Preregistration

Compliance of funder requirements of open access publication preregistration

Henrik Danielsson[1], Gustav Nilsonne[2,3], Lovisa Österlund[1], Johanna Nählinder[1]

[1] Linköping University

[2] Stockholm University

[3] Karolinska Institutet

04. January 2019

Study Information

Title	Compliance of funder requirements of open access publication preregistration
Research questions	The main research question of this project is: how does the introduction of open access publishing mandates by grant agencies affect the degree of open access publication? Additionally, we aim to describe trends in open science publication of Swedish research, with respect to fraction of open access publication and types of open access.

Hypotheses

- H1a: The introduction of open access mandates will be associated with an increase in open access publishing.
- H1b: The introduction of open access mandates will be associated with complete compliance of open access publishing (100%).
- H2: The introduction of open access mandates will be associated with an change in type of open access publishing.

Sampling Plan

Data sources: + We will collect bibliometric information and funding information from web of science via their API. + We will collect open access status of published articles from Unpaywall via the R package 'roadoi' (Jahn 2018). + We will collect funding information from an additional source, Dimensions, in case that we get access to their API.

Inclusion criteria: + Swedish affiliation for the reprint author. If reprint author is missing, then affiliation for the first author is used. This is because every funding agency has different requirements and to make the project feasible we limit ourselves to Swedish funding agencies where we know when different open access requirements were introduced. Swedish affiliation is defined by "Sweden" or "Sverige" in the affiliation information for the article as defined above.

- Data will be collected for the following publication years 2008 to 2017. This peridod of time will cover the introduction of open access policies by major funders. The larger grant agencies in Sweden who has introduced an open access requiremnet are the following:

Vetenskapsrådet - The Swedish research agency, starting from grants awarded 2010.

Forte - The Swedish research agency for health, working life and welfare, starting from grants awarded 2012

Formas - The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, starting from grants awarded 2010

Riksbankens Jubileumsfond, starting from grants awarded 2010

Matching procedure: + The link between articles and funding information is the article doi.

- The funding information is free text in Web of Science. We will identify relevant funders with regular expressions. We will search for grant numbers and use that to indetify which year the grant was awarded. If no year is identified, we will assume that the grant was awarded 2 years before publication.

Based on the information above, articles with Swedish affiliation will be divided into three groups. The first group will have funding information matching the grant agencies above, the second group will have other funding information than the grant agencies above, and the third group has no funding information.

Assumptions:

- Not all articles are published in WoS-journals. On the contrary, there is a rather large difference between academic fields. We are aware of this difference. . . .
- Some articles do not have doi-identificators. The share of articles with doi-identificators have risen over time. We assume that articles with doi-identificator follow the same pattern as articles without doi-identificator.
- Some articles do not have funding information, although they are obliged to.
- We assume that articles incorrectly without funding information follow the same pattern as articles with funding information.
- Most articles are co-authored. Our study will identify all articles with a Swedish reprint author. It is plausible that research awarded to a Swedish researcher will have a Swedish reprint affiliation, but this is not 100% sure.

- We assume that the affiliations are correct, i.e. that the word “Sweden” or “Sverige” is correctly used in affiliations
- OA status of articles change over time. An article that was not OA when published can become OA later. The other way around is not plausible but this means that there is a bias towards OA in our data. The control group comparison will solve with this problem partially.

Existing data	**Registration prior to analysis of the data**. As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.
Explanation of existing data	All data are available, but the combination of data from different sources is not. To avoid harking, one of the authors searched the first 200 entries from web of science to use as pilot data. On the pilot data, two other authors could optimize the regular expression to find different spellings of the grant agencies. Pilot data was also used to make sure that the coverage of funding information was sufficient (need a number?) in Crossref.
Data collection procedures	See above.
Sample size	The sample size will be all data that is available in the combination of data sources above.
Sample size rationale	The data is limited by when the requirement to publish open access by the main Swedish grant agencies was introduced. We collect data from 2 years before that was introduced to establish a baseline of degree of open acceess publications. Then we use all data up to 2017 to make sure that we have data for whole years. Data is also limited to one data source for each type of data. We have chosen the data sources with the best coverage of data that we are interested in by doing pilot searches. Web of Science had the best coverage of bibliometric information that is available via an API (Google Scholar has better coverage, but no easy way to get the data). The exception from one data source is funding information where we will use data from Web of Science that we know is sufficient. We will also use data from Dimensions, but we have not access to their API yet, so we have noy been able to evaluate their coverage. The sample of articles is divided into three groups: 1. have funding info from the

main grant agencies (who has OA demand), 2. have other funding info, and 3. have no funding info. The comparison between group 1 and 2 is our main comparison, but group 3 is also included as a comparison. Group 3 could lack funding info for many reasons and is therefore a weaker comparison than group 2, but it is possible that it is a relatively large proportion of the articles and therefore it is important to include.

Stopping rule No.

Variables

- article publication year
- article open access status
- article funding information
- reprint author affilliation
- article doi
- funding agency startyear of open access publication requirement
- funding agency type of open access publication requirement

Manipulated variables Not applicable.

Measured variables Not applicable.

Indices OA status. Green, Bronze, other categories... Specify!

Use definition from "The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles" av Piwowar, Priem, Larivière , Alperin, Matthias, Norlander, Farley, West, Haustein? citat: "Classifications We classify publications into two categories, OA and Closed. As described above, we define OA as free to read online, either on the publisher website or in an OA repository ; all articles not meeting this definition were defined as Closed. We further divide the OA literature into one of four exclusive subcategories, resulting in a five-category classification system for articles: Gold : Published in an open-access journal (as defined by the DOAJ). Green : Toll-access on the publisher page, but there is a free copy in an OA

repository. Hybrid : Free under an open license in a toll-access journal. Bronze : Free to read on the publisher page, but without a license. Closed : All other articles, including those shared only on an ASN or in Sci-Hub. These categories are largely consistent with their use throughout the OA literature, although a few clarifications are useful. First, we (like many other OA studies) do not include ASN-hosted content as OA. Second, categories are exclusive, and publisher-hosted content takes precedence over self-archived content. This means that if an article is posted in both a Gold journal and an OA repository, we would classify it as Gold, not Green. Put another way, publisher-hosted content can "shadow" archived articles that would otherwise be Green. This definition of Green ("available in a repository but not available from the publisher") is often used in the OA literature (including by Steven Harnad, the coiner of the Green and Gold terms [Harnad et al., 2008]), but this usage is not unanimous. Some studies allow a given article to be both Gold and Green; compared to these, our classification system does undercount Green. Finally, we add novel subcategory, Bronze. Bronze shares attributes of Gold and Hybrid; like both, Bronze OA articles are publisher-hosted. Unlike Gold OA, Bronze articles are not published in journals listed as OA in the DOAJ. Unlike Hybrid, Bronze articles are Gratis OA, carrying no license to extend reuse rights beyond simply reading. It is not clear if Bronze articles are temporarily or permanently available to read for free."

Design Plan

Study type **Observational Study**. Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.

Blinding Data will be collected by two of the authors. The affiliations and oainformation will be given codes. Then two of the other authors will analyze the data without being aware of the meaning of the codes.

Study design	The design of this observational study is a longitudinal study quasi-experimental design.
Randomization	No.

Analysis Plan

Descriptive analyses	We will tabulate and plot the frequency and proportion of OA articles by funder and year.
Statistical models	We will use an interrupted time series model. OA status will be the outcome. Predictors will be presence of policy (yes/no) and time since introduction of policy. The analysis will be performed as in Hardwicke et al. 2017 (http://rsos.royalsocietypublishing.org/content/5/8/180448, R code available at https://osf.io/y2meu/. Thus, we will fit a general linear model with a logistic link function.
Transformations	No.
Follow-up analyses	Follow-up analyses will be made to investigate if the main result is different depending on affiliation (which university in Sweden).
Inference criteria	Inferences will be based on p-values with alpha $<$. 05. Two-sided tests will be used so that inferences can be drawn if the introduction of open access mandates is associated with a decrease in open access publishing.
Data exclusion	No exclusion of data.
Missing data	The degree of missing data from the different data sources will be documented and reported.

Assumptions (optional)	No.
Exploratory analyses (optional)	No.
Analysis scripts (optional)	No.

Other

Other (Optional)	No.

References

@Manual{, title = {roadoi: Find Free Versions of Scholarly Publications via Unpaywall}, author = {Najko Jahn}, year = {2018}, note = {R package version 0.5.1}, url = {https://CRAN.R-project.org/package=roadoi}, }