data Service. SN: 8302, http://doi.org/10.5255/UKDA-SN-8302-1).

  (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213435)

Landscape changes

Well-established Secure Data Facilities like the UKDS SecureLab are increasingly receiving enquiries to enable robust and transparent reproducibility; and facilitate and assist peer reviewers prior to a journal article publication, especially for economics publications.

In the context of the constantly evolving secure access data landscape, reproducibility has become a growing concern for journals, researchers, and data service providers.

We will develop and examine possible solutions as to how Secure Data Facilities could handle the new reproducibility requirements for secure access data, and discuss the very practical implications of the proposed processes.

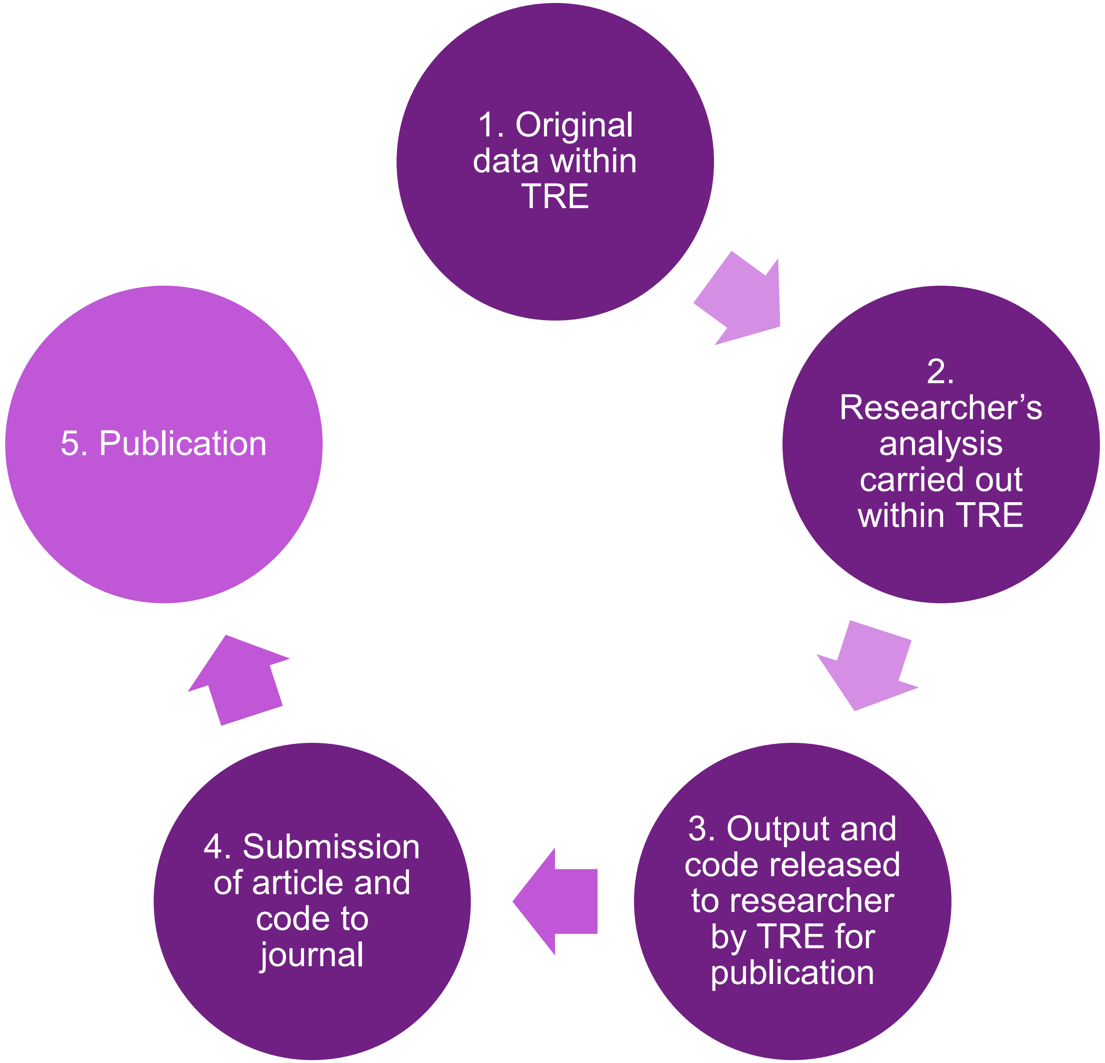
What could potential solutions look like?

- Scenario A: Allowing direct access for peer reviewers in the Secure Data Facility.

- Scenario B: Certified reproducibility provided by a tailor-made service (with-)in the Secure Data Facility.

The main aims of the presentation are,
- in the short run, to outline how Secure Data Facilities can support the peer review process better.
- in the long run, to help pave the way for enabling reproducibility of scientific research based on secure access data.

Current situation UKDS SecureLab



1. Original data within TRE

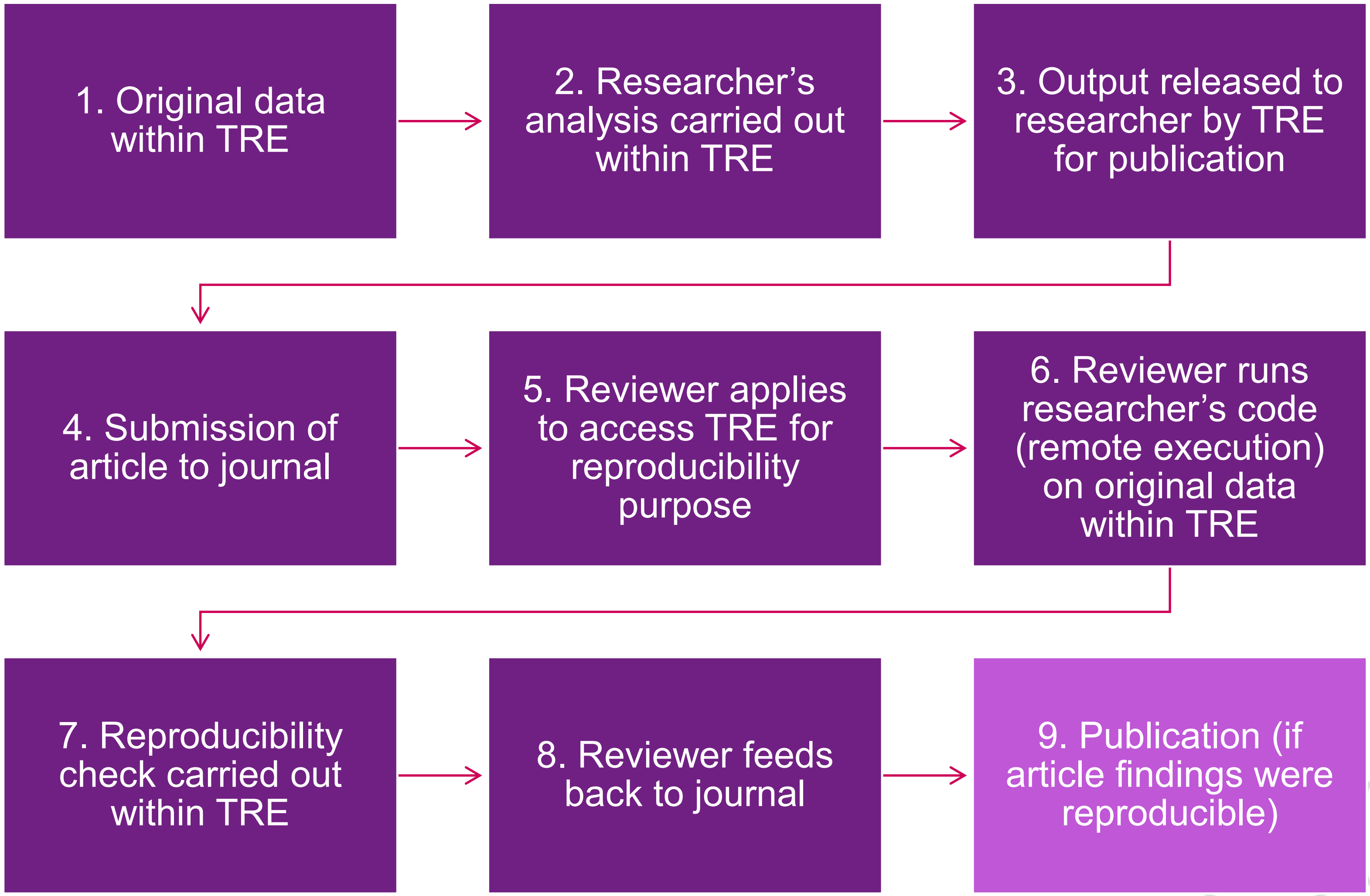
2. Researcher's analysis carried out within TRE

3. Output and code released to researcher by TRE for publication

4. Submission of article and code to journal

5. Publication

Scenario A: Reviewer allowed @ UKDS SecureLab



1. Original data within TRE

2. Researcher’s analysis carried out within TRE

3. Output released to researcher by TRE for publication

4. Submission of article to journal

5. Reviewer applies to access TRE for reproducibility purpose

6. Reviewer runs researcher’s code (remote execution) on original data within TRE

7. Reproducibility check carried out within TRE

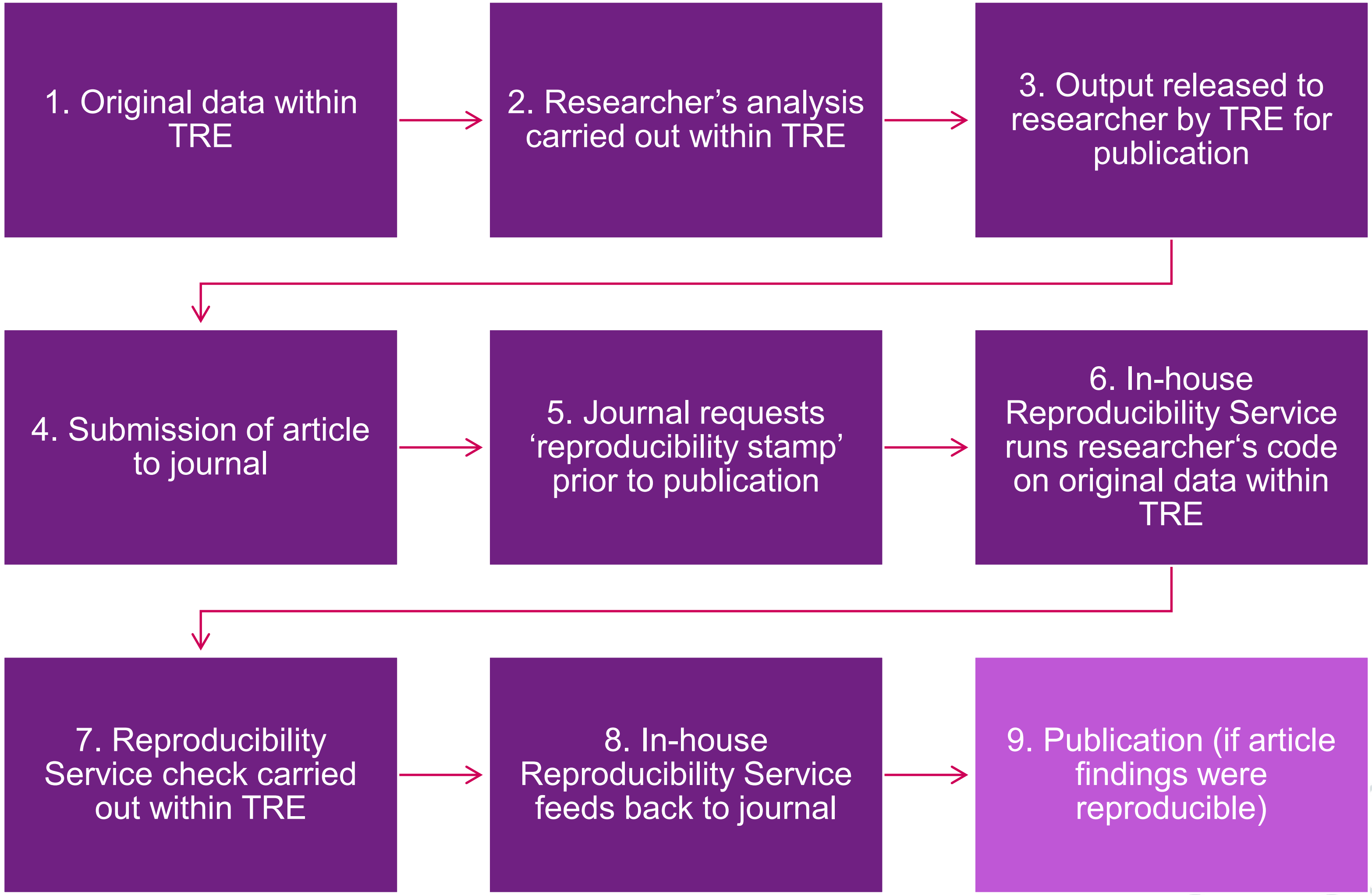
8. Reviewer feeds back to journal

9. Publication (if article findings were reproducible)

Scenario A: Challenges

Challenge	Detailed	Solution
Time.	Application estimate for accessing UKDS SecureLab data: 3 months.	New streamlined process for reviewers (considerations for single and double blind reviews).
Access.	Data owner approval needed to grant access for reproducibility purposes.	Reproducibility as a key component of the deposit licence agreements.
Costs.	Resources needed to train researchers, and establish a process and procedures for allowing reviewers access.	Funding.
Single and double blind reviews.	Public registers, i.e. Accredited Researchers and Accredited Projects.	Leverage legislation e.g. Digital Economy Act 2017; contract law.
Technical arrangements within TRE.	Project set up, access to data, documentation and code.	Establishing new setup for reproducibility process; only providing access to the code (remote execution) necessary to reproduce results used in publication.

Scenario B: Reproducibility Service established @ UKDS SecureLab



1. Original data within TRE

2. Researcher's analysis carried out within TRE

3. Output released to researcher by TRE for publication

4. Submission of article to journal

5. Journal requests 'reproducibility stamp' prior to publication

6. In-house Reproducibility Service runs researcher's code on original data within TRE

7. Reproducibility Service check carried out within TRE

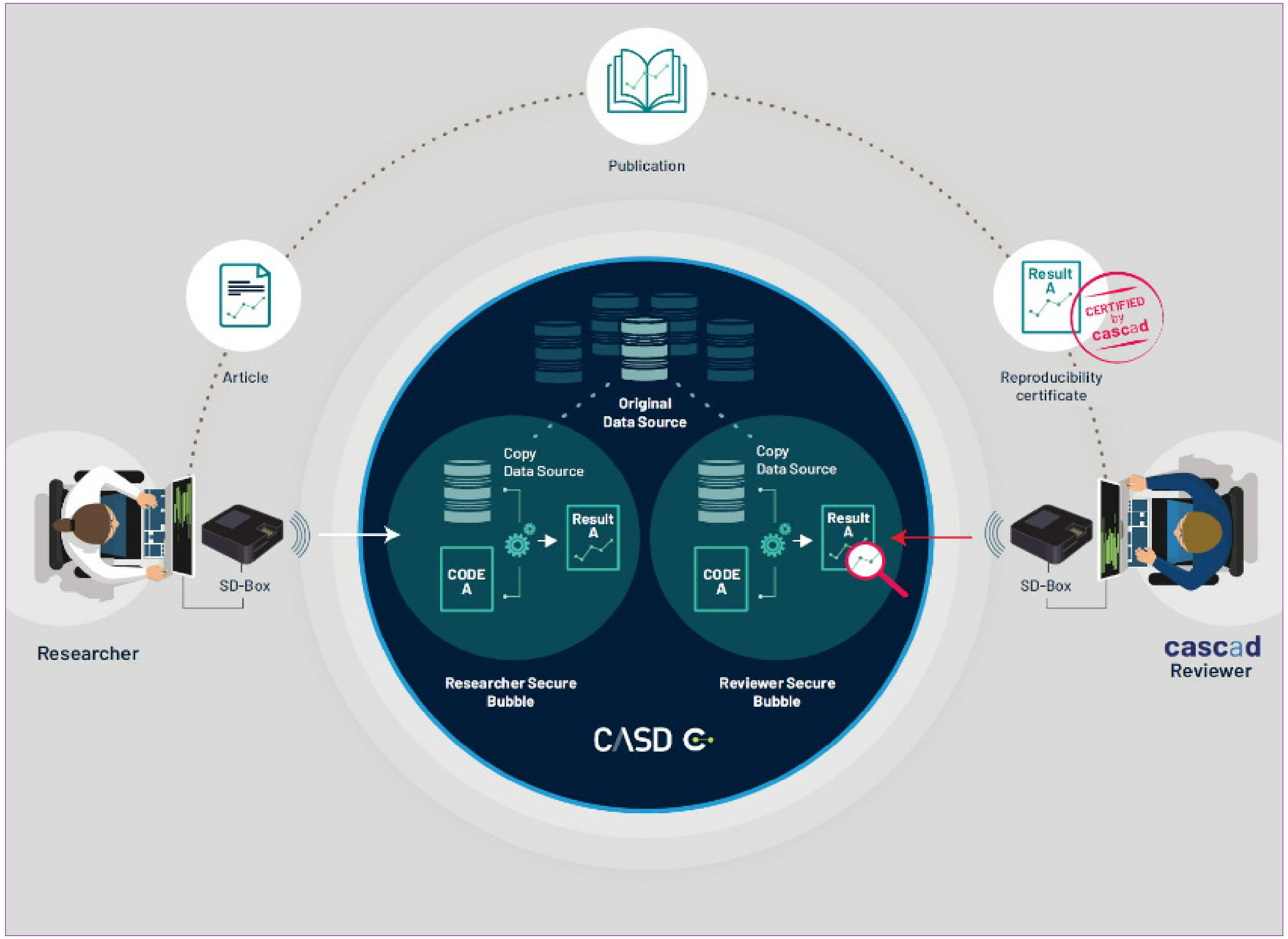
8. In-house Reproducibility Service feeds back to journal

9. Publication (if article findings were reproducible)

Scenario B: Challenges

Challenge	Detailed	Solution
Time.	Application estimate for accessing UKDS SecureLab data: 3 months.	Default access controlled via agreements.
Access.	Data owner approval needed to grant access for reproducibility purposes.	Reproducibility as a key component of the deposit licence agreements.
Costs.	A new sub-service has to be provided (cost intensive).	Funding needed to establish a Reproducibility Service.
Single and double blind reviews.	Public registers, i.e. Accredited Researchers and Accredited Projects; Access to internal systems.	Leverage legislation e.g. the Digital Economy Act 2017; contract law; sub-service not linked to other core services.
Technical arrangements within TRE.	Project set up, access to data, documentation and code.	Establishing new setup for reproducibility process; only providing access to the code necessary to reproduce results used in publication.

Example of a French Reproducibility Certification Agency: cascad-CASD



- 2018 Pilot; 3 year project from 2019, 4 year extension thereafter.
- Average workload: 1.5 days per request.
- Simplified certification-specific accreditation process of just 2 weeks.
- cascad controller will only access data and code for the time of certification.

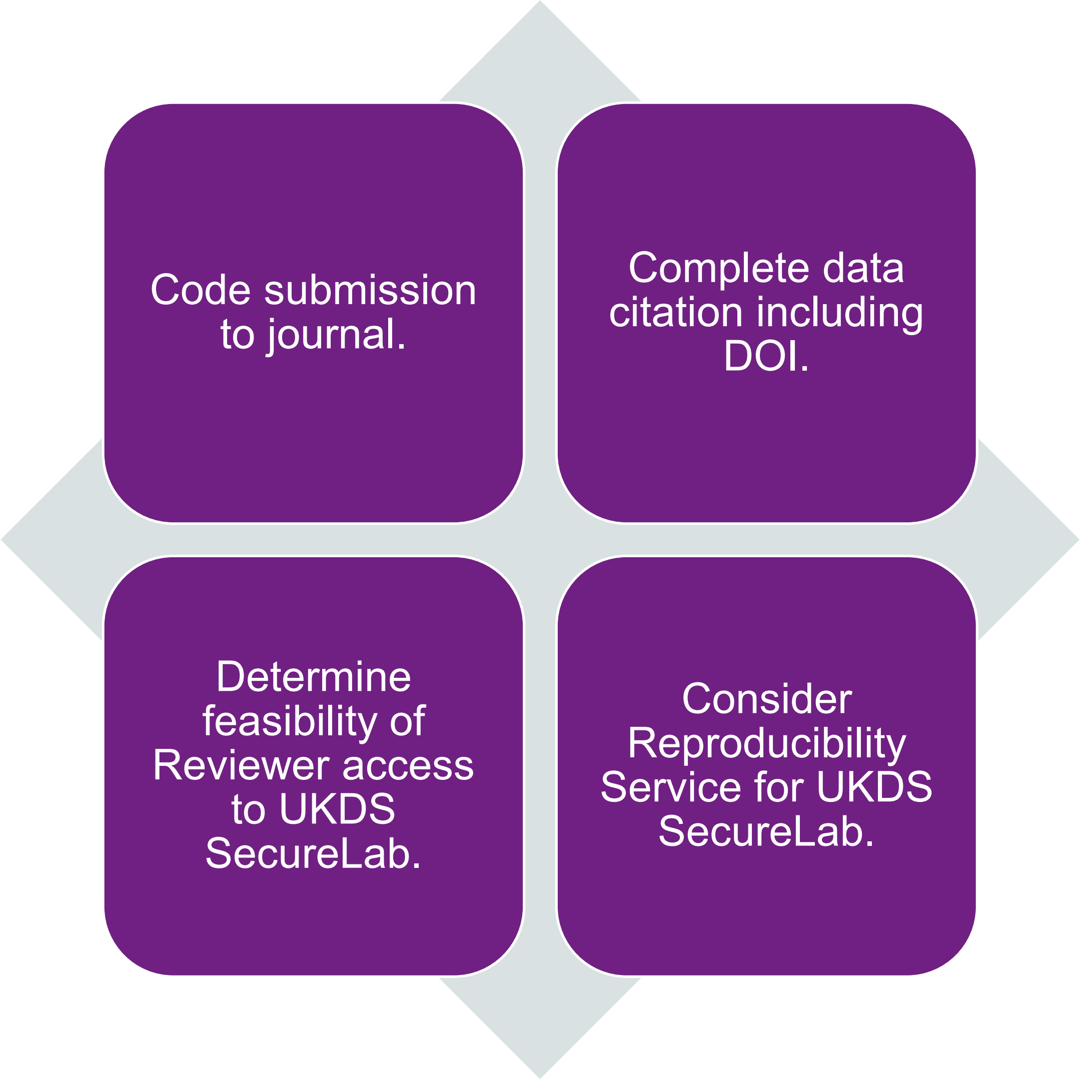
[Reproducibility Certification: cascad-CASD – Le CASD – Centre d'accès Sécurisé aux Données]()

cascad-CASD

The Certification Agency for Scientific Code and Data (www.cascad.tech) or cascad for short, is a non-profit, certification agency created by academics with the support of the French National Science Foundation (CNRS) and a consortium of French research institutions. The goal of this agency is to provide researchers with an innovative tool allowing them to signal the reproducibility of their research.

Source: Eric Debonnel and Roxane Silberman (2022)

Outlook



Code submission to journal.

Complete data citation including DOI.

Determine feasibility of Reviewer access to UKDS SecureLab.

Consider Reproducibility Service for UKDS SecureLab.

Thank you.

Beate Lichtwardt (blicht@essex.ac.uk)

Cristina Magder (dcmagd@essex.ac.uk)