LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

Vol. 23, no. 1 (2013) 48–79 | ISSN: 1435-5205 | e-ISSN: 2213-056X

Data Management in Scholarly Journals and Possible Roles for Libraries — Some Insights from EDaWaX

Sven Vlaeminck

German National Library of Economics/Leibniz Information Centre for Economics (ZBW)
s.vlaeminck@zbw.eu

Abstract

In this paper we summarize the findings of an empirical study conducted by the EDaWaX-Project. 141 economics journals were examined regarding the quality and extent of data availability policies that should support replications of published empirical results in economics. This paper suggests criteria for such policies that aim to facilitate replications. These criteria were also used for analysing the data availability policies we found in our sample and to identify best practices for data policies of scholarly journals in economics. In addition, we also evaluated the journals' data archives and checked the percentage of articles associated with research data. To conclude, an appraisal as to how scientific libraries might support the linkage of publications to underlying research data in cooperation with researchers, editors, publishers and data centres is presented.

Key Words: economics journals; data policies; linking research data and publications

Background and introduction

Empirical studies become increasingly important in many disciplines. This is also the case in economics, where a rising number of contributions to journals

Igitur publishing | http://liber.library.uu.nl/ | URN:NBN:NL:UI:10-1-114595

consist of empirical papers in which authors have collected their own research data or used external datasets for statistical analyses.

In economics, three major types of research data used in scientific papers can be distinguished:

- The most important are econometric studies in which researchers use datasets from multiple sources for verifying theoretical models by the methods of statistical analysis.
- A second type comprises simulations for gauging the behaviour of the economy under emerging conditions or to calculate distributions for statistics.
- A third field includes experiments in which test subjects are confronted with an (economic) challenge to solve. Depending on the results of these experiments, economic assumptions are made as to how stakeholders in economic markets behave.

Therefore research data in economics originates from different sources. In contrast to more empirically focused scientific disciplines, often the datasets used in economics are not collected and aggregated by the researchers themselves. Instead, researchers are using datasets that are part of the official statistics, thus have been collected by specialised research institutions[1] as well as proprietary datasets that are bought from commercially oriented companies (e.g. Thomson Reuters, Bloomberg). One exception is experimental research where researchers often compile their own datasets.

However, there have been few means to replicate the results of economic research within the framework of published journal articles and to verify the results claimed in such an empirical paper. This is not only unsatisfactory from a scientific point of view because replicability is a cornerstone of the scientific method; also on a political and social level, lack of replicability is a problem because political decisions often are justified by economic research.[2]

According to the literature there seem to be at least three principle reasons why economic research often is not replicable:

1. First and very important it is due to a lack of incentives for researchers to share their research data with the community. The academic reward system does not honour this time-consuming kind

of work — in sharp contrast to publications though (as Anderson, Greene, McCullough, and Vinod (2008) pointed out) *"[a]n applied economics article is only the advertising for the data and code that produced the published results*" (p. 5). Therefore a researcher in economics often feels that he or she might suffer a disadvantage if he or she does share his or her data, especially because potential competitors might use an interesting dataset for their own research, without acknowledging the creator of the data.

2. Secondly, economics journals rarely pledge their authors to provide the data and the code of computation of their analyses. Only a few years ago some economics journals just started to implement so called data availability policies[3], which (at least partially) mandated the availability of data and code.
3. A third reason is based on the hardly existing e-infrastructure for publication-related research data in economics.[4] Some journals have implemented data archives for their respective journals, but data availability is often not enforced. Also an overall infrastructure for publication-related research data is currently not yet available at specialized data centres.[5]

All aforementioned topics have been explored in the analysis phase of the project European Data Watch (EDaWaX[6]) that is funded by the German Research Foundation (DFG). Beside other tasks, EDaWaX analysed the data sharing practices among economists (Andreoli-Versbach & Mueller-Langer, 2013), the possibilities to host and store a publication-related data archive in European research data centres[7] and — and this is the purpose of this paper – the amount and quality of data availability policies in economic scholarly journals.

In this explorative study, we wanted to gain knowledge about how many journals in a defined sample are equipped with data availability policies, how these policies are structured, and what requirements authors are pledged to fulfil for complying with them.

Moreover, we wanted to find out the current practices of these journals with the goal of providing the best practices to the community. These findings and experiences of our analysis have been used to generate functional requirements for the current development of a pilot application for publication-related research data.[8]

Replications and data policies

Replication is a cornerstone of the scientific method as the US-economist B.D. McCullough (2009) lines out: "*[…] replication ensures that the method used to produce the results is known. Whether the results are correct or not is another matter, but unless everyone knows how the results were produced, their correctness cannot be assessed. Replicable research is subject to the scientific principle of verification; non-replicable research cannot be verified. Second, and more importantly, replicable research speeds scientific progress. We are all familiar with Newton's quote, 'If I have seen a little further, it is by standing on the shoulders of Giants.' […] Third, researchers will have an incentive to avoid sloppiness. […] Fourth, the incidence of fraud will decrease*" (p. 118f) . But what about the replicability of economics research and the amount of replication attempts in economics?

Replications in economics

According to many studies that have faced replications in economics, the amount of replications conducted is marginal (Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007; Hamermesh, 2007; McCullough & McKitrick, 2009; Evanschitzky & Armstrong, 2010). Also, researchers who systematically tried to replicate the results of economic articles often failed: Dewald, Thursby and Anderson (1986) attempted to replicate the results of 54 empirical papers and were able to replicate only two of them. Other attempts (McCullough, McGeary, & Harrison, 2006) showed almost the same results: only 14 out of 62 articles could be replicated. The same authors confirmed these findings two years later trying to replicate 117 articles succeeding only 7 times (McCullough, McGeary, & Harrison, 2008). Anderson *et al.* (2008) conclude: "*To date, every systematic attempt to investigate this question has concluded that replicable economic research is the exception and not the rule*" (p. 100).

The reason for these poor findings is directly connected to the lack of incentives for researchers to share "their" data and code: A recent paper published in the context of the EDaWaX project shows that only 2.05% of 488 empirical economists fully share their research data (Andreoli-Versbach & Mueller-Langer, 2013). Also the principle "publish-or-perish" seems to be an important component why economic research often is irreproducible. In the researchers' competition for permanent jobs, scientific careers and reputation, a scientist may perceive a strategic advantage in publishing the

results of his or her research while retaining the underlying research data and code (Mirowski & Sklivas, 1991; Anderson *et al.*, 2008). These theses seem to be evident. The motivation of researchers to act in this manner may stand to reason — but additionally the public has "financed" scientists for doing research work as well. One might argue — and that's what we do — that also the public has a right to verify and reuse the fruits of publicly funded research. Moreover, there is no doubt that concerning the progress of science, the process of acquiring important scientific resources is crucial. Scientific progress emerges because researchers may build on findings made by their predecessors.

At this point the journals in Economics come to the fore. Journals have a dominant position in the way researchers provide publication-related research data. According to the research of McCullough, McGeary and Harrison (2008) at least some of the top journals in economics have implemented efficient data policies for authors of empirical or econometric articles as well as for articles dealing with simulations or experiments.

It has been a long way to reach this point: As one of the first journals in Economics — The *Journal of Money, Credit and Banking (JMCB)* adopted a so-called "Replication Policy" in 1982. "Replication policies" are requiring authors to pledge to provide data (and sometimes code, too) to would-be replicators in case of upcoming requests. Dewald *et al.* (1986) showed that these kinds of policies are insufficient. In practice, many studies observed that authors often failed to honour these policies — they are simply ignoring them (McCullough & Vinod, 2003).

The major problem is that the incentives for authors to comply with policies that only rely on the honour system rather than requiring authors to provide the data and code are ineffective: "*The goals of the replication policies were incompatible with the incentive mechanisms implemented (or not) by the journals*" (McCullough *et al.*, 2006, p. 1094). Both in theory and in practice, "replication policies" do not work. Therefore, "replication policies" appear to be window dressing and not a sustainable attempt to enforce the availability of data and code.

A loophole out of the irreproducible research was found with the implementation of mandatory data availability policies that meet the tenet of Gary King's replication standard. King suggests that replications should be able

without the help of the author (King, 1995). Since 2000 some economic journals, including *The American Economic Review* (AER, n.d.)[9], have adopted data availability policies — slowly realizing the ineffectiveness of replication policies. The AER tightened its policy in 2004 towards a mandatory data and code archive after McCullough and Vinod (2003) attempted to replicate all the empirical articles in a single issue of the AER and almost half of the authors failed to honour the replication policy. Some other top journals soon followed the AER's lead.

Research conducted by Glandon (2010) showed that these new policies are suitable for replication purposes: Glandon believed that a total of 31 (79%) out of 39 investigated articles published in the AER were replicable without contacting the editors.

Requirements for data availability policies to enable replications

These comparatively satisfying results could be obtained because the editors of the AER seemed to have learned some lessons. For our project it was important to identify some core requirements for data policies that will facilitate replications. Therefore we consulted several research papers (Dewald *et al.*, 1986; King, 1995; McCullough, 2007; Anderson *et al.*, 2008; McCullough *et al.*, 2008) and used the recommendations we found in the papers as a basis for analysing and assessing the suitability of data availability policies of economic journals in our study.

1. A data availability policy must be mandatory. (Dewald *et al.*, 1986)
2. Besides requiring authors to provide datasets, also the provision of code, programs and detailed descriptions of the data (data dictionaries) are required. Authors have to submit the original data from which the final dataset is derived and all instructions/code necessary to achieve the final results of computation. A README file should list all submitted files with a description of each and indicate which programs correspond to what results in the paper. (McCullough, 2007; McCullough *et al.*, 2008)
3. All required files have to be provided to the journal's editors prior to the publication of an article. (Dewald *et al.*, 1986)
4. All submitted data and files (if not confidential or proprietary) must be made publicly available to interested researchers. (King, 1995)

5. A data policy has to have a procedure in place which allows interested readers to replicate proprietary or confidential datasets in principle, even if the raw dataset cannot be submitted to the journal due to juridical reasons.
6. The journal should have a replication section and encourage the readers to use it for conducting replications of previously published results. Such a section is important because authors must know that journals will publish the results of failed replications (Anderson *et al.*, 2008). Thereby authors will scrutinize their data. The submission to an archive of badly documented data or even junk will most likely be prevented.[10]
7. All data has to be submitted in the ASCII-format or at least in open formats that facilitate the long-term preservation of data as well as the interoperability of the data and code. The code submitted should call these ASCII files. (McCullough, 2007)
8. The version of the operating systems and the software used for obtaining the results should be indicated, because results may seriously differ depending on the used versions of the operating system and software package. (McCullough & Vinod, 2003)

These eight recommendations were used as theoretical background for the analysis of the data policies' quality we found in our sample.

The study

Sample and methodology

For building our sample of journals for our study, we chose the list of 150 journals that have been analysed by the German economists Bräuninger, Haucap and Muck (2011) regarding their relevance and reputation. This list (we will refer to it as the BHM list) comprises the most important economic journals as well as a bigger part of the economic journals published in Germany, Austria and Switzerland. This sample offers many advantages for our analysis, because it enables the comparison of journals published in the German speaking area with the international ones. Furthermore, this sample offers the possibility to compare the rankings of journals with data availability or replication policies to journals without data policies and to determine some other characteristics of them.

In accordance with the focus of the study, we added four more journals with data availability policy that were not part of the sample of Bräuninger *et al.* but were analysed by McCullough (2009). 13 journals were removed afterwards from the sample because it was apparent that these journals are focused on discussing solely economic policy or theoretical research. Altogether a sample of 141 journals remained for our analysis. By having included many of the top journals, we assume that our sample is rated higher than the average in economics.

In our sample journals of all major scientific publishers were included: The largest concentration of analysed journals were published by Elsevier and Wiley-Blackwell (both 23.4%), followed by journals published by Springer (12.8%) and Oxford University Press (5.7%). Almost all journals were subscription journals with the exception of a single open-access journal. Three-fourths (75.2%) of the journals in our sample are present in Thomson Reuters' Journal Citation Reports 2010 (Thomson Reuters, 2011) (abbreviated as JCR in the following) and almost 96% are included in the Handelsblatt Ranking Volkswirtschaftslehre (n.d.) for 2010 — both are very important rankings in economics.[11]

Our analysis started with a desktop research: Both the publisher's and the editor's website of the journals were examined precisely (we did *not* check the printed edition) to evaluate how many of these journals are equipped with a data policy.[12]

To verify the thesis that journals with data availability policies are often among the top rated journals, as McCullough *et al.* (2008) outline, we also examined how these journals are ranked compared to the ones without data policies. For that purpose we compared means and median of the whole sample as well as for subsamples. In addition to testing the theorem that these journals are highly ranked, we conducted regression analyses for clearing potential coherences between the journal's ranking and the availability of a data policy for this journal.

We also qualitatively analysed the policies along the proposed recommendations listed above and summed up some conclusions. We evaluated these policies on the basis of the announcements within the policy. The implementation of these policies in practice must not necessarily accord with these announcements. We therefore investigated the journals' data archives respectively their

websites in order to check how many articles in two single issues are accompanied by research data/code/programs or descriptions.

Descriptive results

Analysing 141 economic scholarly journals (Figure 1), we were able to find 29 journals (20.6%) that are equipped with data availability policies. Another 11 (7.8%) had a replication policy implemented. Looking for the publishers of the 29 journals with data availability policy, we noticed that in total numbers the majority was published by Wiley-Blackwell (6) and Elsevier (4). But when we compared the total number of all single publisher's journals to the portion of journals with data availability policies in our sample, university presses (e.g. Cambridge University Press) or association presses (e.g. from the American Economic Association) are equipped with high to very high portions of journals owning data availability policies.

Evaluating the ranking and the impact factor of these journals (Figure 2) we found out that the average impact factor of journals equipped with data availability policies is rated 0.43 points higher in Thomson Reuters JCR (2011)

Fig. 1: Data policies of Economics journals in our sample.

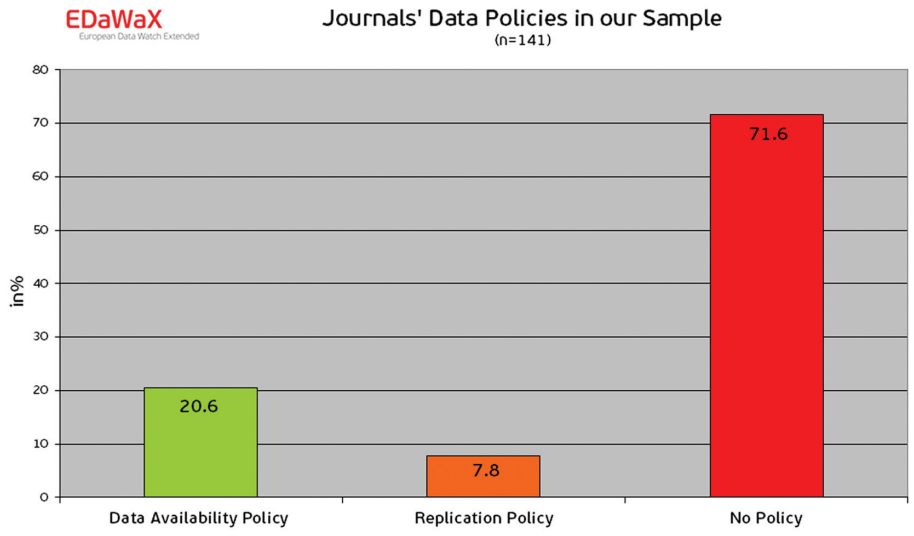
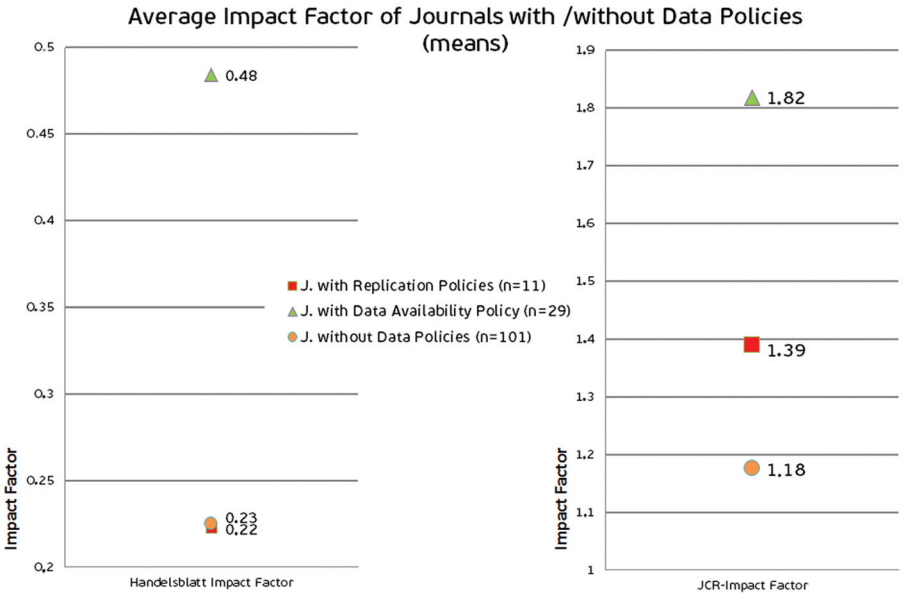


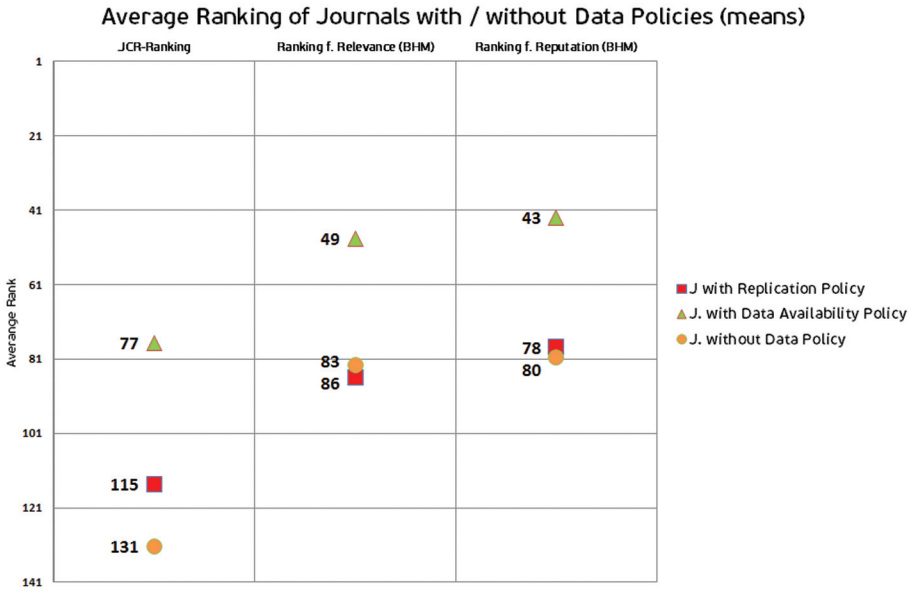
Fig. 2: Average Impact Factor (rounded) of journals with data availability policies, with replication policies and without data policies.



compared to the average impact factor of journals with a replication policy and even 0.64 points higher compared to journals without a data policy. For the Handelsblatt Ranking Volkswirtschaftslehre (n.d.) for 2010 we ascertained that these journals are ranked still 0.26 points higher than the average of journals with replication policy and 0.25 points higher than journals without a data policy.

When comparing the average ranking of journals with data availability policy to those without a data policy (Figure 3), we detected that those with a data availability policy are ranked on average almost 55 places higher in Thomson Reuters JCR (2011), 34 places higher in the ranking of Bräuninger *et al.* (2011) for relevance and even 37 places higher for reputation. Compared to journals equipped with a replication policy, journals with a data availability policy still are ranked 38 places higher in the JCR, 37 places higher in BHM's ranking for relevance and 35 higher for reputation. When conducting a regression analysis we found an average significant correlation (0.296 to 0.4) between the higher ranking of a journal and the existence of a data availability policy.

Fig. 3: Average ranking of journals with data availability policies, with replication policies and without data policies.



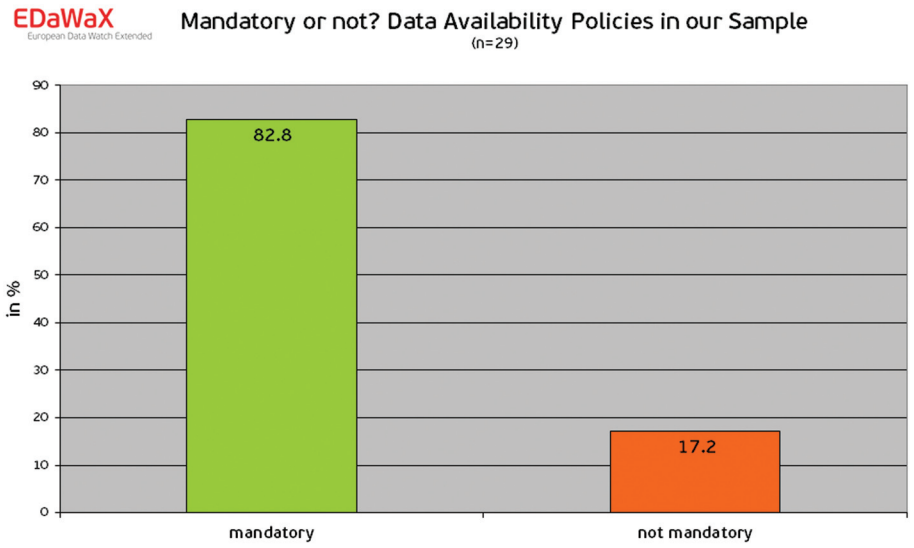
Evaluating the quality of data availability policies

In this chapter we summarize our findings regarding the quality of data availability policies. The quality of these data policies was examined along the eight mentioned requirements in chapter 2.2. The quality and extent of the data availability policies in our sample differed massively: some were just a few sentences long, others comprise several printed pages.[13] But the extent of a policy is not necessarily a proof of good quality. We discovered good examples that are no longer than one-third of a page.

a) Mandatory data availability policies

A policy was evaluated as mandatory when the policy pledged authors to provide data. That means if a policy contained one of the phrases "requirement/condition for publication"/"must be"/"publish papers only if"/"will be expected" in the context of data submission. Consequently a policy was evaluated as not mandatory when one of the phrases "should be/offered the possibility"/"authors are encouraged" were found in the policy's text.

Fig. 4: The extent of mandatory data availability policies in ours Sample.



Following these criteria 82.8% (24) of the 29 analyzed journals with data availability policies were evaluated to be mandatory (Figure 4).

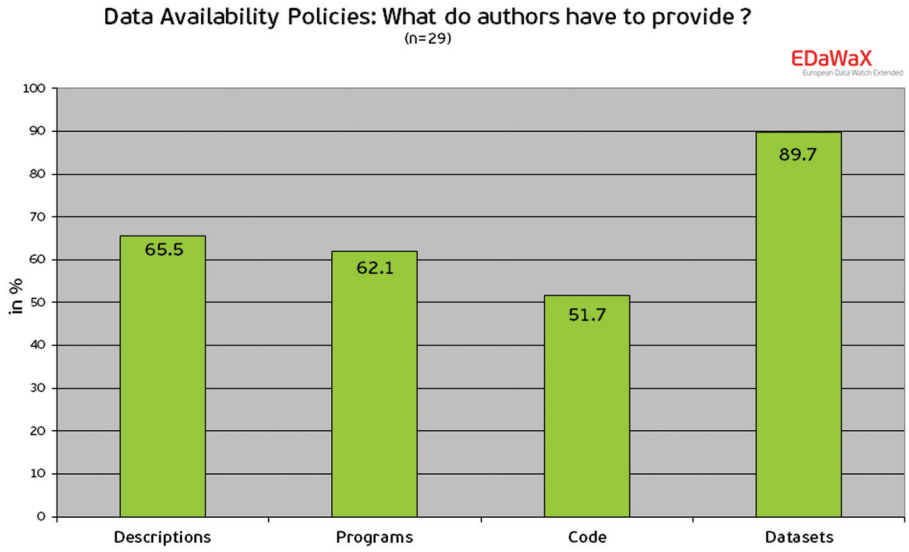
b) Data and files that have to be submitted to the journal

For obtaining results in this section we checked the specifications of the policies (Figure 5). We found out that 26 of 29 policies (89.7%) pledged authors to submit datasets used for the computation of their results.[14]

The submission of (user-written) programs used e.g. for simulation purposes are mandatory for 62% of the policies but only half of them mandated authors to provide the code of their calculations. Due to the importance of code for replication purposes this percentage may be considered as low.

Descriptions of the data submitted and instructions on how to use the single files for replications are obligatory for 65.5% of the policies. The quality of these descriptions differs from very detailed instructions to a few sentences only that might not really help would-be replicators. This finding points out that there is currently no consensus and no standard among economists on how detailed these descriptions have to be and what they have to cover.

Fig. 5: Percentage of journals with data availability policies requiring datasets, code, (user-written) programs and descriptions.



Therefore the quality of descriptions depends entirely on the weal and woe of a single author, which is the opposite of a standard.

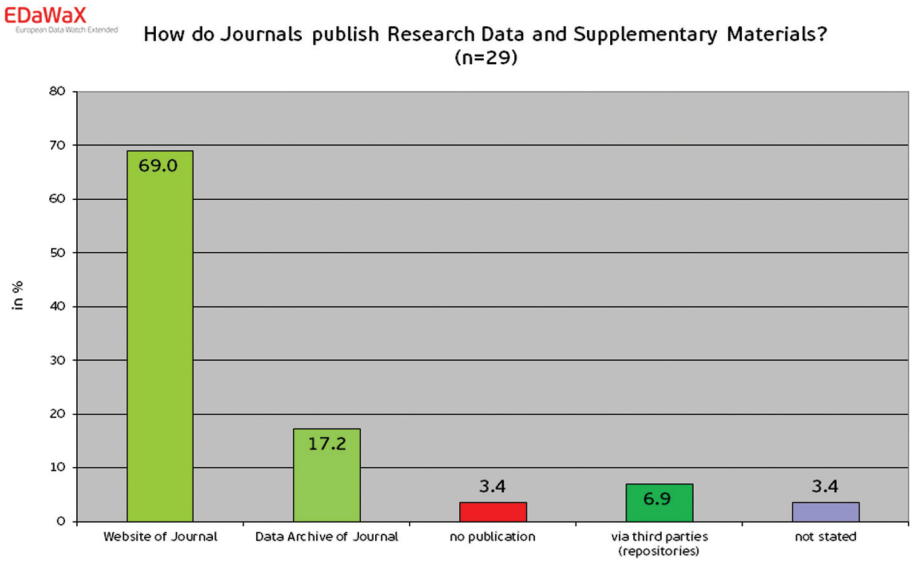
c) Submission of data prior to publication

While examining the data policies in regard to a defined point of time when data has to be submitted, we discovered that almost 90% of the policies required their authors to provide all data prior to publication. A single journal (*The Journal of Law, Economics & Organization*) offered the possibility for authors to provide data within three (!) years after publication.

d) Provision of publication-related research data

In the course of our analysis we noticed that the primary way for publishing publication-related research data and code (Figure 6) was via attaching files to the article on the journal's website: 69% of the journals mentioned in their data policy to use this way for providing research data. The most common way is to attach a zip-file to the article (this zip-file most often is available in the supplementary information section). An interested researcher may download the zip-container and extract the content. When examining some of these zip-files

Fig. 6: Provision of publication-related research data by economic journals equipped with a data availability policy.



the diversity of formats and files within these zip-containers underlines why detailed descriptions are crucial for the effort of replication attempts.

Another 17.2% of these journals used a special website for providing research data. Normally these websites list all issues of a journal and all articles of the single issue. Where datasets (and code) have been provided, a link for downloading the data is available.[15] Other journals used Dataverse[16] for their data archive — in our opinion a very useful practice. Dataverse offers numerous functionalities for searching, citing, downloading and even analysing research data — especially compared to the practice of simply attaching a zip-file to an article.[17]

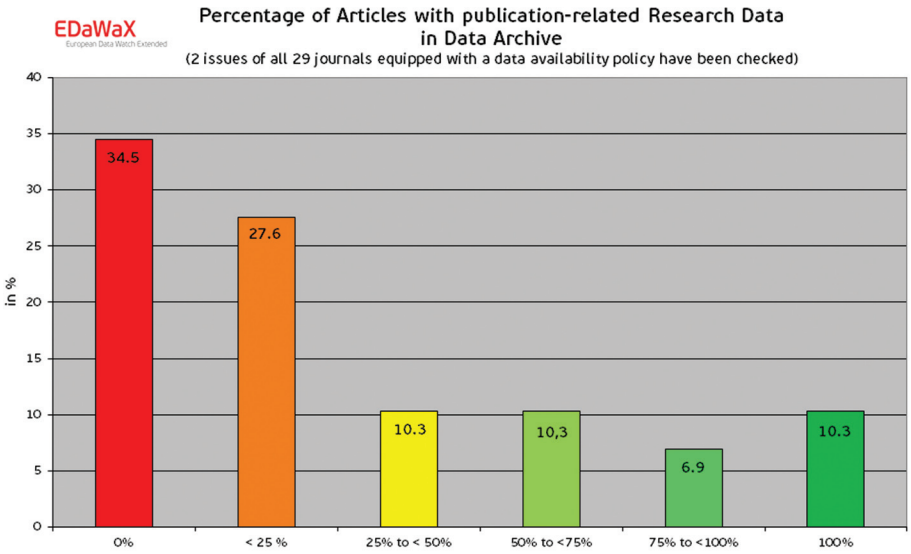
A special way to provide research data is conducted by two cross-disciplinary journals of our sample: Nature and Science are using discipline-specific data repositories for providing datasets and code, descriptions and other files. This is a very useful way to disseminate publication-related research data and code, because the archive is managed by subject specific specialists, who know best what is necessary for a proper documentation of data and code. This approach

also facilitates the provision of data and code, especially for editors of scholarly journals: the archive is managed externally, and the editors only have to present the URL to these data and materials in their journals.

A single journal of our sample does not provide data at all — the files provided by the authors are used for internal evaluation by specialised referees only.

However, the statements within the data policies are just one side of the coin. Besides examining the text of these policies we were also interested in the current practices of these economic journals. Do really all of them have a data archive in place? Is the data policy enforced so that almost every (empirical) article is equipped with its underlying research data and code? We investigated the journals' data archives (respectively the supplements of all articles) for the issues 1/2010 and 1/2011 and checked how many articles provide datasets, code etc. (Figure 7). We did not categorize the focus of the articles, so that our investigation is not a systematic approach for analysing these data and code archives but a snapshot.

Fig. 7: Percentage of articles that are associated with accompanied research data/code and/or descriptions.



Nevertheless, the results we obtained suggest that the current practice paints a far different picture than the warm words stated within the data policies suggest: only 19 out of 29 journals (65.5%) with a data availability policy had something that may be called (with reservations) an archive. And even for the remaining 19 journals we have to state that the archives are filled highly differently: While some of the journals are taking their policies quite seriously, (e.g. *Brooking Papers, Nature, Science, American Economic Journal: Applied Economics, Proceedings of the National Academy of Sciences*) many others seem to be relatively apathetic about them: We found eight journals with a data availability policy where less than 25% of all articles were equipped with anything related to the data policy — in four of these cases even less than 10%. In the light of these findings the portion of functional data availability policy considerably diminishes.

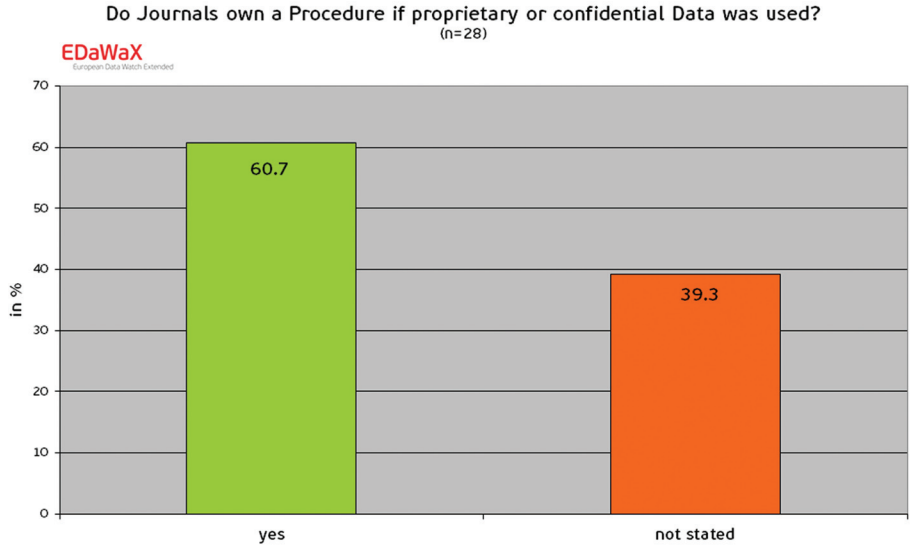
e) Defined procedure in case of exceptions to the data policy

As mentioned above, many data sources in economics derive from companies or research data centres and are therefore proprietary or even confidential as in the case of micro data. Because research using these sources is common, a defined procedure in case of exceptions to the data policy is relevant for the general ability to replicate even results of research conducted with those data sources. In the course of our research we found out (Figure 8) that 72.4% of the data availability policies allowed exceptions to their data policy (one journal explicitly did not permit exceptions). But only 60.7% of all of these journals had a procedure in place about how authors have to conduct in the case of proprietary or confidential data. In such cases authors often still have to provide code and descriptions. In addition, they have to state how to obtain data in principle (e.g. name and address of the company/institution, contact details, version of the dataset, …).

f) Replication sections

There are only very few economics journals equipped with a replication section — and none of them has been part of our sample.[18] One of these journals is the *Journal of Applied Econometrics* (JAE), which has introduced a replication section in January 2003. This section was initially devoted exclusively to the issue of replication of empirical results published in papers of the *Journal of Applied Econometrics*. Surprisingly the JAE decided to extend the coverage of the section and also invites authors to submit replication attempts

Fig. 8: Percentage of journals owning a defined procedure in cases where authors have used proprietary or confidential data for their research.



for empirical research that has been published in the following additional journals (Pesaran, n.d.):

- Econometrica
- American Economic Review
- Journal of Political Economy
- Quarterly Journal of Economics
- Review of Economics and Statistics
- Review of Economic Studies
- Journal of Econometrics
- Journal of Business and Economic Statistics
- Economic Journal

This is a surprising result: within our sample of 29 journals we were not able to find a single journal that explicitly claims to have a replication section. Regarding the replication initiative by JAE, it is not clear whether their approach is coordinated with the other economic journals or not.

Instead of having a dedicated replication section, 6 of the 29 journals equipped with a data availability policy at least own a section for comments. This is especially the case for journals using Dataverse, because these comments are part of the features Dataverse offers. In principle it is possible to report failed replication attempts by using this comment section.

The absence of a replication section on the contrary does not imply that these journals do not publish replication studies, but in general published replication studies are rare among all journals we investigated.

g) Format specifications

In our sample only two journals (6.9%) made proposals regarding the formats of datasets, programs and descriptions. Both recommended the usage of plain ASCII (text) files. None of the other journals did make a statement on this topic. The journals that have adopted the data policy of the AER, e.g., are allowing any format "using any statistical package or software" (AER, n.d.). The only constraint is related to the README-file, which is often recommended to be in PDF- or ASCII-format.

h) Operating system and software used for generating results

In our full sample, we were not able to find any clear recommendations regarding the operating system used for the calculations. Also regarding descriptions of the software used for statistical analyses only the journals that have adopted the policy of the AER are declaring that the README-file should "*list […] all included files and document […] the purpose and format of each file provided*." (AER, n.d.). Detailed requirements were not stated.

Summary and conclusion

In summary, it can be stated that the management of publication-related research data in economics is still in its early stages. We were able to find 29 journals with data availability policies. At first glance that is much more than McCullough (2009) found some years ago. In the field of economics, editors and journals seem to be in motion. This seems to be a positive signal and it will be interesting to see whether and how this upward growth continues.

Also, the fact that a large portion of the analysed data availability policies are mandatory is useful and may be observed as a sign that editors consider the availability of research data to be important. Moreover, the finding that 90% of the journals are urging their authors to submit the data prior to the publication of an article shows that many of them have understood the importance of providing data at an early stage in the publication process. The fact that more than 60% of the journals define exact procedures for describing what kind of material has to be provided in the case of exceptions to the policy can also be read as a development towards the reproducibility of research conducted with proprietary or confidential data sets. Nevertheless, there is a need for improving the quantity of policies that define a procedure in case of proprietary or confidential datasets.

But this is just one side of the coin. The flip side is the amount of data policies that are merely window dressing. Part of these window dressers are all journals equipped with a replication policy. Many studies concluded that these policies do not work in practice — nevertheless they are still in use.

But of the 29 journals equipped with a data policy only half of them mandate the availability of data and code. If we take into account that even among journals with such a policy only slightly more than a third offers data (and even less code) for more than half of all papers we investigated, it seems obvious that only a small portion of journals really enforces the availability of research data and code. Therefore a lot of journals, even those with a data availability policy, seem to pay lip service to replicable research.

Among the journals with data availability policies we noticed that 10 out of these 29 used the data availability policy implemented at first by the *American Economic Review* (AER, n.d.). These journals either used exactly the same policy or a slightly modified version of it. In our opinion, this policy suits as best practice. The amenities of the AER policy comprise that

- the policy is mandatory,
- the journal provides policies for econometric papers, papers that are based on simulations as well as for experimental work,
- not only datasets are required to be made available, but also the code for computations, programs and a detailed README-file are mandatory parts of the submission,

- the policy has a defined procedure in case of granted exceptions to the policy for confidential or proprietary data,
- the AER pledges authors to provide all data prior to the publication of an article,
- the journal has a special website (a data archive) that provides the datasets and other files to interested readers (other journals with the same policy used even Dataverse),
- the journal mandates authors to describe the formats of the files they provided — and therefore some kind of information about the software used for computation.

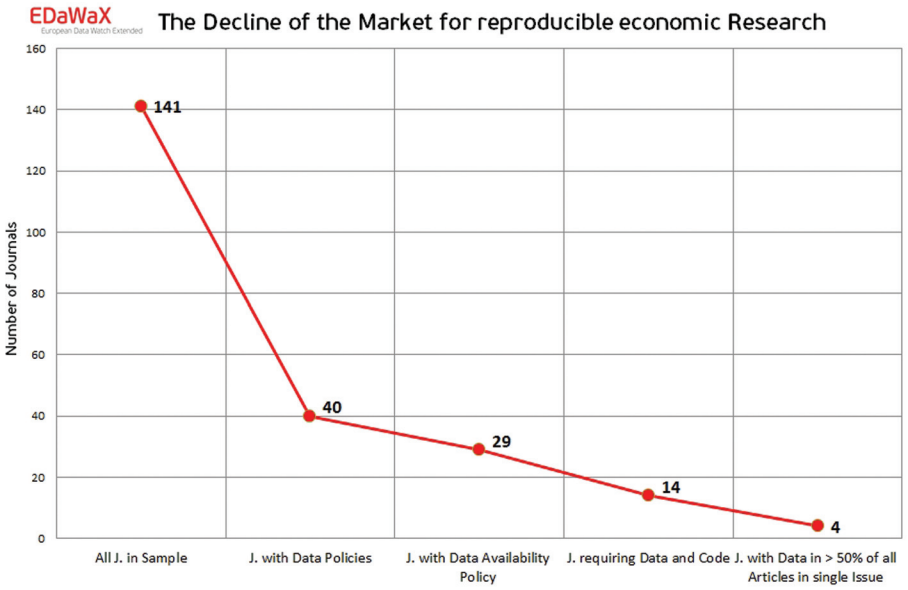
Although we are able to acknowledge some progress, it is still a small part of journals that are requiring their authors to provide the data and code they have used for analyses. Due to the fact that only half of the journals recommend the submission of code and only two–thirds mandate the authors to provide detailed descriptions and programs, this does not enable other researchers to 'stand on the shoulders of Giants'.

Especially checking the reality of data provision to would-be replicators was deflating (Figure 9): only 19 (65.5%) of the 29 journals actuate a data archive — which is a shattering result. And of these journals almost a quarter only had a humble percentage of articles with supplementary data.

In total we were able to find 4 journals that both mandate their authors to provide data plus code and that had at least every second article in one of the two issues we assessed equipped with accompanied data. This equates 2.8% of our full sample — not a glorious chapter of economic research. But even for these journals with a mandatory data and code archive we see both a growing demand for standardization as well as for the development of infrastructural components and additional features. The demand for standardisation is visible in the proliferation of accepted formats for research data that normally do not support interoperability or long-term preservation. Additionally other metadata — as for example the operating system and the version of software used for computation — are missing all along the line.

Additional features for these publication-related research data archives are also missing: to enable crediting researchers for documenting and sharing their data the datasets and code have to be citable. Therefore the assignment of a persistent identifier is urgently needed for these data archives.

Fig. 9: The graphic shows the total amount of journals in our sample, those journals equipped with a data availability policy and a replication policy, the journals with only a data availability policy; those who are both requesting data and code and finally those journals who had more than 50% of a single issue accompanied by research data/code/programs or descriptions.



Furthermore, it would be a useful feature to make these data searchable to facilitate the reuse of these data also for other research activities. Therefore the creation of additional metadata is highly important in order to have the possibility to establish the integrations of these important scientific resources in subject-specific repositories.

Linking data and publications — a new task for scientific libraries

Based on the results of our study, we see an urgent need for infrastructural solutions that go beyond attaching supplements to articles. In our opinion, the linkage between publications and their underlying research data is an interesting role that libraries could fulfil in the future. The success of discipline-specific repositories such as PANGAEA[19] or Dryad[20] exemplifies which kinds of solutions for publication-related research data are realizable.

With the following suggestions we want to intervene in the discussion on how to link datasets and publications — with a focus on the current situation in economics. Our proposal is designated to suit as basis for further discussions. Many of our thoughts on this topic are influenced by the paper *Dealing with data: Roles, rights and responsibilities* that was published by Liz Lyon in 2007 (Lyon, 2007). She lined out the roles and responsibilities of different stakeholders for managing research data.

In our opinion, the relevant stakeholders for implementing a publication-related data archive in economics consist of researchers, journal editors, publishers, research libraries and data centres. Other stakeholders comprise founders and the users of research data — but for the implementation of a publication-related data archive, the first mentioned are crucial. Each of these stakeholders has a special role to play for succeeding in building up a publication-related data archive (Table 1).

The part of the researchers as *creators* of data seems to be clear: Researchers have to meet the standards of good scientific practice and have to work up data for use by others. They have to comply with the journal's data policy and have to deposit the data they used for obtaining the results of their research papers. In addition to their data, authors have to submit at least some core metadata for their datasets — for example: author, name and version of the dataset, a short description of the dataset, some keywords etc.

Researchers as (re)*users* of data have to abide by licence conditions and have to acknowledge and to cite the creator of the dataset in their own publications when using the data of other researchers.

Editors of scholarly journals play an important role on the forefront: They are the responsible stakeholders for implementing data availability policies AND enforcing data availability for their respective journals. This is an important first step — without mandatory data policies there is little hope to receive a multitude of research data used for claiming results in publications.

For establishing a data archive editors should seek ways to cooperate with research libraries as well as with data centres for building up the necessary e-infrastructure. After establishing and using a sustainable infrastructure for publication-related data, editors should assist in managing the archive and check whether the data submitted by the author complies with their data policy.

To enable the linking of research data and publication, it is important that editors negotiate with publishers to assure them to link from the journal's website to the respective dataset and code in an external data archive. After deciding to publish an article, the editor or his/her staff has to add some core metadata (e.g. ISSN, volume, issue, page number references) to dataset(s), code and other materials. Given these core metadata libraries have the ability to link data and publication.

Thereby the major role of the publishers is outlined as well. Often publishers do not see the need to implement data archives for journals on their own (De Waard, 2012). It may raise the costs of publication and publishers do not benefit from managing a data archive as long as there are no gains to be earned for doing this task. Nevertheless, it is important that the publishers are linking datasets to the article on the journal's website. The expenses for linking data are marginal and the advantages of linking data and publications consist of a higher usage of these articles (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011). This higher usage exhibits an additional incentive for commercially orientated publishers.

The roles of libraries and data centres are not easy to delimit. Traditionally positioned at opposite ends of the research lifecycle, the convergence of data and publications and independencies between both has modified this traditional scope of duties. Both libraries and data centres are in a transition process. Today the tasks of research libraries and data centres are starting to partially overlap, but are generally in complementary roles (Reilly *et al.*, 2011). A good example of this overlapping is the creation of da|ra[21], a DOI registration agency for economic and social science research data that cooperates with the DataCite[22] consortia. Managed by GESIS[23], the Leibniz Institute for the Social Sciences and ZBW, the National Library of Economics/Leibniz Information Centre for Economics, da|ra provides persistent identifiers for datasets to make them citable.

Research data centres are skilled in the treatment of discipline specific data; they represent an important way to ensure effective data sharing and reuse (Research Information Network, 2011). Data centres have a lot of experience with these types of data and the technical know how to manage it — even for the long term. Also, data centres are knowledgeable in legal questions regarding the publication of datasets, privacy protection and access controls.[24] Therefore, data centres are predestined to take over the hosting of research data (in accordance to IPR and legal requirements), the long-term

preservation of data and code and the creation of technical metadata. Beyond this, data centres might support the research community by providing tools for the re-use of data. The problem here is that so far many data centres provide only their own data to the research community and have not opened up for external datasets (e.g. from scholarly journals).[25]

Data centres can advise researchers on how to reconfigure data for reuse by offering advice, guidance, standards and structures (Research Information Network, 2011) — but this is already a task that can also be carried out partially by research libraries. Both stakeholders could also facilitate the data submission processes by building up or adapting a user frontend[26] for depositing the data and providing training for deposit.

Libraries have been specialized in categorizing, recording, cataloguing, and provenance of publications for hundreds of years. Therefore, libraries are very experienced in their respective fields and may offer a multitude of services to the research communities. Among others, these services comprise the creation of additional descriptive and administrative metadata for research data. Besides, the cataloguing of research data and publications is a task libraries could fulfil as well as the content acquisition of datasets. In this context, libraries should open up their catalogues to research data sets; they should index them and treat them as a normal resource of the knowledge economy (Reilly *et al.*, 2011).

In addition, our profession may provide consultancy for developing and providing interoperable (metadata) standards as well as policies. Offering training opportunities or even giving lectures about replication and data availability for doctoral candidates as the Mantra project[27] at the University of Edinburgh does, is another opportunity for libraries to get involved in these future tasks.

To conclude, it is to be indicated that also the funders assign an important superordinate role in the context of linking data and publications: generally, funders set public policy drivers. Amongst others, they participate in policy coordination, joint planning and fund service delivery. In this position funders have an enormous influence in the way researchers handle their data. If funders require the publication of research funded by the public authorities as a condition for receiving grants, the whole question of obtaining research data would be processed under widely changed conditions.

Table 1: Roles, rights, responsibilities and relations in the process of linking data and publications.

Role	Rights	Responsibility	Relations
Scientist — as creator of data	To be acknowledged. To expect IPR to be honoured. To receive training and advice.	Meet standards for good practice. Work up data for use by others. Comply with journal's data policies. Submit data to journal's data archive. Submit core metadata.	With subject community With data centre/research library With founder of work With editorial office of journal
Scientist/user community — as user of data	To re-use data (non-exclusive licence). To access quality metadata to inform usability.	Abide by licence conditions. Acknowledge data creators/curators.	With research library for finding data(sets) With data centre as supplier.
Editor — creation and enforce data policies	To receive all data and materials necessary to enable replications. To receive training and advice. To select data of long-term value.	Implement data policies for journal. Monitor and enforce data availability. Ensure that data is stored in a trustworthy place or repository. Promote the repository service. Negotiate with publishers to link to journal's data archive.	With scientists as data originator With data centres as data hosts of data archive With research library for cataloguing and retrieval
Publisher — link datasets and article	To request pre-publication data deposit in data repository (-> data centre).	Link to research data to support publication standards. Support uniform data citation standards.	With scientist as creator, author and reader With data centres and research libraries as suppliers With editors as content provider

Table 1 continued

Role	Rights	Responsibility	Relations
Data Centre — curation of and access to data	To be offered a copy of data. To select data of long-term value (in accordance with editor/researcher).	Develop easy to use user front-ends to facilitate data submission. Creation of technical metadata. Manage data (and software) for the long-term. Provide training for deposit. Promote the repository service. Protect rights of data contributors. Manage data access according to IPR. Provide tools for re-use of data. Creation of persistent identifiers.	With scientist as client With user communities With research libraries through expert staff With founder of service
Research Library — cataloguing, retrieval, content acquisition	To be offered a copy of metadata.	Develop easy to use user front-ends to facilitate data submission. Creation of further descriptive and administrative metadata. Provide interoperable metadata (schemata). Creation of persistent identifiers. Provide training for deposit. Promote the repository service. Cataloguing research data and publication. Integrate research data in retrieval services and link data and publications. Content acquisition of datasets.	With scientist as client With subject community as client With data centre as data host With Editor as client With founders

Table 1 continued

Role	Rights	Responsibility	Relations
Founder - set/react to public policy drivers	To implement general data policies. To require those they fund to meet policy obligations.	Consider wider public-policy perspective & stakeholder needs. Participate in strategy co-ordination. Develop policies with stakeholders. Participate in policy co-ordination, joint planning & fund service delivery. Resource post-project long-term data management. Act as advocate for data curation & fund expert advisory service(s). Support workforce capacity development of data curators.	With scientist as founder With data centre as founder With research libraries as founder With other founders With other stakeholders as policy-maker and founder of services

Source: *Lyon (2007). Adapted by the EDaWaX-Project for the purpose of showing the role assignment for linking data and publications.*

References

AER (n.d.). *The American Economic Review - Data availability policy*. Retrieved April 20, 2013, from http://www.aeaweb.org/aer/data.php.

Anderson, R.G., Greene, W.H., McCullough, B.D., & Vinod, H.D. (2008). The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, *15*, 99–119

Andreoli-Versbach, P., & Mueller-Langer, F. (2013, February). Open access to data: An ideal professed but not practised. *RatSWD Working Papers, 215*. Retrieved April 20, 2013, from http://www.ratswd.de/dl/RatSWD_WP_215.pdf.

Bräuninger, M., Haucap, J., & Muck, J. (2011). Was lesen und schätzen Ökonomen im Jahr 2011? *DICE Ordnungspolitische Perspektiven 18*. Retrieved April 20, 2013, from http://www.econstor.eu/bitstream/10419/49023/1/667448497.pdf.

Burman, L.E., Reed, W.R., & Alm, J. (2010). A call for replication studies. *Public Finance Review 38*(6): 787–793. doi:10.1177/1091142110385210. Retrieved May 16, 2013, from http://www.sagepub.com/upm-data/36845_Replication_Studies11PFR10_787_793.pdf.

De Waard, A. (2012). Linking data to publications: Towards the execution of papers. In: P. Uhlir (Ed.), *For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop* (pp.157-159). Retrieved April 20, 2013 from https://download.nap.edu/catalog.php?record_id=13564.

Dewald, W., Thursby, J., & Anderson, R. (1986). Replication in empirical economics: The Journal of money, credit and banking project. *The American Economic Review*, *76*, 587–603.

Evanschitzky, H., & Armstrong, J. (2010). Replications of forecasting research. *International Journal of Forecasting, 26*(1), 4–8.

Evanschitzky, H., Baumgarth, C., Hubbard, R. & Armstrong, J. (2007). Replication research's disturbing trend. *Journal of Business Research, 60*, 411–415.

Glandon, P. (2010). *Report on the American economic review data availability — compliance project.* Vanderbilt University. Retrieved April 20, 2013, from http://www.aeaweb.org/aer/2011_Data_Compliance_Report.pdf.

Hamermesh, D. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d'Economique*, *40*, 715–732.

Handelsblatt Ranking Volkswirtschaftslehre - Journalliste 2010 (n.d.). Retrieved April 20, 2013, from http://tool.handelsblatt.com/tabelle/?id=33

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28*, 443–499. Retrieved April 20, 2013, from http://gking.harvard.edu/gking/files/replication.pdf.

Lyon, L. (2007): *Dealing with data - Roles, rights, responsibilities and relationships. Consultancy report.* Retrieved April 20, 2013, from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf.

McCullough, B.D. (2007). Got replicability? The Journal of Money, Credit and Banking archive. *Econ Journal Watch*, *4*, 326–337.

McCullough, B.D. (2009). Open access economics journals and the market for reproducible economic research. *Economic Analysis and Policy*, *39*(1), 117–126.

McCullough, B.D., McGeary, K.A., & Harrison, T.D. (2006). Lessons from the JMCB archive. *Journal of Money, Credit, and Banking*, *38*, 1093–1107.

McCullough, B.D., McGeary, K.A., & Harrison, T.D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics*, *41*, 1406–1420.

McCullough, B.D., & McKitrick, R. (2009*). Check the numbers: The case for due diligence in policy formation.* Fraser Institute. Retrieved April 20, 2013, from http://www.terry.uga.edu/~mustard/courses/e8420/Frasier.pdf.

McCullough, B.D., & Vinod, H.D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, *93,* 873–892.

McCullough, B.D., & Vinod, H.D. (2008). The role of data/code archives in the future of economic research. *The Journal of Economic Methodology*, *15*, 99–119.

Mirowski, P., & Sklivas, S. (1991). Why econometricians don't replicate (although they do reproduce). *Review of Political Economy*, *3*, 146–163.

Pesaran, M.H. (n.d.). Journal of Applied Econometrics - News. Retrieved April 20, 2013, from http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-1255/homepage/News.html.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications, Opportunities for Data Exchange project (ODE).* Retrieved April 20, 2013, from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf.

Research Information Network (2011). *Data centres: their use, value and impact.* Retrieved April 20, 2013, from http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf.

Thomson Reuters (2011). *2010 Journal Citation Reports (Social Sciences Edition)* [Database]. Retrieved March 12, 2012, from http://admin-apps.webofknowledge.com/JCR/JCR?RQ=LIST_SUMMARY_JOURNAL&cursor=181.

Notes

[1] A very common example for this type of research data is the German Socio-Economic Panel Study (SOEP): http://www.diw.de/en/diw_02.c.221178.en/about_soep.html

[2] A prominent example for irreproducible/incorrect research in the recent past that massively influenced policy-makers is the case of Rogoff and Reinert which leads to my controversial discussions in the community. (see http://www.bloomberg.com/news/2013-04-28/refereeing-the-reinhart-rogoff-debate.html).

[3] In economic journals, two types of data policies can be distinguished: A replication policy requires authors to provide data and code to would-be replicators. In contrast a data availability policy mandates authors to provide data and code to the journal. The journal provides this information to would-be replicators by attaching the data and code to the article (often in "supplementary information"). The terms replication policy and data availability policy have been defined in: McCullough, McGeary, & Harrison (2008).

[4] A useful attempt has been implemented by the NEREUS-Network in the course of the "Network of European Economists Online" (NEEO) project (http://www.economistsonline.org/home): NEEO had a runtime from 2007 till 2010. In this project existing research resources such as RePEc (http://www.repec.org) and new content of excellence from over 1000 top economics scholars, are made available through the Economists Online portal. Though the portal provides bibliographic information for more than 900,000 research articles and access to several thousand full-texts, there are only 142 datasets (and even a much smaller number of datasets accompanied by the code of computation) available in the NEEO Dataverse (http://dvn.iq.harvard.edu/dvn/dv/NEEO).

[5] The sole publication-related data archive we were able to find in the course of the project is currently available at the ICPSR (Inter-university Consortium for Political and Social Research): http://www.icpsr.umich.edu/icpsrweb/deposit/pra/index.jsp

[6] http://www.edawax.de

[7] Results of this work package are available at http://www.edawax.de/wp-content/uploads/2013/01/EN-EDaWaX-Online-Survey-Hosting-Options_blog.pdf

[8] For additional information on the EDaWaX project please visit the project web blog: www.edawax.de

[9] Data archive of the AER: http://www.aeaweb.org/aer/contents/index.php

[10] For the background of this recommendation I also want to refer to the case of Oberholzer-Gee/Strumpf vs. Liebowitz. http://regulation2point0.org/wp-content/uploads/downloads/2010/05/The-Oberholzer-Gee-Strumpf-File-Sharing-Instrument-Fails-the-Laugh-Test.pdf

[11] To determine the average impact factor and ranking for groups of journals with different data policies, the following numbers of journals have been included in our analyses:

- For analyzing the ranking within the JCR 21 journals equipped with a data availability policy (72.4% of all journals with such a policy in our sample), 11 journals equipped with a replication policy (100% of all journals with such a policy in our sample) and 74 journals without a data policy (73.3% of all journals with such a policy in our sample) have been included.
- For the analyses regarding the Handelsblatt ranking 28 journals with a data availability policy (96.6%), 11 journals (100%) with a replication policy and 96 journals without a data policy (95.1%) were included.
- For the analyses regarding the ranking of BHM, 26 journals with a data availability policy (89.7%), 11 journals (100%) with a replication policy and 100 journals without a data policy (99%) were included.

The journals not included could not be used for these calculations, because they were not listed.

[12] In the course of our analysis we found a case where the data policy is available in the printed edition only (*German Economic Review*). For other journals we were able to find a data archive, but not a data policy (e.g. *Jahrbücher Nationalökonomie und Statistik*, *Journal of Financial Economics*, *The Federal Reserve Bank of St. Louis Review*). These cases were not included into the analysis of the data availability policies but were categorized as journals without data policy.

[13] The original wording of all data availability and replication policies we found in the course of our analyses is available on the project blog: http://www.edawax.de/wp-content/uploads/2012/07/Data_Policies_WP2.pdf

[14] The remaining journals did not request authors to submit datasets because they are focussed on experimental data. In these journals providing of data was optional.

[15] An example for a data archive in economic scholarly journals (here the data archive of the *American Economic Review*) is available here: http://www.aeaweb.org/aer/contents/index.php

[16] Readers interested in Dataverse should visit the Dataverse homepage (www.thedata.org) for more information. An interesting overlook on Dataverse was provided by Mercè Crosas, Director of Product Development, in a workshop on Persistent Identifiers in Berlin at May the 8th 2012. Her presentation is available here: http://www.ratswd.de/ver/docs_PID_2012/Crosas_PID2012.pdf

[17] A good example for a dataset and code in Dataverse is available here: http://dvn.iq.harvard.edu/dvn/dv/arzheimer/faces/study/StudyPage.xhtml?globalId=hdl:1902.1/12092&studyListingIndex=0_e757de6b960f442ef22a63c6b03a

[18] Beside the JAE, also the *Journal of Economic and Social Measurement* and the *International Journal of Research in Marketing* own a replication section. In addition the *Public Finance Review* published a call for replication studies in 2010 (Burman, Reed, & Alm, 2010).

[19] The information system PANGAEA (http://www.pangaea.de) is operated as an open-access library aimed at archiving, publishing and distributing georeferenced data from earth system research. Inter alia PANGAEA was able to conclude an agreement with Elsevier with the result that the research data used within an article and available at PANGAEA is shown on the website of the research article published in an Elsevier Journal.

[20] Dryad is a discipline-specific research data repository for the basic and applied biosciences: www.datadryad.org.

[21] The homepage of da | ra is http://www.da-ra.de/en/home/

[22] Further Information on Datacite, that is part of the international DOI-Foundation (IDF) may be obtained on the website http://datacite.org/

[23] http://www.gesis.org

[24] For Germany, the RatSWD (German Data Forum: http://www.ratswd.de) has formulated some criteria for research data centres. For obtaining certification from the German Data Forum they have to fulfil some requirements, listed on the webpage http://www.ratswd.de/download/publikationen_rat/RatSWD_FDZCriteria.pdf

[25] An online-survey among German and European research data centres supports this finding. For more information on the results of this survey, please visit http://www.edawax.de/wp-contentt/uploads/2013/01/EN-EDaWaX-Online-Survey-Hosting-Options_blog.pdf

[26] Examples of software solutions comprise Nesstar (http://www.nesstar.com/), Dataverse (http://thedata.org/) or CKAN (http://ckan.org/) exemplarily

[27] Some information on Mantra can be obtained via the website of the University of Edinburgh: http://www.ed.ac.uk/schools-departments/information-services/about/organisation/edl/data-library-projects/mantra