Reproducible Research for Confidential Data

Christophe Bontemps

Porto, December 2019

What are we talking about?

- Definition of « confidential » is still fussy
  - Internally (to a firm) data are not confidential
  - Some versions of confidential data are not confidential (scale, aggregation)

- Nature of confidential data:
  - Produced by National Statistical Offices (NSOs) → “Administrative.”
  - Produced by private firms (Banks, insurance firms, HomeScanned data, Nielsen, …) →“Proprietary”
  - Produced by researchers (experiments..) → «Experimental »

AER Submission policy

Data and Code Availability Policy

It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented, and access to the data and code is clearly and precisely documented and is non-exclusive to the authors.

Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

Last accessed December 17, 2019

AER Submission policy

Data

For data, enough information should be provided (a) to accurately describe the data so that somebody who doesn't have knowledge of the data can understand its principal (and salient) characteristics (INFORMATION); (b) to be able to acquire the data (whether by download, by contract, by application process, etc.) (ACCESSIBILITY); and (c) to assure the reader that the data is available for a sufficiently long period of time (PERSISTENCE).

The data files can be provided in any format compatible with any commonly used statistical package or software. Authors are encouraged to provide data files in open, non-proprietary formats.

3 key requirements

a) Information on the data (data description, salient characteristics,…)

- Metadata

b) Accessibility of the data *(“by download, by contract, by application process, etc.”*)

- Application protocols, (national) restrictions, technical requirements, (national) regulations

c) Persistence of the data (Sustainably)

→What happens if these requirements are not fulfilled?

Reproducible Research with Confidential data

Current situation

- Researcher send a paper
- Journal evaluate the scientific content of the paper
- Journal (Editor) asks for code + data
- Researcher states that data is confidential
- Paper is still published.

Reproducible Research with Confidential data

Better situation

- Researcher send a paper
- Journal evaluate the scientific content of the paper
- Journal (Editor) asks for code + data
- Researcher states that data is confidential

  Somebody (Journal , CASCAD, …) checks that the results are reproducible
- Paper is still published

Access to data (for reproducibility purpose)

"The key to reproducible confidential data is mechanisms to facilitate non-exclusive access"

Vilhuber & Lagoze (2017)

- To whom?
  - Researchers (which ones?)
  - Academic institutions (universities, NSF, ANR, …)
  - Journal editors/referees
- Why?
  - To ensure trustable results (that will not be reproducible)
  - To make it reproducible by "others"

→ is there a sort of "ExecAndShare" on confidential data?

Two different goals

- At what cost?
  - Cost of access? (Who pays?)
  - Cost for reproducibility? (Who pays?)

Security and access requirements

Datatags as a means of specifying security and access requirements for sensitive data.
Sweeney *et al.* (2015)

→see also [Data Privacy Lab](#) at Harvard

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Data curation & data cycle management

- Differences in data curation between data providers
  - NSOs
  - Private firms
  - Researchers

    → Datajournals, Metadata?
- Differences in data versioning
  - NSOs
  - Private firms
  - Researchers

    → MetaData (XML, DDI), DOI?

A Revolution On Its Way?

- Barriers to entry?
  - To data ?
  - To journals?
  - To publication….
- Open journals excluded?
- Contracts everywhere, regulation everywhere, costs everywhere?
- What about sustainability (data, regulations, contracts…)

First mover disadvantage?