

# Analysis of data policies in scholarly journals – a case study



Thomas Seyffertitz  
Vienna University of Economics and Business  
Library Center – Research Data Management

[researchdata@wu.ac.at](mailto:researchdata@wu.ac.at)

112. BiblioCon, Hamburg, June 4-7, 2024



- Introduction & Background
- Objectives
- Method
- Journal data policies
  - Formal features (examples)
  - Content elements (examples)
- Conclusions & Recommendations

- **Journal data policies – a working definition**
  - document provided by a journal or its editor/owner that
  - contains **instructions, recommendations and regulations** on how to handle (research) **data and code** on which the **submitted paper** is based.
    - e.g. regulations on data availability, reproducibility, data sharing, data depositing etc.
- **Different origins of data policies**
  - **publisher policies:** publisher/research organization/associations; valid for specific group or all journals in journal portfolio ([Springer Nature](#), [AEA](#))
  - **journal specific data policy** (e.g. created by the editorial board)

# Introduction & Background

## WU Star-journal-list

- Purpose: Incentive to publish in top journals in the relevant subject areas → monetary award (bonus)
- 2022/23 as part of an extensive selection process
- → sample of 84 scholarly journals
- more than 270 publications by WU researchers since 2017 in these journals

Subjects	No. of journals
Economics	17
Business	13
Management	12
Psychology	9
Environmental Research	7
Political Science	7
Math./Stat./Comp.Sci.	6
Educational Research	4
Information Systems	3
Multidisciplinary Science	3
Sociology	2
Communication	1

- **We are interested in**
  - **Formal features** relate to **presentation**, **preparation** and **clarity**.
  - **Content elements** of the data **policy** for handling research data
- **Objectives**
  - present some **key features** of policies together with **practical examples** and
  - **raise awareness** for data policies of scholarly journals
  - develop **advisory services** to (early career) researchers

- **Collected** policies/guidelines from the journal/publishers or learned societies' websites
- **Converted** to pdf
  - using a predefined naming convention
  - resulted in **111 documents for 84 journals**
- Used R to produce **text corpora** from the pdfs
- **Selection of policy elements** of interest
  - used predefined terms/phrases to look for content elements (e.g. data availability, reproduction)
  - created some basic frequency tables
- **Manual review** of policies matching elements

# Journal data policies

## Formal features overview

Category	Description
Location	<ul style="list-style-type: none"><li>○ <b>publisher's</b> (journal) website</li><li>○ website of <b>association</b> (e.g. <a href="#">AMA</a>)/learned society or <b>university</b></li></ul>
Naming	How are the data <b>policies called</b> ? Examples: <ul style="list-style-type: none"><li>○ "<u><i>Code Sharing Policy</i></u>"</li><li>○ "<u><i>Reporting standards and ...</i></u>"</li><li>○ "<u><i>Data and Code Availability Policy</i></u>"</li><li>○ "<u><i>Guide for authors</i></u>" ..., etc.</li></ul>
Presentation, Format and clarity	<ul style="list-style-type: none"><li>○ <b>separate</b> data policy documents, or</li><li>○ <b>part</b> of author guidelines or other policy documents</li><li>○ (mostly) <b>HTML</b>, but also <b>pdf</b></li></ul>
Timing	<b>When</b> do data guidelines become <b>effective</b> for the author: submission, acceptance or publication

# Journal data policies – formal features

## Not unambiguous location

- „**missing link**“: at the [Wiley website](#) there is no link to the policy at the [journal owner's website](#) (University of Pennsylvania)

The screenshot shows the Wiley website for the *International Economic Review*. The navigation bar includes 'HOME', 'ABOUT', 'CONTRIBUTE', and 'BROWSE'. The 'CONTRIBUTE' dropdown menu is open, showing 'Author Guidelines', 'Open Access', 'Submit a Manuscript', and 'For Referees'. A red arrow points from the 'Open Access' option to the 'Data Availability Policy' link in the main content area. The main content area features a header for the journal, a navigation bar with 'EDITORIAL INFORMATION', 'ISSUES', 'SUBMISSIONS', and 'SUBSCRIPTIONS', and a section titled 'Data Availability Policy' with the text: 'EFFECTIVE January 1, 2022. To promote consistent standards and requirements among ge... Review has adopted the AER data availability policy. We are, policy. Data Availability Policy. It is the policy of the IER to publish papers only if the data us available to any researcher for purposes of replication. Authc'.

→ may be challenging to find



# Journal data policies – formal features

## Clarity & presentation example

- Publisher: <https://www.elsevier.com/researcher/author/submit-your-paper/sharing-and-promoting-your-article> and
- <https://www.sciencedirect.com/journal/journal-of-financial-economics/publish/guide-for-authors>
- and **at a separate journal's site**: <https://www.jfinec.com/news/data-and-code> → JFE's „News“ page **leads to the entire policy**:

Effective April 3, 2023, all manuscripts that receive an “accept pending uploads” decision must follow the new [data and code policy](#), which is identical to the old policy except for two additions. The data editor will review all code and data packets to ensure that the packets are complete. In those cases in which the actual data cannot be disclosed, a pseudo data set must be provided. Articles will not be accepted outright until the data editor approves them. Articles that have received an “accept pending uploads” decision before April 3, 2023, must comply with the old data and code policies. Articles that were initially submitted before July 1, 2021, are exempt.

**Attention:** different guidelines at different locations for the same journal (!?)

# Journal data policies – Content elements I

Category	Description
Data availability statement (DAS)	<p>Is a DAS mentioned in the guidelines or not?</p> <p>It confirms the presence or absence of shared data.</p> <p>It is used in publications to describe where the data directly underpinning the publication can be found and under what conditions they can be accessed(*)</p>
Data repository	<p>Mentioning of the subject „data repository“ or notes on certain data repositories</p> <p>Example: <u><a href="#">QJE Dataverse repository for replication packages and raw data</a></u></p>

# Journal data policies – Content elements II

Category	Description
Data & code sharing	Statements about data/code sharing in general, e.g. mandatory or only recommending
Pseudo-data	If original data can not be provided due to certain restrictions
Reproducibility/ Replication	Statements about reproducibility, its requirements, recommendations, and how-to descriptions (replication package requirements)  Is it required? What must a replication package contain?
Data editor	governing the data policy compliance, advising authors in data related aspects; in charge of reproducibility issues

# Content elements: DAS & Data repository

In how many policy documents do terms occur?

Term	Occurrence*	Note
data availability	42	all disciplines
code availability	16	all disciplines
data availability statement	30	all disciplines
data statement	9	economics/business

Term	Occurrence*	Note
repository	75	all disciplines
dataverse	22	all disciplines
code repository	13	economics

# Content element: Data & code sharing

- Example from the „Guide for Authors“ of the **Journal of Accounting and Economics**

... This **journal encourages** and enables you to share data that supports your research publication where appropriate, and enables you to interlink the data with your published articles. .... To **facilitate reproducibility** and data reuse, this journal also **encourages** you to share your software, code, models, algorithms, protocols, methods and ....

- Example from the „Guide for Authors“ **Accounting, Organizations and Society**

... This **journal requires** and enables you to share data that supports your research publication where appropriate and enables you to interlink the data with your published articles. ...

Term	Occurrence*
data sharing	36
code sharing	8
share data	33
share code	2
sharing policy	12

\*) the term occurs at least once in  $n$  policy documents ( $N=111$ )

# Content element: Pseudo-data

- Journal of Finance: "... Authors ... encouraged to include the data... If authors choose not to provide the data, **they must include a pseudo-data set** that illustrates the format of the files read **by the code** so that users can understand and **check the functionality of the code**. ..."
- Journal of Financial Economics: "... data cannot be **disclosed**, the authors **must supply pseudodataset(s)** to **demonstrate that the code runs**. ..."
- Review of Financial Studies: "... **encourage** authors to include the data ... source code. ... authors **choose** not to provide ..., or ... **restricted** ... because of copyright or confidentiality ..., ... **required** to **include a pseudo-data** ...".

This element occurs **in only six documents**, but is of importance as it may cause additional work for researchers!

# Content element: Replication – an example

- **Problem:** licensed/proprietary data not to be shared
- **Solution:** in-depth description how data can be derived
- contact (data) editor at or prior to submission (!)

## Data for Replication of Hébert and Schreger “The Costs of Sovereign Default: Evidence from Argentina”

This document outlines how to collect the replication data for the paper. Much of the data used in the paper is from commercial data vendors and cannot be shared. In this case, this document will outline how to acquire the data and a skeleton file will be posted so a reader can verify the acquired data is in the correct format for the replication files to run. In the below instructions, this will be indicated as [Raw Data Unavailable].

When publicly available data is used, the raw file is posted along with information on where it was acquired. This case will be noted as [Raw Data Available].

VOL. 107, NO. 10, OCTOBER 2017  
(pp. 3119-45)

[Download Full Text PDF](#)

### Article Information

#### Abstract

We estimate the causal effect of sovereign default on the equity returns of Argentine firms. We identify this effect by exploiting changes in the probability of Argentine sovereign default induced by legal rulings in the case of NML Capital, Ltd. v. Republic of Argentina. We find that a 10 percent increase in the probability of default causes a 6 percent decline in the value of Argentine equities and a 1 percent depreciation of a measure of the exchange rate. We examine the channels through which a sovereign default may affect the economy.

#### Citation

Hébert, Benjamin, and Jesse Schreger  
*Economic Review*, 107 (10): 3119-45

DOI: 10.1257/aer.20151667

Choose Format:

#### Additional Materials

[Replication Package \(72.84 MB\)](#)

[Online Appendix \(773.00 KB\)](#)

[Author Disclosure Statement\(s\) \(1\)](#)

Term	Occurrence
reproducibility	37
replication	54
replicability	10
replication package	12
reproduce	50
replicate	47

# Content element: Data editor – examples

- **Review of Economic Studies:** "... Requests for **exemptions** should be clearly stated **when the article is first submitted**. The article will then be **reviewed** at the discretion of the Managing Editors and the **Data Editor**. Exceptions will not be considered later ..."
- **Management Science:** "... task of ... **Data Editor** is to help authors to get their published research **compliant** with the ... **Data Policy**. ... **advises authors of published papers** on the data, materials, and information to be provided to allow other researchers to replicate the original results. ..."

➔ **American Economic Review and Journal of Economic Literature:** "... **data editor** will **assess compliance** with this policy and will **verify** the **accuracy** of the information **prior to acceptance** by the Editor. ... private (not to be published) version of the data should be provided to the AEA Data Editor and/or a designated third-party replicator who can provide a third-party **reproducibility** report. ... **After the data and code deposit is accepted** by the AEA Data Editor, it will become the version of record associated with the paper. ..."



# Conclusions

- Most **comprehensive** and **stringent data policies** are **community-based**: AEA, AERA, INFORMS, APS etc.
- Different **disciplines** have **different priorities** in their data policies
- Commercial publishers sometimes provide **framework policies**
  - Journal guidelines may refer to a publisher's policy concerning data availability (e.g. *Review of Finance* by OUP: Level1- to Level4)
- **Data availability statements** have become quasi standard in policies
- **Lack of awareness** about data policies can lead to significant **time and effort wasted** (revisions, delay in publication) or **increase the likelihood** of submitted paper being **rejected**

**@analysis: frequency of terms does not say much about the mandatory nature of the policy**

# Recommendations

- **Start as early as possible** reading data guidelines of the journal; at least **before submitting** the paper
- **Research community**: Participate in academic discussions, conferences, or forums related to data-sharing and reproducibility
- **Journal Websites**: Explore the websites of journals, especially those known for stringent data policies.
  - Some journals provide information on their editorial decisions, including instances where papers were rejected due to data-related issues (or see, e.g. <https://retractionwatch.com/>)
- **Familiarize** with (relevant) **tools** (e.g. code checking)
- **Contact** (data) **editor at** or **prior to submission**, if not sure which policy features may apply

**important for  
researchers per se,  
BUT  
research supporting  
staff may build  
information or  
advisory services  
upon**

# Excursus: Journal data policies on generative AI ?

- regulating use of generative AI tools **in writing submissions**
- AI bots (e.g. ChatGPT) **should not** be listed as **an author** in the submission;
- Authors are required to provide a list of
  - **sources** used to generate content and **citations generated** by language models
- Clearly **indicate** the use of **language model** (which model, what purpose)
- **Examples:**
  - [Guide for authors - Global Environmental Change](#) (Elsevier)
  - [Association for Psychological Science - Editorial Policies](#) (APS)

# Questions!