metajelo: A metadata package for journals to support external linked objects

Lars Vilhuber[1] Carl Lagoze[2]

[1]Cornell University
[2]University of Michigan

2019 Feb 05

Reproducibility of research

Reproducibility and replicability of scientific findings has been given great scrutiny in recent years

(Camerer et al. 2016; Collaboration 2015; Fanelli 2018; Klein et al. 2014) .

Historically, it has been difficult to find the materials required to conduct reproducibility or replication exercises

(Dewald, Thursby, and Anderson 1986; McCullough, McGeary, and Harrison 2006; McCullough and Vinod 2003) .

Reproducibility of research

Journals are supporting the endeavor with “Data [and Code] Availability Policies” [DAP]

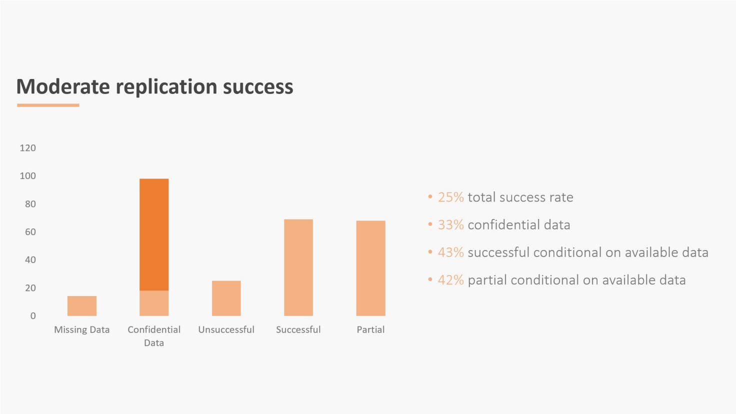
with variable success rates (Höffler 2017a; Stodden, Guo, and Ma 2013; Stodden, Seiler, and Ma 2018; Stodden et al. 2016) .

Not a panacea

Despite DAPs, researchers often find that studies do not reproduce (Chang and Li 2017, 2015; Höffler 2017b; Stodden, Seiler, and Ma 2018)

Replication rates

Our own study



Moderate replication success

- 25% total success rate
- 33% confidential data
- 43% successful conditional on available data
- 42% partial conditional on available data

Restricted-access Data

Confidential Government Data

- U.S. Census Bureau
- Statistisk sentralbyrå (Statistics Norway)
- many others

Confidential sub-national data

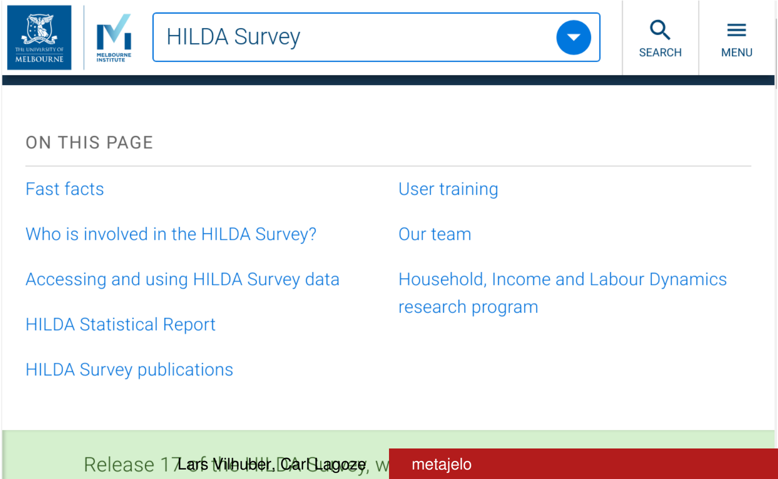
Just in the US:

- North Carolina Education Data
- Ohio Earnings and Education Statistics
- Oregon Health Data
- etc.

Restricted-access Data

Here in Australia

Household, Income and Labour Dynamics in Australia (HILDA) Survey . . .



ON THIS PAGE

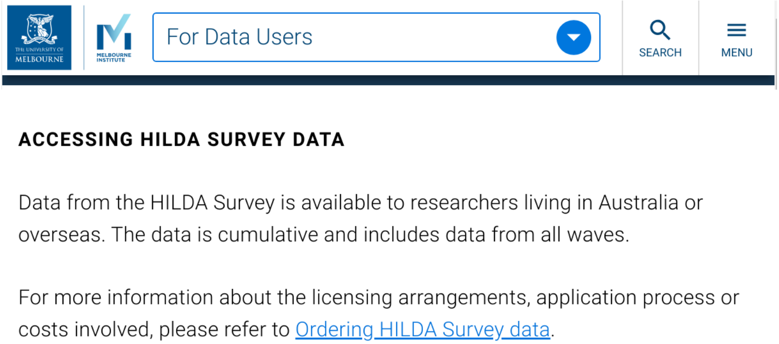
- Fast facts
- Who is involved in the HILDA Survey?
- Accessing and using HILDA Survey data
- HILDA Statistical Report
- HILDA Survey publications
- User training
- Our team
- Household, Income and Labour Dynamics research program

Release 17 of the HILDA Survey, w

Restricted-access Data

Here in Australia

Household, Income and Labour Dynamics in Australia (HILDA) Survey . . . which is **restricted-access**



ACCESSING HILDA SURVEY DATA

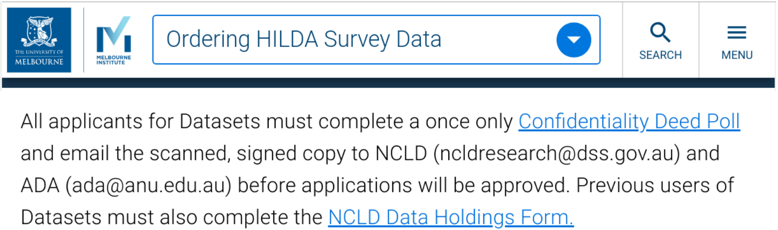
Data from the HILDA Survey is available to researchers living in Australia or overseas. The data is cumulative and includes data from all waves.

For more information about the licensing arrangements, application process or costs involved, please refer to Ordering HILDA Survey data.

Restricted-access Data

Here in Australia

Household, Income and Labour Dynamics in Australia (HILDA) Survey . . . which is **restricted-access**



All applicants for Datasets must complete a once only Confidentiality Deed Poll and email the scanned, signed copy to NCLD (ncldresearch@dss.gov.au) and ADA (ada@anu.edu.au) before applications will be approved. Previous users of Datasets must also complete the NCLD Data Holdings Form.

Requirements for Reproducibility

What do we need?

- Description of the analysis (possibly as code)
- Access to the data

Is restricted-access data compatible with reproducibility?

Answer: yes

- Theoretical access to confidential data:
  - More than **1000** users (not just Germans from Germany) have been granted access to German confidential labor market data (Müller and Möller 2019)
  - More than **1500** users currently have access to confidential French data (`https://casd.edu`)
  - More than **700** researchers were active at the end of 2017 in the US Federal Statistical Research Data Centers (U.S. Census Bureau, 2018)

Is restricted-access data compatible with reproducibility?

Answer: yes

- Replications actually do occur with restricted-access data

  For instance, see exchange between Olivier (2016) and Chemin and Wasmer (2017) regarding Chemin and Wasmer (2009), using French (Réseau Quetelet) data

Access Conditions

Is access **easy**?

- No

Is access **fast**?

- No

Can **others** access the data?

- **Yes**!

Is there a **process** for granting access?

- **Yes**!

No data, OK - surely we have metadata on all those things?

No.

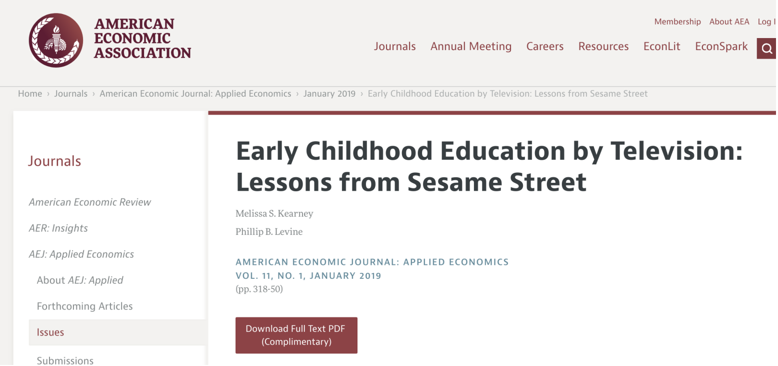
Lack of reliable metadata

There is a

pervasive lack of consistent, reliable metadata

on the materials provided to journals, and in particular those provided through third-party locations.

Example 1: AEA journal



AMERICAN ECONOMIC ASSOCIATION

Membership About AEA Log I

Journals Annual Meeting Careers Resources EconLit EconSpark

Home › Journals › American Economic Journal: Applied Economics › January 2019 › Early Childhood Education by Television: Lessons from Sesame Street

Journals

American Economic Review

AER: Insights

AEJ: Applied Economics

About *AEJ: Applied*

Forthcoming Articles

Issues

Submissions

Early Childhood Education by Television: Lessons from Sesame Street

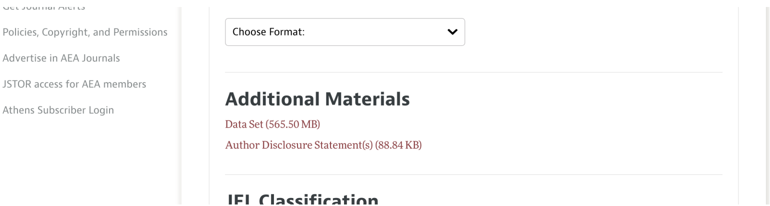
Melissa S. Kearney

Phillip B. Levine

AMERICAN ECONOMIC JOURNAL: APPLIED ECONOMICS
VOL. 11, NO. 1, JANUARY 2019
(pp. 318-50)

Download Full Text PDF (Complimentary)

Example 1: AEA journal


Get Journal Alerts
Policies, Copyright, and Permissions
Advertise in AEA Journals
JSTOR access for AEA members
Athens Subscriber Login
Choose Format:

Additional Materials

Data Set (565.50 MB)

Author Disclosure Statement(s) (88.84 KB)

JEL Classification

Source code

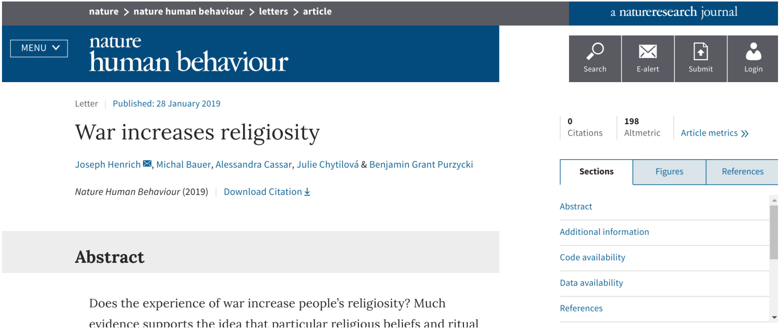
```
<ul id=’additionalMaterials’>
  <li>
    <a href=”/doi/10.1257/app.20170300.data”>
          Data Set  (565.50 MB)
     </a>
   </li>
 </ul>
```

Example 1: AEA journal

AEJ: Applied Economics

- **Supplementary data** as ZIP file
- **Accessibility method**: download (obvious)
- **Accessibility conditions** or **license**: none stated (but in fact, Copyright with “all rights reserved”!)
- **Persistence**: assumed to be “permanent” because on journal website

Example 2: Nature journal



Letter | Published: 28 January 2019

War increases religiosity

Joseph Henrich ✉, Michal Bauer, Alessandra Cassar, Julie Chytilová & Benjamin Grant Purzycki

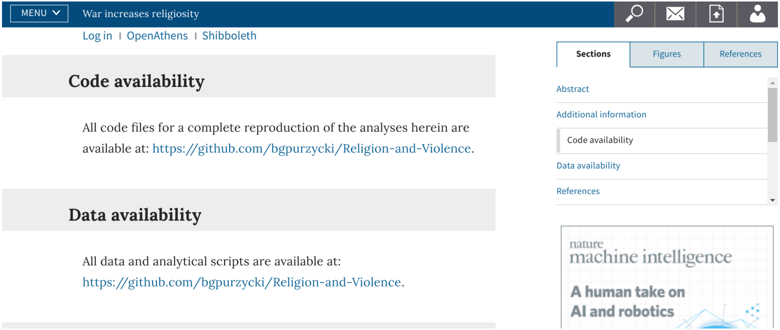
Nature Human Behaviour (2019) | Download Citation ⤓

0 Citations | 198 Altmetric | Article metrics »

Abstract

Does the experience of war increase people's religiosity? Much evidence supports the idea that particular religious beliefs and ritual

Example 2: Nature journal



MENU
War increases religiosity
Log in | OpenAthens | Shibboleth
Sections
Figures
References
Abstract
Additional information
Code availability
Data availability
References
nature machine intelligence
A human take on AI and robotics

Code availability

All code files for a complete reproduction of the analyses herein are available at: https://github.com/bgpurzycki/Religion-and-Violence.

Data availability

All data and analytical scripts are available at: https://github.com/bgpurzycki/Religion-and-Violence.

Source code

```
<section aria-labelledby=”data-availability”>
 <div class=”...” id=”data-availability-section”>
  <h2 class=”...” id=”data-availability”>Data avail
   <div class=”...” id=”data-availability-content”>
    <p>All data and analytical scripts are availabl
    <a href=”https://github.com/bgpurzycki/Religion
    https://github.com/bgpurzycki/Religion-and-Viole
    </p>
   </div>
 </div>
</section>
```

Example 2: Nature journal

Nature Human Behavior

- **Supplementary data** as (manually!) linked Github archive
- **Accessibility method**: unknown, but download presumed
- **Accessibility conditions** or **license**: none stated on journal website, and in fact, none stated on Github site. Therefore: Copyright with “all rights reserved”!
- **Persistence**: none stated, but use of Github is troublesome, as deletion is nearly instantaneous (at whim of author) and permanent

Current Metadata is Problematic

For **easily** accessible data

- Unstructured
- Opaque
- Leads to imperfect, unreliable, failing replications

Current Metadata is Problematic

For **difficult to access** data

- Highly unstructured (*prose*) or inexistant
- Opaque
- Leads to imperfect, unreliable, failing replications

Mission of journals: Better transparency

metajelo

metadata (package) for
journals to support
external
linked
objects

Basic motivation

Publication-oriented

Provide journals with the metadata needed to assess the robustness and reliability of supporting materials.

Minimal information

Requests no more information than (in theory) is currently being requested from authors.

- Name
- Location
- Accessibility (conditions, license)
- Persistence

Basic motivation

Extensive re-use

Leverage existing metadata *schemas* and *infrastructure* as much as possible

- Re-use of DataCite and re3data metadata elements
- Overlapping elements map into Dublin Core, CrossRef, DDI

Structure of `metajelo`

Each package is a linkage record

- conceptually models a linkage between a publication and its supplementary materials
- A record has an **identity** (Digital Object Identifier (DOI)), a **date created**, a **last modified date**, and the **identity** (DOI) **of the research objects** (papers) that are associated with the supplementary products
- Then an unlimited number of `supplementaryProducts`

Structure of `metajelo`

Each `supplementaryProducts` has

- an identifier,
- a description of its **type**,
- linkages to full metadata available elsewhere that fully describes the product.
- an associated location block (institutional archive)
- the set of possible policies (**access, license, preservation**), with a boolean designation flagging the relevant policy for the particular object
- Each policy instance structured to allow for verbatim answers if necessary

Detailed Structure

Full annotated schema

github.com/labordynamicsinstitute/metajelo

About that *infrastructure* . . .

Shortcoming of existing metadata

Why not directly collect the information from *infrastructure*

Much information lacking from current infrastructure

Paper goes into detail about the **failures of the current infrastructure** to provide information even when the objects are registered by knowledgeable institutions, and despite the ability to do so **within existing metadata schemas.**

How to generate a `metajelo` package?

Enable authors

Develop an application to allow users to generate a `metajelo` package

- Leveraging infrastructure where possible
- Querying the user where necessary

Portability

Once created, a `metajelo` package should be of use at multiple journals: saved locally

How to generate a `metajelo` package?

Encourage data providers

All information can be provided by data providers in a static format

- Generate once, deposit on website
- (optional) leverage to display suggested data citations

How to **use** a `metajelo` package?

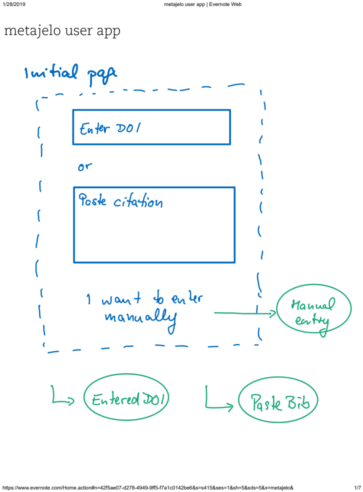
Enable journals

The information should be leverable with minimal modifications to current journal infrastructure

- Simplest: attach `metajelo` package, leverage CSS and JS to display contents
- More robust: ingest in journal management system

Short-term implementation can be made, without preventing future robust implementation

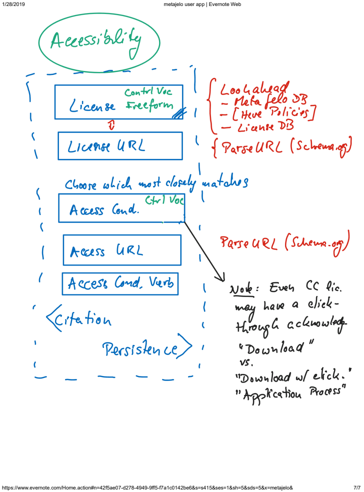
Sketch of a user-facing app



1/28/2019 metajelo user app | Evernote Web

metajelo user app

https://www.evernote.com/Home.action#n=42f5ae07-d278-4949-9ff5-f7a1c0142be6&s=s415&ses=1&sh=5&sds=5&x=metajelo& 1/7

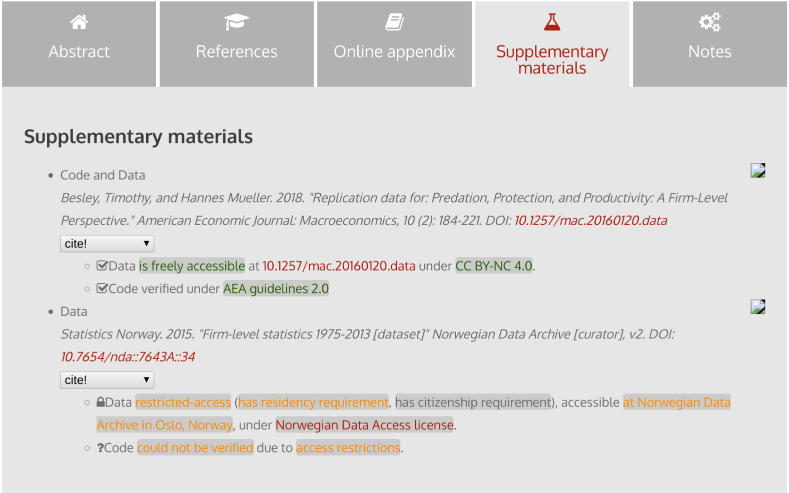


1/28/2019 metajelo user app | Evernote Web

Note: Even CC lic. may have a click-through acknowledg. "Download" vs. "Download w/ click." "Application Process"

https://www.evernote.com/Home.action#n=42f5ae07-d278-4949-9ff5-f7a1c0142be6&s=s415&ses=1&sh=5&sds=5&x=metajelo& 7/7

Sketch of a website



Supplementary materials

- Code and Data

  Besley, Timothy, and Hannes Mueller. 2018. "Replication data for: Predation, Protection, and Productivity: A Firm-Level Perspective." American Economic Journal: Macroeconomics, 10 (2): 184-221. DOI: 10.1257/mac.20160120.data

  cite! ▾

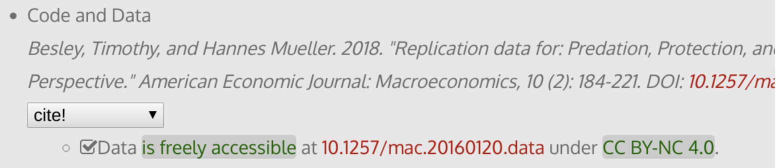
  - ☑Data is freely accessible at 10.1257/mac.20160120.data under CC BY-NC 4.0.
  - ☑Code verified under AEA guidelines 2.0
- Data

  Statistics Norway. 2015. "Firm-level statistics 1975-2013 [dataset]" Norwegian Data Archive [curator], v2. DOI: 10.7654/nda::7643A::34

  cite! ▾

  - 🔒Data restricted-access (has residency requirement, has citizenship requirement), accessible at Norwegian Data Archive in Oslo, Norway, under Norwegian Data Access license.
  - ?Code could not be verified due to access restrictions.

Sketch of a website



- Code and Data

  Besley, Timothy, and Hannes Mueller. 2018. "Replication data for: Predation, Protection, an
  Perspective." American Economic Journal: Macroeconomics, 10 (2): 184-221. DOI: 10.1257/ma

  - ☑Data is freely accessible at 10.1257/mac.20160120.data under CC BY-NC 4.0.

Next steps

Community input

We want to hear from a broad community about utility, extensions, etc.

Development of apps

We have started development of both **user-facing apps** and **journal-oriented toolkit**

Implementation

This is part of a broader strategy to improve transparency and reproducibility at the American Economic Association and in economics in general

Thank you

Merci

References

Camerer, Colin F., et al. 2016. “Evaluating Replicability of Laboratory Experiments in Economics”. *Science*: aaf0918. doi:`10.1126/science.aaf0918`.

Chang, Andrew C., and Phillip Li. 2017. “A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time”. *American Economic Review* 107 (5): 60–64. doi:`10.1257/aer.p20171034`.

— . 2015. *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ”Usually Not”*. Finance and Economics Discussion Series 2015-83. Board of Governors of the Federal Reserve System (U.S.) `https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf`.

Chemin, Matthieu, and Etienne Wasmer. 2017. “Erratum”. *Journal of Labor Economics* 35 (4): 1149–1152. doi:`10.1086/693983`.

— . 2009. “Using Alsace-Moselle Local Laws to Build a Difference-in-Differences Estimation Strategy of the Employment Effects of the 35-Hour Workweek Regulation in France”. *Journal of Labor Economics* 27 (4): 487–524. doi:`10.1086/605426`.

Collaboration, Open Science. 2015. “Estimating the Reproducibility of Psychological Science”. *Science* 349 (6251): aac4716–aac4716. doi:`10.1126/science.aac4716`.

Dewald, William G, Jerry G Thursby, and Richard G Anderson. 1986. “Replication in Empirical Economics: The Journal of Money, Credit and Banking Project”. *American Economic Review* 76 (4): 587–603.

Fanelli, Daniele. 2018. “Opinion: Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?” *Proceedings of the National Academy of Sciences* 115 (11): 2628–2631. doi:`10.1073/pnas.1708272114`.

Höffler, Jan H. 2017a. “Replication and Economics Journal Policies”. *American Economic Review* 107 (5): 52–55. doi:`10.1257/aer.p20171032`.

— . 2017b. “ReplicationWiki: Improving Transparency in Social Sciences Research”. *D-Lib Magazine* 23 (3/4). doi:`10.1045/march2017-hoeffler`.

Klein, Richard A., et al. 2014. “Investigating Variation in Replicability: A “Many Labs” Replication Project”. *Social Psychology* 45 (3): 142–152. doi:`10.1027/1864-9335/a000178`.

References

McCullough, Bruce D., Kerry Anne McGeary, and Teresa D. Harrison. 2006. "Lessons from the JMCB Archive". *Journal of Money, Credit and Banking* 38 (4): 1093–1107. `http://ideas.repec.org/a/mcb/jmoncb/v38y2006i4p1093-1107.html`.

McCullough, Bruce D., and Hrishikesh D. Vinod. 2003. "Econometrics and Software: Comments". *Journal of Economic Perspectives* 17 (1): 223–224. `http://EconPapers.repec.org/RePEc:aea:jecper:v:17:y:2003:i:1:p:223-224`.

Müller, Dana, and Joachim Möller. 2019. "Giving the International Scientific Community Access to German Labor Market Data: A Success Story". In *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*, ed. by Nuno Crato and Paolo Paruolo, 101–117. Springer International Publishing. ISBN: 978-3-319-78461-8. doi:`10.1007/978-3-319-78461-8_7`.

Olivier, Godechot. 2016. *L'Alsace-Moselle peut-elle décider des 35 heures ?* Notes et documents 2016-04. OSC.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma. 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals". Ed. by Dmitri Zaykin. *PLoS ONE* 8 (6): e67111. doi:`10.1371/journal.pone.0067111`.

Stodden, Victoria, Jennifer Seiler, and Zhaokun Ma. 2018. "An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility". *Proceedings of the National Academy of Sciences*: 201708290. doi:`10.1073/pnas.1708290115`.

Stodden, Victoria, et al. 2016. "Enhancing reproducibility for computational methods". *Science* 354 (6317): 1240–1241. doi:`10.1126/science.aah6168`.